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

(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,

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

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

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

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 not imply that $\liminf_{n\to\infty} n \operatorname{Var}_\theta(\tilde{\theta}_n) \geq I(\theta)^{-1}$ 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

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

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

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* $\mathrm{Cov}(X) = \Sigma$ *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 $X_1, \dots, X_n$ 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 $l(\theta) = \mathbb{E}_{\theta_0}\left[\log f(X, \theta)\right]$.
- Assumption 5 implies that continuity of the function $\theta \mapsto \log f(x, \theta)$ carries over to continuity of $\theta \mapsto \mathbb{E}_\theta\left[\log f(X, \theta)\right] = l(\theta)$, 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 $X_1, \dots, X_n$ 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. □

Remark: the proof can be simplified 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 $X_1, \dots, X_n$ 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 $\{\theta_1, \dots, \theta_M\} \subseteq \Theta$, we have

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

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

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

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

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

□

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 $1 \le j \le N(\varepsilon)$*

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 $X_1, \dots, X_n$ 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 $n \geq n_0(\varepsilon)$. 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 $\{\theta_1, \dots, \theta_N\} \subseteq \Theta$ 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

$$\bar{A}_n \xrightarrow{P} \mathbb{E}_{\theta_0}[\mathbf{H}_\theta \log f(X,\theta)] = -I(\theta_0)$$

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 $\bar{A}_n \xrightarrow{P} \mathbb{E}_{\theta_0}[\mathbf{H}_\theta \log f(X,\theta)] = -I(\theta_0)$ 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). □

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 $n \operatorname{Var}_\theta(\tilde{\theta}_n) \to I(\theta)^{-1}$ for all $\theta \in \operatorname{int}(\Theta)$ (if $p = 1$) or when $p > 1$ $n \operatorname{Cov}_\theta(\tilde{\theta}_n) \to I(\theta)^{-1}$ fo all $\theta \in \operatorname{int}(\Theta)$.

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 minimax risk.

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. □

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 $\hat{\theta}_{\text{MLE}} = \bar{X}_n$. SO $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1)$. Suppose $\Phi(\theta) = \theta^2$. Then the Delta Method implies $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0^2) \xrightarrow{d} 2\theta_0 \mathcal{N}(0,1) = \mathcal{N}(0, 4\theta_0^2)$. 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

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

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

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

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 $\hat{\theta}_{\text{MLE}}$.

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 $\hat{i}_n = i_n(\hat{\theta}_{\text{MLE}})$ 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

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

The first term is upper bounded by

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

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 $\hat{j}_n(\theta) = j_n(\hat{\theta}_{\mathrm{MLE}})$, 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

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

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:

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta \setminus \Theta_0 \end{aligned}$$

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 $\theta_0 \in \operatorname{int}(\Theta)$. 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 $\hat{\theta}_n \in \operatorname{int}(\Theta)$. 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 $\Lambda_n(\Theta, \Theta_0) \xrightarrow{d} \chi^2_p$. □

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 hypotheses, 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 $\Theta = \{\theta_1, \dots, \theta_k\}$, 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 $\theta | X$.

Let $\{f(\cdot,\theta) : \theta \in \Theta\}$ be a statistical model.

Definition. For a sample space $\mathcal{X}$ of the observation X, consider the product space $\mathcal{X} \times \Theta$ with corresponding pdf/pmf $q(x,\theta) = f(x,\theta)\pi(\theta)$. The distribution π is the *prior distribution* of θ. The *posterior distribution* is the conditional probability

$$\theta|X \sim \frac{f(x,\theta)\pi(\theta)}{\int_\Theta f(x,\theta')\pi(\theta')\mathrm{d}\theta'} := \Pi(\theta|X)$$

Note that the conditional probability of X given θ is

$$X|\theta \sim \frac{f(x,\theta)\pi(\theta)}{\int_{\mathcal{X}} f(x',\theta)\pi(\theta)\mathrm{d}x'} = f(x,\theta)$$

Also, if $X_1,\ldots,X_n$ are iid samples with common law $f(x,\theta)$, then

$$\theta|X_1,\ldots,X_n \sim \frac{\prod_{i=1}^n f(x_i,\theta)\pi(\theta)}{\int_\Theta \prod_{i=1}^n f(x_i,\theta')\pi(\theta')\mathrm{d}\theta'} := \Pi(\theta|X_1,\ldots,X_n)$$

The last expression is just a reweighted and renormalised version of the likelihood function. In fact, the denominator is often ignored in calculations.

Example. Suppose $X|\theta \sim \mathcal{N}(0,1)$, with prior $\theta \sim \mathcal{N}(0,1)$. The numerator of the posterior distribution $\theta|X_1,\ldots,X_n$ is proportional to (as a function of θ)

$$\begin{aligned}
\exp\left(-\frac{\theta^2}{2}\right)\prod_{i=1}^n \exp\left(-\frac{(X_i-\theta)^2}{2}\right) &\propto \exp\left(n\theta\bar{X} - \frac{n\theta^2}{2} - \frac{\theta^2}{2}\right) \\
&= \exp\left(n\theta\bar{X} - \frac{(n+1)\theta^2}{2}\right) \\
&\propto \exp\left(-\frac{\left(\theta\sqrt{n+1} - \frac{n\bar{X}}{\sqrt{n+1}}\right)^2}{2}\right) \\
&= \exp\left(-\frac{\left(\theta - \frac{n\bar{X}}{n+1}\right)^2}{\frac{2}{n+1}}\right)
\end{aligned}$$

Thus, $\theta|X_1,\ldots,X_n \sim \mathcal{N}\left(\frac{1}{n+1}\sum_{i=1}^n X_i, \frac{1}{n+1}\right)$. For the more general case of $\mathcal{N}(\theta,\sigma^2)$ with prior $\mathcal{N}(\mu,\nu^2)$, see Example sheet.

Note that in the above example, the posterior distribution is in the same distribution class as the prior (both normal), except the parameters have been updated based on the X_i's.

Definition. In a statistical model $\{f(\cdot,\theta) : \theta \in \Theta\}$, if the prior $\pi(\theta)$ and posterior $\pi(\theta|X)$ belong to the same family of distributions, then $\pi(\theta)$ is a *conjugate prior.*

The following examples of conjugate priors are in the Example sheet:

1. Normal prior and normal sampling - gives normal posterior
2. Beta prior and binomial sampling - gives beta posterior
3. Gamma prior and Poisson sampling - gives Gamma posterior

Definition. A prior non-negative function with indefinite integral $\int_\Theta f(x,\theta)\pi(\theta)\mathrm{d}\theta$ is an *improper prior.*

As an example, consider the prior $\pi(\theta) = 1$ for all θ. But as long as the posterior is a valid pdf, a prior such as the uniform prior can still be used for estimation.

Jeffreys prior

A potential issue with the uniform prior is that it is not invariant to reparametrisation. Suppose $X|\theta \sim \text{Binomial}(n,\theta)$. Suppose the prior on θ is Uniform$[0,1]$. However, suppose we reparametrise with $\phi = \theta^{100}$.

Definition. The prior $\pi(\theta)$ proportional to $\sqrt{\det(I(\theta))}$ is called the Jeffreys prior.

Consider a monotonic function h, and let $\phi = h(\theta)$. We can write the Fisher information as

$$\begin{aligned}
I_\phi(\phi) &= -\mathbb{E}\left[\frac{\mathrm{d}^2 \log p(x|\phi)}{\mathrm{d}\phi^2}\right] \\
&= -\mathbb{E}\left[\frac{\mathrm{d}^2 \log p(X|\phi)}{\mathrm{d}\theta^2}\left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2 + \frac{\mathrm{d}\log p(X,\theta)}{\mathrm{d}\theta}\frac{\mathrm{d}^2\theta}{\mathrm{d}\phi^2}\right] \\
&= -\mathbb{E}\left[\frac{\mathrm{d}^2 \log p(X,\theta)}{\mathrm{d}\theta^2}\right]\left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2 - \underbrace{\mathbb{E}\left[\frac{\mathrm{d}\log p(X,\theta)}{\mathrm{d}\theta}\right]}_{=0}\frac{\mathrm{d}^2\theta}{\mathrm{d}\phi^2} \\
&= I_\theta(\theta)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2 \\
&= \frac{I_\theta(\theta)}{(h'(\theta))^2}
\end{aligned}$$

(assuming that θ, ϕ are univariate). Thus $\pi_\phi(\phi) = \frac{\pi_\theta(\theta)}{|h'(\theta)|} \propto \frac{\sqrt{I_\theta(\theta)}}{|h'(\theta)|} = \sqrt{I_\phi(\phi)}$.

As a concrete example, the Example sheet considers the $\mathcal{N}(\mu,\sigma^2)$ model with $\theta = \begin{pmatrix}\mu \\ \sigma^2\end{pmatrix} \in \mathbb{R} \times (0,\infty)$. Then the Fisher information matrix is $I(\theta) =$

$\begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}$ so the Jeffreys prior is $\pi(\theta) \sim \frac{1}{\sigma^2}$, which is constant in μ, and therefore improper. However the posterior marginal of μ (conditioned on σ^2) is $\mathcal{N}(\bar{X}_n, \frac{\sigma^2}{n})$, which makes sense as a posterior. This can be compared with the posterior under the prior $\pi \sim \mathcal{N}(\theta, 1)$, which would be $\mathcal{N}\left(\frac{\bar{X}_n + \frac{\theta\sigma^2}{n}}{1+\frac{\sigma^2}{n}}, \frac{\sigma^2}{n+\sigma^2}\right)$.

Other priors motivated by other notions of "low-informativeness" lead to difference familes of reference priors.

Statistical inference with the posterior

The distribution $\pi(\cdot|X_1, \ldots, X_n)$ is a random probability measure on the parameter space Θ.

The posterior can be used for various tasks such as estimation, uncertainty quantification, and hypothesis testing (from a Bayesian perspective).

Statistical inference with the posterior

Based on the posterior distribution $\Pi(\cdot|X_1,\ldots,X_n)$, we might address the following statistical questions:

1. Estimation: posterior mean $\bar{\theta}(X_1,\ldots,X_n)=\mathbb{E}_\Pi[\theta|X_1,\ldots,X_n]$.
2. Uncertainty quantification: take a subset $\mathcal{C}_n\subseteq\Theta$ such that $\Pi(\mathcal{C}_n|X_1,\ldots,X_n)=1-\alpha$. This is a *credible set.*
3. Hypothesis testing: as in the motivating example from Lecture 10, the Bayes factor is defined as

$$\frac{\Pi(\Theta_0|X_1,\ldots,X_n)}{\Pi(\Theta|X_1,\ldots,X_n)}=\frac{\int_{\Theta_0}\prod_{i=1}^n f(X_i,\theta)\pi(\theta)\mathrm{d}\theta}{\int_{\Theta_1}\prod_{i=1}^n f(X_i,\theta)\pi(\theta)\mathrm{d}\theta}$$

Frequentist behaviour of credible sets

Suppose $X_i\sim^{\text{iid}} f(x,\theta_0)$. We will discuss the behaviour of $\mathcal{C}_n$ as $n\to\infty$. As an example, suppose $X|\theta\sim\mathcal{N}(0,1)$ and we take the prior $\theta\sum\mathcal{N}(0,1)$.

We have seen that

$$\theta|X_1,\ldots,X_n\sim\mathcal{N}\left(\frac{1}{n+1}\sum_{i=1}^n X_i,\frac{1}{n+1}\right)$$

Thus the posterior mean is

$$\bar{\theta}_n=\mathbb{E}_\pi[\theta|X_1,\ldots,X_n]=\frac{1}{n+1}\sum_{i=1}^n X_i$$

This is not the MLE, but it is very close. Also, it is consistent: if $X_i\sim^{\text{iid}}\mathcal{N}(\theta_0,1)$, then $\bar{\theta}_n=\frac{n}{n+1}\frac{1}{n}\sum_{i=1}^n X_i\xrightarrow{P}\theta_0$ by the LLN and Slutsky's lemma.

Now take $\sqrt{n}\left(\bar{\theta}_n-\theta_0\right)$. If we write

$$\sqrt{n}\left(\bar{\theta}_n-\theta_0\right)=\sqrt{n}\left(\hat{\theta}_n-\theta_0\right)+\sqrt{n}\left(\bar{\theta}_n-\hat{\theta}_n\right)$$

We see that $\sqrt{n}\left(\hat{\theta}_n-\theta_0\right)\xrightarrow{d}\mathcal{N}(0,1)$ and

$$\sqrt{n}\left(\bar{\theta}_n-\hat{\theta}_n\right)=\sqrt{n}\left(\frac{1}{n+1}-\frac{1}{n}\right)\sum_{i=1}^n X_i=-\frac{\sqrt{n}}{n+1}\bar{X}_n\xrightarrow{P}0$$

Since $\bar{X}_n\xrightarrow{P}\theta_0$ and $\frac{\sqrt{n}}{n+1}\to 0$ so apply Slutsky's lemma. So altogether we have $\sqrt{n}\left(\bar{\theta}_n-\theta_0\right)\xrightarrow{d}\mathcal{N}(0,1)$. Thus, we could construct frequentist confidence intervals of the form

$$\mathcal{C}_n=\left\{\nu:|\nu-\bar{\theta}_n|\leq\frac{z_\alpha}{\sqrt{n}}\right\}$$

i.e $\hat{\theta}_n$ is replaced by $\bar{\theta}_n$as the centre of the confidence interval.

However, in Bayesian analysis, credible sets are based on the posterior distribution, not the asymptotic distribution of an estimator.

We should build a set of the form $\mathcal{C}_n = \{\nu : |\nu - \bar{\theta}_n| \leq \frac{R_n}{\sqrt{n}}\}$ where R_n is taken such that $\Pi(\mathcal{C}_n|X_1, \dots, X_n) = 1 - \alpha$. Note that R_n is in general a random variable based on the X_i's.

The main result we will show is that under appropriate assumptions, $\mathbb{P}_{\theta_0}(\theta_0 \in \mathcal{C}_n) \to 1 - \alpha$ as $n \to \infty$, when $\mathcal{C}_n$ is a credible set, so these are valid frequentist condidence intervals.

Theorem 1.24 (Bernstein - von Mises)**.** *Consider a parametric model $\{f(\cdot, \theta) : \theta \in \Theta\}$ for $\Theta \subseteq \mathbb{R}$ satisfying the usual regularity conditions, a prior with continuous density π at θ_0 with $\pi(\theta_0) > 0$, and the associated posterior $\Pi(\cdot|X_1, \dots, X_n)$. Let $\hat{\phi}_n$ denote the random distribution $\mathcal{N}(\hat{\theta}_n, \frac{I(\theta_0)^{-1}}{n})$ (where $\hat{\theta}_n$ is the MLE). As $n \to \infty$ under P_{θ_0}, we have*

$$\left\|\Pi_n - \hat{\phi}_n\right\|_{TV} = \int_\Theta |\Pi_n(\theta) - \hat{\phi}_n| \mathrm{d}\theta \xrightarrow{a.s} 0$$

Remark: total variation distance (TV), integrates the absolute difference between two pdf's.

Proof. We will only prove the special case $X|\theta \sim \mathcal{N}(0, 1)$ and $\theta \sim \mathcal{N}(0, 1)$. The proof for the more general result is beyond the scope of the course.

Note that since Π_n and $\hat{\phi}_n$ are probability distributions, they both integrate to 1, so $\int_\Theta \left(\Pi_n(\theta) - \hat{\phi}_n(\theta)\right) \mathrm{d}\theta = 0$. Thus the integrals of the positive and negative parts are the same, so

$$\int_\Theta |\Pi_n(\theta) - \hat{\phi}_n(\theta)| \mathrm{d}\theta = 2 \int_\Theta \left(\hat{\phi}_n(\theta) - \Pi_n(\theta)\right)^+ \mathrm{d}\theta$$

$$= 2 \int_\Theta \left(1 - \frac{\Pi_n(\theta)}{\hat{\phi}_n(\theta)}\right)^+ \hat{\phi}_n(\theta) \mathrm{d}\theta$$

In this specific setting, the distribution of $\hat{\phi}_n$ is $\mathcal{N}(\bar{X}_n, \frac{1}{n})$ and the distribution of Π_n is $\mathcal{N}(\bar{\theta}_n, \frac{1}{n+1}) = \mathcal{N}(\frac{n}{n+1}\bar{X}_n, \frac{1}{n+1})$.
Thus

$$\frac{\Pi_n(\theta)}{\hat{\phi}_n(\theta)} = \frac{\frac{1}{\sqrt{\frac{2\pi}{n+1}}} \exp\left(-\frac{n+1}{2}\left(\theta - \bar{\theta}_n\right)^2\right)}{\frac{1}{\sqrt{\frac{2\pi}{n+1}}} \exp\left(-\frac{n}{2}\left(\theta - \hat{\theta}_n\right)^2\right)}$$

Making the substitution $v = \sqrt{n}\left(\theta - \hat{\theta}_n\right)$, we can write the integral as

$$2\int_{\Theta}\left(1 - \sqrt{\frac{n+1}{n}}\exp\left(-\frac{n+1}{2n}\left(v + \sqrt{n}\left(\hat{\theta}_n - \bar{\theta}_n\right)\right)^2 + \frac{v^2}{2}\right)\right)^+ \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{v^2}{2}\right)\mathrm{d}v$$

Since the function $x \mapsto (1-x)^+$ is bounded by 1 for $x \geq 0$, if it can be show that almost surely (with respect to P_{θ_0}) the inner term converges to 0 for all v, then the dominated convergence theorem implies convergence to 0.

As argued before, $\sqrt{n}(\bar{\theta}_n - \hat{\theta}_n) \xrightarrow{\text{a.s}} 0$. Thus it is easy to see that for any fixed v, the term in the exponential converges to 0, concluding the proof. □

Theorem 1.25. *Under the same assumptions as the Bernstein - von Mises theorem, for any $\alpha \in (0,1)$, we have $\mathbb{P}_{\theta_0}(\theta_0 \in \mathcal{C}_n) \to 1-\alpha$ as $n \to \infty$ where $\mathcal{C}_n = \{\nu : |\nu - \hat{\theta}_n| \leq \frac{R_n}{\sqrt{n}}\}$ and R_n is choses such that $\Pi_n(\mathcal{C}_n) = 1-\alpha$.*

Proof. We will prove this in the general case (not just for normal distributions). First we will show that $R_n \xrightarrow{\text{a.s}} \Phi_0^{-1}(1-\alpha)$, where Φ_0 is the cdf of a standard normal. Concretely, Φ_0 is defined by $\Phi_0(t) = \mathbb{P}(|Z_0| \leq t) = \int_{-t}^{t} \varphi_0(x)\mathrm{d}x$, where $Z_0 \sim \mathcal{N}(0, I(\theta_0)^{-1})$. Note that Φ_0 has a well-defined, continuous inverse denoted Φ_0^{-1}.

We can write

$$\Phi_0(R_n) = \int_{-R_n}^{R_n} \varphi_0(v)\mathrm{d}v = \int_{\hat{\theta}_n - \frac{R_n}{\sqrt{n}}}^{\hat{\theta}_n - \frac{R_n}{\sqrt{n}}} \hat{\phi}_n(\theta)\mathrm{d}\theta = \hat{\phi}_n(\mathcal{C}_n)$$

Where we made the substitution $\theta = \hat{\theta}_n + \frac{v}{\sqrt{n}}$ and $\hat{\phi}_n$ is the pdf of $\mathcal{N}(\hat{\theta}_n, \frac{I(\theta_0)^{-1}}{n})$.

Furthermore, $\hat{\phi}_n(\mathcal{C}_n) = (\hat{\phi}_n(\mathcal{C}_n) - \Pi_n(\mathcal{C}_n)) + 1 - \alpha$. By the Bernstein - von Mises theorem, $\hat{\phi}_n(\mathcal{C}_n) - \Pi_n(\mathcal{C}_n) \xrightarrow{\text{a.s}} 0$. Therefore $\Phi_0(R_n) \xrightarrow{\text{a.s}} 1-\alpha$. So by the Continuous Mapping Theorem, $R_n \xrightarrow{\text{a.s}} \Phi_0^{-1}(1-\alpha)$.

Next, using Slutsky's lemma and noting that $\Phi_0^{-1}(1-\alpha) > 0$, we have $\frac{\Phi_0^{-1}(1-\alpha)}{R_n}\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$. Thus

$$\begin{aligned}
\mathbb{P}_{\theta_0}(\theta_0 \in \mathcal{C}_n) &= \mathbb{P}_{\theta_0}\left(|\hat{\theta}_n - \theta_0| \leq \frac{R_n}{\sqrt{n}}\right) \\
&= \mathbb{P}_{\theta_0}\left(\frac{\Phi_0^{-1}(1-\alpha)}{R_n}\sqrt{n}|\hat{\theta}_n - \theta_0| \leq \Phi_0^{-1}(1-\alpha)\right) \\
&\to \mathbb{P}(|Z_0| \leq \Phi_0^{-1}(1-\alpha)) \\
&= \Phi_0(\Phi_0^{-1}(1-\alpha)) \\
&= 1-\alpha
\end{aligned}$$

□

Decision Theory

Given a statistical model $\{f(\cdot,\theta) : \theta \in \Theta\}$, and an observed sample $X \in \mathcal{X}$, we can phrase many statistical problems as decision problems, with an *action space* $\mathcal{A}$ and *decision rules* $\delta : \mathcal{X} \to \mathcal{A}$.
Examples:

1. Hypothesis testing: $\mathcal{A} = \{0,1\}$ and the decision $\delta(X)$ is a test

2. Estimation: $\mathcal{A} = \Theta$ and $\delta(x) = \hat{\theta}(X)$ is an estimator

3. Inference: $\mathcal{A} = \mathcal{P}(\Theta)$ (powerset of Θ) and $\delta(X) = \mathcal{C}(X)$ is a confidence set

The performance of a decision rule is assessed by a *loss function* $L : \mathcal{A} \times \Theta \to [0, \infty)$. For example, in hypothesis testing, the answer is right or wrong, so taking $\theta \in \{0, 1\}$ representing the index of the hypothesis, we could have $L(a, \theta) = 1\{a \neq \theta\}$ (0-1 loss).

In an estimation problem, we might take $L(a, \theta) = |a - \theta|$, or $(a - \theta)^2$ (in one dimension).

Definition. For a loss L, and a decision rule $\delta(X)$, with $X \sim P_\theta$, we define the *risk* $R(\delta, \theta) = \mathbb{E}_\theta[L(\delta(X), \theta)] = \int_{\mathcal{X}} L(\delta(x), \theta) f(x, \theta) \mathrm{d}x$.

Examples:

1. In hypothesis testing, $R(\delta, \theta) = \mathbb{E}_\theta[1\{\delta(X) \neq \theta\}] = \mathbb{P}_\theta(\delta(X) \neq \theta)$. This is the probability of type I/type II error.

2. In estimation, the quadratic risk corresponds to the mean squared error (MSE): $R(\hat{\theta}, \theta) = \mathbb{E}_\theta[(\hat{\theta}(X) - \theta)^2]$.

   As a special case, for $X \sim \mathrm{Bin}(n, \theta)$ and $\theta \in [0, 1]$, if we take $\hat{\theta}(X) = \frac{X}{n}$, we can check that $R(\hat{\theta}, \theta) = \frac{\theta(1-\theta)}{n}$. If we instead consider the estimator $\hat{\eta}(X) = \frac{1}{2}$, we have

   $$R(\hat{\eta}, \theta) = \mathbb{E}_\theta[(\hat{\eta}(X) - \theta)^2] = \left(\theta - \frac{1}{2}\right)^2$$

   From this we can see that it is generally impossible to uniformly compare two estimators, with respect to the risk function.

Definition. Given a prior π on Θ, the π-*Bayes risk* of δ for the loss function L is defined by

$$R_\pi(\delta) = \mathbb{E}_\pi[R(\delta, \theta)] = \int_\theta R(\delta, \theta)\pi(\theta)\mathrm{d}\theta = \int_\theta \int_{\mathcal{X}} L(\delta(x), \theta)\pi(\theta) f(x, \theta)\mathrm{d}x\mathrm{d}\theta$$

A π-Bayes decision rule δ_π is any decision rule that minimises the π-Bayes risk $R_\pi(\delta)$.

Example: for the binomial example above, if we take the uniform prior on $[0, 1]$ for π, then the π-Bayes risk is

$$R_\pi\left(\frac{X}{n}\right) = \mathbb{E}_\pi\left[\frac{\theta(1-\theta)}{n}\right] = \frac{1}{n}\int_0^1 \theta(1-\theta)\mathrm{d}\theta = \frac{1}{6n}$$

Definition. For a Bayesian model, the *posterior risk* is defined as the average loss under the posterior distribution for an observation $x \in \mathcal{X}$:

$$\mathbb{E}_\Pi\left[L(\delta(x), \theta)|x\right] = \int_\Theta L(\delta(x), \theta)\Pi(\theta|z)\mathrm{d}\theta$$

Theorem 1.26. *An estimator δ that minimises the Π-posterior risk also minimises the π-Bayes risk R_π.*

Proof. The π-Bayes risk may be written as

$$\begin{aligned} R_\pi(\delta) &= \int_{\mathcal{X}}\int_{\Theta} L(\delta(x),\theta) f(x,\theta)\pi(\theta)\mathrm{d}\theta\mathrm{d}x \\ &= \int_{\mathcal{X}}\int_{\Theta} L(\delta(x),\theta)\underbrace{\frac{f(x,\theta)\pi(\theta)}{\int_\Theta f(x,\theta')\pi(\theta')\mathrm{d}\theta'}}_{\Pi(\theta|X)}\underbrace{\int_\Theta f(x,\theta')\pi(\theta')\mathrm{d}\theta'}_{:=m(x)\geq 0}\mathrm{d}\theta\mathrm{d}x \\ &= \int_{\mathcal{X}} \mathbb{E}_\Pi[L(\delta(x),\theta)|x]m(x)\mathrm{d}x \end{aligned}$$

If δ_Π is a decision rule minimising posterior risk, then for any other decision rule δ, and for all $x \in \mathcal{X}$, we have

$$\mathbb{E}_\Pi[L(\delta_\Pi(x),\theta)|x] \leq \mathbb{E}_\Pi[L(\delta(x),\theta)|x]$$

Multiplying by $m(x)$ and integrating with respect to x, shows

$$R_\pi(\delta_\Pi) \leq R_\pi(\delta)$$

□

Example. For the quadratic risk (with squared loss), the posterior risk can be written as

$$\mathbb{E}_\Pi[(\delta(x)-\theta)^2|x] = \mathbb{E}_\Pi[\delta(x)^2 - 2\delta(x)\theta + \theta^2|x] = \delta(x)^2 - 2\delta(x)\mathbb{E}_\Pi[\theta|x] + \mathbb{E}_\Pi[\theta^2|x]$$

Since this is a quadratic in $\delta(x)$, the posterior risk is minimised by taking the posterior mean

$$\delta_\Pi(x) = \mathbb{E}_\Pi[\theta|x]$$

Other losses give other ways to minimise the posterior risk, leading to other decision rules. One can also show that this is the unique π-Bayes rule for the quadratic risk (see Example sheet).

The following theorem shows that unbiased estimators are typically disjoint from Bayes estimators.

Theorem 1.27. *Let δ be an unbaised decision rule for θ, i.e $\mathbb{E}_\theta[\delta(X)] = \theta$ for all $\theta \in \Theta$. If δ is also a Bayes rule for some prior π in the quadratic risk, then $\mathbb{E}_Q[(\delta(X)-\theta)^2] = 0$ where $\mathbb{E}_Q$ is the expectation taken with repect to the joint distribution $q(x,\theta) = f(x,\theta)\pi(\theta)$. In particular, $\delta(X) = \theta$ with probability 1 under Q.*

Proof. Recall that for any random variable $Z(X,\theta)$, by applying the "tower rule" of expectation in two different ways, we have

$$\mathbb{E}_Q[Z(X,\theta)] = \mathbb{E}[\mathbb{E}_\Pi[Z(X,\theta)|X]] = \mathbb{E}[\mathbb{E}_\theta[Z(X,\theta)]] \qquad (*)$$

From the calculation in the previous example, the unique π-Bayes rule (which minimises $\mathbb{E}_Q[(\delta(X)-\theta)^2]$) is given by $\delta(X)=\mathbb{E}_\Pi[\theta|X]$. Taking $Z(X,\theta)=\theta\delta(X)$ in $(*)$, we have

$$\mathbb{E}_Q[\theta\delta(X)]=\mathbb{E}[\mathbb{E}_\Pi[\theta\delta(X)|X]]=\mathbb{E}[\delta(X)\mathbb{E}_\Pi[\theta|X]]=\mathbb{E}[\delta(X)^2]$$

$$\mathbb{E}_Q[\theta\delta(X)]=\mathbb{E}[\mathbb{E}_\theta[\theta\delta(X)]]=\mathbb{E}[\theta\mathbb{E}_\theta[\delta(X)]]=\mathbb{E}[\theta^2]$$

Hence

$$\mathbb{E}_Q[(\delta(X)-\theta)^2]=\mathbb{E}_Q[\delta(X)^2]-2\mathbb{E}_Q[\theta\delta(X)]+\mathbb{E}_Q[\theta^2]=0$$

□

Remark: the following statements can be derived from the above theorem

1. In a $\mathcal{N}(\theta,1)$ model, the MLE $\bar{X}_n$ is not a Bayes estimator for any prior π.
2. In a binomial model, the MLE $\frac{X}{n}$ is only Bayes in degenerate cases (see Example sheet).

Minimax Risk

Bayes risk takes the average risk of a decision rule with respect to a prior. We might alternatively be interested in worst-case risk.

Definition. The *maximal risk* of the decision rule δ over the parameter space Θ is defined by $R_m(\delta,\Theta)=\sup_{\theta\in\Theta}R(\delta,\theta)$. The *minimax risk* is defined as the infimum of the maximum risk:

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

A decision rule that attains the minimax risk is called *minimax*.

Going back to the $X\sim\text{Binom}(n,\theta)$ example from before, let $\Theta=[0,1]$ and $\hat{\theta}(X)=\frac{X}{n}$ and $\hat{\eta}=\frac{1}{2}$.

$$R(\hat{\theta},\theta)=\mathbb{E}_\theta[(\hat{\theta}(X)-\theta)^2]=\frac{\theta(1-\theta)}{n}$$

$$R(\hat{\eta},\theta)=(\theta-\frac{1}{2})^2$$

Thus $R_m(\hat{\theta},\Theta)=\frac{1}{4n}$ and $R_m(\hat{\eta},\Theta)=\frac{1}{4}$.

Theorem 1.28. *For any prior π and decision rule δ, we have*

$$R_\pi(\delta)\leq R_m(\delta,\Theta)$$

Proof. $R_\pi=\mathbb{E}_\pi[R(\delta,\theta)]\leq\sup_{\theta\in\Theta}R(\delta,\theta)=R_m(\delta,\Theta)$ (i.e average is bounded above by worst case). □

Definition. A prior λ is *least favourable* if for every prior λ', we have

$$R_\lambda(\delta_\lambda) \geq R_{\lambda'}(\delta_{\lambda'})$$

where δ_λ and $\delta_{\lambda'}$ are Bayes estimators with respect to the priors λ and λ'.

Theorem 1.29. *Let λ be a prior on Θ such that $R_\lambda(\delta_\lambda) = \sup_{\theta\in\Theta} R(\delta_\lambda, \theta)$, where δ_λ is a λ-Bayes rule. Then*

1. *δ_λ is minimax*
2. *If δ_λ is the unique λ-Bayes rule, then it is the unique minimax rule*
3. *The prior λ is least favourable*

Corollary 1.30. *If a (unique) Bayes rule δ_λ has constant risk in θ, then it is the (unique) minimax rule.*

Proof. If a Bayes rule δ_λ has constant risk, then

$$R_\lambda(\delta_\lambda) = \mathbb{E}_\lambda[R(\delta_\lambda, \theta)] = \sup_{\theta\in\Theta} R(\delta_\lambda, \theta)$$

so the hypothesis of the previous theorem is satisfied. □

Proof of Theorem. We want to show that $\inf_\delta \sup_{\theta\in\Theta} R(\delta, \theta) = \sup_{\theta\in\Theta} R(\delta_\lambda, \theta)$

1. Let δ be any decision rule. Then

$$\sup_{\theta\in\Theta} R(\delta, \theta) \geq \mathbb{E}_\lambda[R(\delta, \theta)] \geq R_\lambda(\delta_\lambda) = \sup_{\theta\in\Theta} R(\delta_\lambda, \theta) \tag{$*$}$$

Thus $\inf_\delta \sup_{\theta\in\Theta} R(\delta, \theta) \geq \sup_{\theta\in\Theta} R(\delta_\lambda, \theta)$. Hence $\inf_\delta \sup_{\theta\in\Theta} R(\delta, \theta) = \sup_{\theta\in\Theta} R(\delta_\lambda, \theta)$, so δ_λ is minimax.

2. If δ_λ is the unique λ-Bayes rule, then the 2nd inequality in $(*)$ is strict for any $\lambda' \neq \delta_\lambda$, so $\sup_{\theta\in\Theta} R(\delta', \theta) > \sup_{\theta\in\Theta} R(\delta_\lambda, \theta)$ and δ' is not minimax.

3. For any prior λ', we have

$$R_{\lambda'}(\delta_{\lambda'}) = \mathbb{E}_{\lambda'}[R(\delta_{\lambda'}, \theta)] \leq \mathbb{E}_{\lambda'}[R(\delta_\lambda, \theta)] \leq \sup_{\theta\in\Theta} R(\delta_\lambda, \theta) = \mathbb{E}_\lambda[R(\delta_\lambda, \theta)]$$

So δ_λ is least favourable.

□

Example. In a $\mathrm{Bin}(n, \theta)$ model, let $\pi_{a,b}$ be a $\mathrm{Beta}(a, b)$ prior on $\theta \in [0, 1]$. It can be shown that the unique Bayes rule for $\pi_{a,b}$ over the quadratic risk is the posterior mean $\delta_{a,b} = \bar{\theta}_{a,b}$. If we solve the equation $R(\delta_{a,b}, \theta) =$ constant over all $\theta \in [0, 1]$, we can find a prior π_{a^*,b^*} and corresponding Bayes rule δ_{a^*,b^*} of constant risk. Therefore it is minimax (see Example Sheet).

In this example, the minimax rule is unique and different from the MLE. However, in the next section we will show that in a $\mathcal{N}(\theta, 1)$ model, the MLE $\bar{X}_n$ is minimax.

Admissibility

Definition. A decision rule δ is *inadmissible* if there exists δ' such that

$$R(\delta', \theta) \le R(\delta, \theta),\ \forall \theta \in \Theta$$

and $R(\delta', \theta) < R(\delta, \theta)$ for some $\theta \in \Theta$. In this case, δ' is said to *dominate* δ.

Theorem 1.31.

1. *A unique Bayes rule is admissible*
2. *If δ is admissible and has constant risk, then it is minimax*

Proof. Example Sheet. □

We now study a case where an estimator which is not Bayes with respect to any prior can be proved to be admissible.

Theorem 1.32. *Let $X_1, \dots, X_n$ be iid samples from a $\mathcal{N}(\theta, \sigma^2)$ model, where σ^2 is known and $\theta \in \Theta = \mathbb{R}$. Then $\hat{\theta}_{MLE} = \bar{X}_n$ is admissible and minimax for estimating θ in quadratic risk.*

Proof. For simplicity, let $\sigma^2 = 1$, the general case is analogous. Note that the MLE has constant risk:

$$R(\hat{\theta}_{\mathrm{MLE}}, \theta) = \mathbb{E}_\theta[(\bar{X}_n - \theta)^2] = \mathrm{Var}_\theta(\bar{X}_n) = \frac{1}{n}$$

Thus by the previous theorem, it remains to show that $\bar{X}_n$ is admissible.

For any decision rule δ we have

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

("Bias-Variance decomposition of MSE"). Let $B(\theta) = \mathbb{E}_\theta[\delta(X)] - \theta$ denote the bias. Recall the general Cramer-Rao lower bound

$$\mathrm{Var}_\theta(\delta(X)) \ge \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta[\delta(X)]\right)^2}{nI(\theta)} = \frac{(1 + B'(\theta))^2}{n}$$

Hence if δ dominates $\bar{X}_n$¡ we have $R(\delta, \theta) \leq \frac{1}{n}$ for all $\theta \in \mathbb{R}$, hence

$$B(\theta)^2 + \frac{(1 + B'(\theta))^2}{n} \leq \frac{1}{n} \tag{$*$}$$

(Differentiability of B can be justified using regularity of the Gaussian model). From $(*)$, we can see that $B(\theta)$ is bounded above and below, and $B'(\theta) \leq 0$ for all θ, so B is (non-strictly) decreasing.

Now we can argue that $B(\theta) = 0$ for all $\theta \in \mathbb{R}$: we construct two sequences $\{\theta_n\}_{n \geq 1}$, one going to $+\infty$, and one going to $-\infty$, such that $B'(\theta_n) \to 0$. Otherwise, $B'(\theta)$ would be bounded away from 0 for θ small or large enough, making $B(\theta)$ unbounded. By $(*)$ we have $B(\theta_n) \to 0$ for both sequences. Since B is non-increasing, this means $B(\theta) = 0$ for all θ.

Thus, plugging into the usual Cramer-Rao bound gives $\mathrm{Var}_\theta(\delta(X)) \geq \frac{1}{n}$, implying that $R(\delta, \theta) \geq \frac{1}{n}$ for all θ. Thus $\bar{X}_n$ is admissible. □

Remark: the decision rule studied here is not Bayes for any prior. However, it is the limit of the Bayes rule δ_{ν^2} for priors $\mathcal{N}(0, \nu^2)$ when $\nu \to \infty$. It can be shown that all minimax rules are limits of Bayes rules (a result of Wald).

The result shown in the theorem can be extended to dimension $p = 2$, with the model $\mathcal{N}(\theta, I_2)$, where $\theta \in \Theta = \mathbb{R}^2$. However it is false for $p > 2$, leading to "Stein's paradox".

James-Stein phenomenon

Consider the model $X \sim \mathcal{N}(\theta, I_p)$ where $p \geq 3$. We will consider the case of a single observation for simplicity.

Definition. For a vector $X \in \mathbb{R}^p$, the *James-Stein estimator* is

$$\delta^{\mathrm{JS}}(X) = \left(1 - \frac{p-2}{||X||_2^2}\right) X$$

We will show that the James-Stein estimator dominates the MLE in this model.

First we calculate the risk of the MLE:

$$R(\hat{\theta}, \theta) = \mathbb{E}_\theta[||X - \theta||_2^2] = \sum_{j=1}^{p} \mathbb{E}[(X_j - \theta_j)^2] = p$$

We now explicitly calculate the risk of δ^{JS} and show that $R(\delta^{\mathrm{JS}}, \theta) < R(\hat{\theta}_{\mathrm{MLE}}, \theta)$ for all θ.

Lemma 1.33 (Stein's Lemma)**.** *Let $X \sim \mathcal{N}(\theta, 1)$ and let $g : \mathbb{R} \to \mathbb{R}$ be a bounded differentiable function such that $\mathbb{E}|g'(X)| < \infty$. Then $\mathbb{E}[(X-\theta)g(X)] = \mathbb{E}[g'(X)]$*

Proof. Write (using integration by parts)

$$\begin{aligned}
&\mathbb{E}[(X - \theta)g(X)] \\
&= \int_{\mathbb{R}} g(x)(x - \theta)\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \theta)^2}{2}\right) \mathrm{d}x \\
&= -\int_{\mathbb{R}} g(x) \left(\frac{\mathrm{d}}{\mathrm{d}x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \theta)^2}{2}\right)\right) \mathrm{d}x \\
&= -\left[g(x)\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \theta)^2}{2}\right)\right]_{-\infty}^{\infty} + \int_{\mathbb{R}} g'(x)\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \theta)^2}{2}\right) \mathrm{d}x \\
&= \mathbb{E}[g'(X)]
\end{aligned}$$

Where in the final step, the first term vanishes since g is bounded. □

Remark: Stein's Lemma is the basis of a proof technique in probability theory called "Steins method". It turns out that X must be normally distributed if it satisfies the equality in Steins lemma for all such g. In fact, this is also "approximately" true, which can be used, e.g, to show that sums of dependent random variables converge to a normal distribution.

Theorem 1.34. *Let $X \sim \mathcal{N}(\theta, I_p)$, for $p \geq 3$. Then the risk of the James-Stein estimator satisfies $R(\delta^{JS}, \theta) < p$ for all $\theta \in \mathbb{R}^p$.*

Proof. We write

$$\begin{aligned}
R(\delta^{\mathrm{JS}}, \theta) &= \mathbb{E}_\theta[||\delta^{\mathrm{JS}} - \theta||_2^2] \\
&= \mathbb{E}_\theta\left[\left\|X - \theta - \frac{p-2}{||X||_2^2}X\right\|^2\right] \\
&= \mathbb{E}_\theta[||X - \theta||_2^2] + (p-2)^2\mathbb{E}_\theta\left[\left\|\frac{X}{||X||_2^2}\right\|^2\right] - 2(p-2)\mathbb{E}_\theta\left[\frac{X^T(X - \theta)}{||X||_2^2}\right] \\
&= p + (p-2)^2\mathbb{E}_\theta\left[\frac{1}{||X||_2^2}\right] - 2(p-2)\mathbb{E}_\theta\left[\frac{X^T(X - \theta)}{||X||_2^2}\right]
\end{aligned}$$

For the last term:

$$\mathbb{E}_\theta\left[\frac{X^T(X-\theta)}{||X||_2^2}\right] = \sum_{j=1}^p \mathbb{E}_\theta\left[\frac{X_j(X_j-\theta_j)}{||X||_2^2}\right] = \sum_{j=1}^p \mathbb{E}_\theta\left[\mathbb{E}_j\left[\left.\frac{X_j(X_j-\theta_j)}{X_j^2+\sum_{i\neq j}X_i^2}\right| X_{(-j)}\right]\right]$$

where we use the tower property of expectation and condition on

$$X_{(-j)} = (X_1,\dots,X_{j-1},X_{j+1},\dots,X_p)$$

Now we write

$$\mathbb{E}_j\left[\left.\frac{X_j(X_j-\theta_j)}{X_j^2+\sum_{i\neq j}X_i^2}\right| X_{(-j)}\right] = \mathbb{E}[(X_j-\theta_j)g_j(X_j)]$$

Where $X_j \sim \mathcal{N}(\theta_j,1)$ and $g_j(x) = \frac{x}{x^2+\sum_{i\neq j}X_i^2}$ is bounded as long as $\sum_{i\neq j}X_i^2 \neq 0$, which happens with probability 1. We can also check that the derivative of g_j is bounded, so $\mathbb{E}|g'(X)| < \infty$. Now Stein's lemma implies

$$\mathbb{E}[(X_j-\theta_j)g_j(X_j)] = \mathbb{E}[g_j'(X_j)] = \mathbb{E}_j\left[\left.\frac{X_j^2+\sum_{i\neq j}X_i^2-2X_j^2}{(X_j^2+\sum_{i\neq j}X_i^2)^2}\right| X_{(-j)}\right]$$

Thus

$$\begin{aligned}\mathbb{E}_\theta\left[\mathbb{E}_j\left[\left.\frac{X_j(X_j-\theta_j)}{X_j^2+\sum_{i\neq j}X_i^2}\right| X_{(-j)}\right]\right] &= \mathbb{E}_\theta\left[\mathbb{E}_j\left[\left.\frac{X_j^2+\sum_{i\neq j}X_i^2-2X_j^2}{(X_j^2+\sum_{i\neq j}X_i^2)^2}\right| X_{(-j)}\right]\right]\\ &= \mathbb{E}_\theta\left[\frac{1}{||X||_2^2}-\frac{2X_j^2}{||X||_2^4}\right]\end{aligned}$$

Summing from 1 to p gives

$$\mathbb{E}_\theta\left[\frac{X^T(X-\theta)}{||X||_2^2}\right] = \sum_{i=1}^p \mathbb{E}_\theta\left[\frac{1}{||X||_2^2}-\frac{2X_j^2}{||X||_2^4}\right] = (p-2)\mathbb{E}_\theta\left[\frac{1}{||X||_2^2}\right]$$

Putting everything together gives

$$\begin{aligned}R(\delta^{\mathrm{JS}},\theta) &= p+(p-2)^2\mathbb{E}_\theta\left[\frac{1}{||X||_2^2}\right]-2(p-2)^2\mathbb{E}_\theta\left[\frac{1}{||X||_2^2}\right]\\ &= p-(p-2)^2\mathbb{E}_\theta\left[\frac{1}{||X||_2^2}\right] < p\end{aligned}$$

□

Remarks:

1. One can lower-bound the term $\mathbb{E}_\theta\left[\frac{1}{||X||_2^2}\right]$ as follows: let ϕ be the pdf of the p-variate standard normal. Then

$$\begin{aligned}\mathbb{E}_\theta\left[\frac{1}{||X||_2^2}\right] &= \int_{\mathbb{R}} \frac{1}{||x||_2}\phi(x-\theta)\mathrm{d}x \\ &\geq \int_{c_1 \leq ||x||_2 \leq c_2} \frac{1}{||x||_2}\phi(x-\theta)\mathrm{d}x \\ &\geq \frac{1}{c_2^2}\mathbb{P}_\theta\left(||X||_2 \in [c_1, c_2]\right)\end{aligned}$$

   On the Example sheet, a similar type of argument can be used to upper bound $\mathbb{E}_\theta\left[\frac{1}{||X||_2^2}\right]$, which can be used to show that $R(\delta^{\mathrm{JS}}, \theta) \to p$ as $||\theta||_2 \to \infty$. This result then implies that even though $\hat{\theta}_{\mathrm{MLE}}$ is not admissible, $\hat{\theta}_{\mathrm{MLE}}$ and δ^{JS} have the same maximal risk.

2. The James-Stein estimator is also itself not admissible. It is dominated by

$$\delta^{\mathrm{JS}+}(X) = \left(1 - \frac{p-2}{||X||_2^2}\right)^+ X$$

   However, even $\delta^{\mathrm{JS}+}$ is inadmissible, since it can be shown that admissible estimators must satisfy a certain smoothness condition.

3. In practice, the MLE X can be much easier to work with, e.g, designing hypothesis tests or constructing confidence regions.

4. It might seem counterintuitive that one can design a better estimator in multiple dimensions by using other coordinates to estimate a certain coordinate. One explanation of Stein's paradox can be made via the "bias-variance tradeoff": MSE = Bias^2 + Var

$$\sum_{j=1}^{p} \mathbb{E}[(\hat{\theta}_j - \theta_j)^2] = \sum_{j=1}^{p}\left(\mathbb{E}[\hat{\theta}_j] - \theta_j\right)^2 + \sum_{j=1}^{p}\mathbb{E}\left[\left(\hat{\theta}_j - \mathbb{E}[\hat{\theta}_j]\right)^2\right]$$

   If we consider $\hat{\theta}_\lambda = \lambda X$, we can check that the terms are

$$\mathrm{MSE}(\hat{\theta}_\lambda) = (\lambda - 1)||\theta||_2^2 + \lambda^2 p$$

   If $\lambda = 1$, bias is 0 but variance is p. For large p, shrinking the estimator may decrease MSE.

Classification problems

A decision problem of great practical importance is classification. We observe joint observations (X, Y). If we observe a new X_i, we want to predict the correponding $Y_i \in \{0, 1\}$.

The joint distribution can be characterised in one of two ways:

1. First generate X according to P_X, then generate Y according to
$$\mathbb{P}(Y = 1|X = x) = \mathbb{E}(Y|X = x) = \eta(x)$$
2. First generate Y from $\{0, 1\}$, with probability distribution (π_0, π_1), then generate X according to
$$X|Y = 0 \sim f_0(x) \quad \text{and} \quad X|Y = 1 \sim f_1(x)$$
(In practice f_0 and f_1 are unknown and need to be estimated from data; but here we assume that f_0, f_1 are known)

Definition. A *classification rule* δ is a function $\delta : \mathcal{X} \to \{0, 1\}$. It is equivalently defined by a region $\mathcal{R} \subseteq \mathcal{X}$ such that

$$\delta(x) = \delta_{\mathcal{R}}(x) = \begin{cases} 1 & \text{if } x \in \mathcal{R} \\ 0 & \text{if } x \in \mathcal{R}^c \end{cases}$$

The probability of error of $\delta_{\mathcal{R}}$ is characterised by two quantities:

1. Probability of misclassifying $X \sim f_0$
$$\mathbb{P}(X \in \mathcal{R}|Y = 0) = \int_{\mathcal{R}} f_0(x)\mathrm{d}x = \mathbb{P}_0(X \in \mathcal{R})$$
2. Probability of misclassifying $X \sim f_1$
$$\mathbb{P}(X \in \mathcal{R}^c|Y = 1) = \int_{\mathcal{R}^c} f_1(x)\mathrm{d}x = \mathbb{P}_1(X \in \mathcal{R}^c)$$

Under the prior (π_0, π_1), the Bayes classification risk is

$$\begin{aligned} R_\pi(\delta_{\mathcal{R}}) &= R(\delta_{\mathcal{R}}, 0)\pi(0) + R(\delta_{\mathcal{R}}, 1)\pi(1) \\ &= \pi_0\mathbb{P}_0(X \in \mathcal{R}) + \pi_1\mathbb{P}_1(X \in \mathcal{R}^c) \end{aligned}$$

By the theorem from Lecture 14, we can find a Bayes rule by minimising the posterior risk. We therefore calculate the posterior distribution:

$$\begin{aligned} \Pi(Y = 0|X = x) &= \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} \\ \Pi(Y = 1|X = x) &= \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} \end{aligned}$$

The posterior risk is

$$1\{\delta(x)=1\}\Pi(Y=0|X=x)+1\{\delta(x)=0\}\Pi(Y=1|X=x)$$

A corresponding Bayes rule therefore takes

$$\delta(x)=\begin{cases}0 & \text{if } \frac{\pi_0 f_0(x)}{\pi_1 f_1(x)}>1\\ 1 & \text{otherwise}\end{cases}$$

Definition. For a prior (π_0,π_1), with $\pi_1\in(0,1)$, the *Bayes classifier* is given by $\delta_\pi=\delta_{\mathcal{R}}$, where

$$\delta_{\mathcal{R}}(x)=\begin{cases}1 & \text{if} x\in\mathcal{R}\\ 0 & \text{if } x\in\mathcal{R}^c\end{cases}$$

where $\mathcal{R}=\{x\in\mathcal{X}:\frac{\pi_1 f_1(x)}{\pi_0 f_0(x)}\geq 1\}$.

Theorem 1.35. *The Bayes classifier $\delta_{\mathcal{R}}$ is a decision rule that minimises the Bayes classification risk. If $\mathbb{P}\left(\frac{\pi_1 f_1(x)}{\pi_0 f_0(x)}=1\right)=0$, then the Bayes rule is unique.*

Proof. Let $\mathcal{J}\subseteq\mathcal{X}$ be a classification region. The classification error associated to the region is

$$\begin{aligned}\mathcal{R}_\pi(\delta_{\mathcal{J}})&=\pi_0\int_{\mathcal{J}}f_0(x)\mathrm{d}x+\pi_1\int_{\mathcal{J}^c}f_1(x)\mathrm{d}x\\ &=\int_{\mathcal{J}}(\pi_0 f_0(x)-\pi_1 f_1(x))\,\mathrm{d}x+\pi_1\underbrace{\int_{\mathcal{X}}f_1(x)\mathrm{d}x}_{=1}\end{aligned}$$

Which is minimised when the first integrand is non-positive, i.e $\mathcal{J}$ corresponds to the region $\{\pi_0 f_0\leq\pi_1 f_1\}$, or equivalently, $\delta_{\mathcal{J}}$ is the Bayes classification rule. It is the unique minimiser when the boundary has probability 0. □

Remarks:

1. Since a unique Bayes rule is admissible, the Bayes classifier is admissible if $\mathbb{P}\left(\frac{\pi_1 f_1(x)}{\pi_0 f_0(x)}=1\right)=0$.

2. We can also use the theorem to find a minimax classifier: for any $q\in(0,1)$, let δ_q be the associated Bayes classifier for the prior $(\pi_0,\pi_1)=(q,1-q)$, and let $\mathcal{R}_q$ denote the corresponding classification region. Then the risk consists of two values, the probabilities of error $\mathbb{P}(\mathcal{R}_q^c|1)$ and $\mathbb{P}(\mathcal{R}_q|0)$, so finding q such that

$$\mathbb{P}(\mathcal{R}_q^c|1)=\mathbb{P}(\mathcal{R}_q|0)=\text{ constant}$$

yields a (unique) Bayes rule that is also (the unique) minimax.

Example. Consider the case of two normal distributions:

$$X \sim f_0 = \mathcal{N}(\mu_0, \Sigma) \quad \text{or} \quad X \sim f_1 = \mathcal{N}(\mu_1, \Sigma)$$

where $\mu_0, \mu_1 \in \mathbb{R}^p$ and Σ is a $p \times p$ covariance matrix. One can show that any Bayes classifier depends on the data X only through the discriminant function $D(X) = X^T \Sigma^{-1}(\mu_1 - \mu_0)$, which is linear in X (see example sheet). This is called linear discriminant analysis. The more general case of unequal covariance matrices does not lead to a linear boundary, and is called quadratic discriminant analysis.

Multivariate analysis

Recall that for two real-valued random variables X and Y, their *covariance* is $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$, and the *correlation* is

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

Definition. Given observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, the *sample correlation coefficient* is

$$\hat{\rho}_{X,Y} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2}\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y}_n)}}$$

It follows from standard results (LLN and Slutsky's lemma) that the sample versions of $\mathrm{Var}(X), \mathrm{Var}(Y)$ and $\mathrm{Cov}(X,Y)$ are consistent, so $\hat{\rho}_{X,Y}$ is also a consistent estimator of the correlation.

Theorem 1.36.

1. *Suppose* $X = \begin{pmatrix} X^{(1)} \\ \vdots \\ X^{(p)} \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$, *with* $\mathrm{Cov}(X^{(i)}, X^{(j)}) = \Sigma_{ij}$ *and* Σ *is positive definite. Then* $\rho_{X^{(i)},X^{(j)}} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}}\sqrt{\Sigma_{jj}}}$. *Furthermore,* $X^{(i)}$ *and* $X^{(j)}$ *are independent if and only if* $\Sigma_{ij} = 0$.

2. *For any random vector* $X = \begin{pmatrix} X^{(1)} \\ \vdots \\ X^{(p)} \end{pmatrix}$, *the matrix* $\rho \in \mathbb{R}^{p \times p}$ *with entries* $\rho_{ij} = \rho_{X^{(i)},X^{(j)}}$ *has all entries in* $[-1, 1]$, *with diagonal entries equal to* 1, *and is positive semidefinite.*

Proof. Not given. □

Remark: if the model is $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$, then $\hat{\rho}_{X,Y}$ (which is the plug-in MLE, as in Lecture 8), is equal to the MLE $\hat{\rho}_{\mathrm{MLE}}$.

Theorem 1.37. *Under a* $\mathcal{N}(0, I_2)$ *model, the distribution of* $\hat{\rho}_{X,Y}$ *is given by the pdf*

$$f_{\hat{\rho}}(r) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-2}{2}\right)}(1-r^2)^{\frac{n-4}{2}}, \text{ for } -1 \leq r \leq 1$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\mathrm{d}t$ *is the gamma function.*

Proof. Not given. □

In particular, knowing the density allows us to construct hypothesis tests for whether Gaussian variables are uncorrelated (independent).

More complicated exact distributions can be derived under nonzero correlation, from which it is possible to test more general hypotheses and/or construct confidence intervals.

Note that as $n \to \infty$, one can also use the Delta method from Lecture 8 to establish asymptotic normality of $\hat{\rho}_{X,Y}$ (without necessarily assuming (X, Y) are jointly Gaussian). Thus, we can derive asymptotically valid hypothesis tests and confidence intervals.

Partial correlation

A basic fact about multivariate Gaussians is that conditioning on a subset of coordinates still gives a Gaussian distribution.

Theorem 1.38. *For a given random vector $X \in \mathbb{R}^p$ such that $X \sim \mathcal{N}(\mu, \Sigma)$, suppose we partition $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$, where $X^{(1)} \in \mathbb{R}^q$ and $X^{(2)} \in \mathbb{R}^{p-q}$, and denote the mean vector and covariance matrix by*

$$\mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Then the conditional distribution of $X^{(1)}|(X^{(2)} = x^{(2)})$ is Gaussian with mean $\mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\left(x^{(2)} - \mu^{(2)}\right)$ and covariance $\Sigma_{11|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Proof. Not given. □

Since the covariance matrix of $X^{(1)}|X^{(2)}$ does not depend on the value of $X^{(2)}$, we can make the following definition.

Definition. The *partial correlation* of $X^{(1)(i)}$ and $X^{(1)(j)}$, for $1 \leq i, j \leq q$, is given by

$$\rho_{ij|2} = \frac{(\Sigma_{11|2})_{ij}}{\sqrt{(\Sigma_{11|2})_{ii}}\sqrt{(\Sigma_{11|2})_{jj}}}$$

The quantity can be estimated consistently using the plug-in MLE

$$\hat{\rho}_{ij|2} = \frac{(\hat{\Sigma}_{11|2})_{ij}}{\sqrt{(\hat{\Sigma}_{11|2})_{ii}}\sqrt{(\hat{\Sigma}_{11|2})_{jj}}}$$

known as the *sample partial correlation coefficient*, where

$$\hat{\Sigma}_{11|2} = \hat{\Sigma}_{11} - \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{21}, \text{ and } \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)\left(X_i - \bar{X}_n\right)^T$$

One can also derive exact and asymptotic distributions of $\hat{\rho}_{ij|2}$.

Example. As a concrete example of when partial correlations might be used, here is an example from Hooker (1907) on yield of hay, spring rainfall, and temperature in an English area over 20 years.

Suppose we observer the following sample correlation matrix:

$$\begin{pmatrix} 1 & 0.8 & -0.4 \\ 0.8 & 1 & -0.6 \\ -0.4 & -0.6 & 1 \end{pmatrix}$$

Of a random variable $X = (X^{(1)}, X^{(2)}, X^{(3)})$ with coordinates representing yield, rainfall and temperature respectively. Then we see that

- Yield and rainfall are positively correlated
- Yield and temperature are negatively correlated
- Rainfall and temperature are negatively correlated

Based on these observations, one might wonder if the negative relationship between yield & temperature is because:

1. High temperature → low yield, or
2. High temperature is associated with low rainfall, hence with low yield

Indeed, computing the sample partial correlation coefficient (conditioning on rainfall), gives a value of 0.1, so the 2nd explanation is more likely, i.e both more rainfall and higher temperature tend to increase yield.

Principle component analysis (PCA)

PCA is a common method in data analysis/machine learning for reducing the dimension of a dataset, e.g, for visualisation. It does this by trying to "capture as much variance" as possible.
PCA Algorithm: start with a dataset $\{X_i\}_{i=1}^n$ with $X_i \in \mathbb{R}^p$.

- Construct the sample covariance matrix

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(X_i - \bar{X}_n)^T$$

- Compute an orthonormal basis of the top $k \leq p$ eigenvectors $\{\hat{v}_i\}_{i=1}^k$ of $\hat{\Sigma}$ corresponding to the top k eigenvalues of the matrix
- Project the n data points into the new coordinate space formed by $\{\hat{v}_i\}_{i=1}^k$.

Statistical interpretation

We are performing an estimation problem, with the goal of estimating leading eigenspaces of Σ. For a random vector $X \in \mathbb{R}^p$ such that $\mathbb{E}X = 0$ and $\mathrm{Cov}(X) = \mathbb{E}[XX^T] = \Sigma$, a standard result in linear algebra (SVD) implies that there is an orthogonal matrix V such that $\Sigma = V\Lambda V^T$, where

$$\Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix}$$

and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. This can be rewritten as $\Sigma = \sum_{i=1}^p \lambda_i v_i v_i^T$, where the v_i are the columns of V.

Further note that if we take an arbitrary unit vector $w = \sum_{i=1}^p \alpha_i v_i$, the variance of $w^T X$ can be computed as

$$\mathrm{Var}(w^T X) = \mathbb{E}[w^T XX^T w] = w^T \mathbb{E}[XX^T]w = w^T \Sigma w = \sum_{i=1}^{p} \alpha_i^2 \lambda_i$$

Thus, the variance is maximised when $\alpha_1 = 1$ and $\alpha_2 = \ldots = \alpha_p = 0$, so $w = v_1$. Similarly, we can show that for all $k > 1$, the vector v_k is a unit vector, orthogonal to $\{v_1, \ldots, v_{k-1}\}$, upon which the projection of X has maximal variance.
Remark: although X does not need to be Gaussian for any of this to be valid, in the Gaussian case, orthogonality of the v_i's implies that the coefficients

$\{v_i^T X\}_{i=1}^p$ are independent. If we define $U = V^T X$, the vector of coefficients of X in the basis of the v_i's, then

$$U = \begin{pmatrix} — & v_1^T & — \\ — & v_2^T & — \\ & \vdots & \\ — & v_p^T & — \end{pmatrix} X = \begin{pmatrix} — & v_1^T X & — \\ — & v_2^T X & — \\ & \vdots & \\ — & v_p^T X & — \end{pmatrix}$$

$$\text{Cov}(U) = \mathbb{E}[UU^T] = \mathbb{E}[V^T X X^T V] = V^T \mathbb{E}[XX^T]V = V^T \Sigma V = V^T V \Lambda V^T V = \Lambda$$

In the Gaussian case, when the λ_i's are unique, the eigenvectors $\{\hat{v}_i\}_{i=1}^p$ of $\hat{\Sigma}$ and the corresponding eigenvalues $\{\hat{\lambda}_i\}_{i=1}^p$ are maximum likelihood estimators of the corresponding eigenvectors and eigenvalues of Σ (this follows from the plug-in MLE theory).

Further statistical analysis of PCA involves deriving asymptotic distributions of the $\hat{\lambda}_i$'s and the $\hat{v}_i$'s. This can be useful in hypothesis testing, e.g, $H_0 : \lambda_i = 0$ for $i > k$.

Resampling and the bootstrap

An informal intuition from our study of statistics so far is that as $n \to \infty$, we can get more information about a distribution from random samples (e.g iid data). In resampling, we take a dataset and redraw (partial) samples on which we recompute a statistic.

Let $T_n = T(X_1, \dots, X_n)$ be an estimator of a parameter θ, with bias $B_n(\theta) = \mathbb{E}_\theta(T_n) - \theta$. If the estimator is biased, we first show a method for reducing its bias.

Definition. Let $T_{(-i)} = T(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ be the estimator with ith observation removed. The *jackknife bias estimator* is defined as

$$\hat{B}_n = (n-1)\left(\frac{1}{n}\sum_{i=1}^n T_{(-i)} - T_n\right)$$

and the *jackknife bias-corrected estimate* of θ is

$$\tilde{T}_{\text{JACK}} = T_n - \hat{B}_n$$

Theorem 1.39. *Suppose the bias function $B_n(\theta)$ can be approximated as*

$$B_n(\theta) = \frac{a}{n} + \frac{b}{n^2} + \mathcal{O}\left(\frac{1}{n^3}\right)$$

for some $a, b \in \mathbb{R}$. Then

$$\mathbb{E}[\tilde{T}_{JACK}] - \theta = \mathcal{O}\left(\frac{1}{n^2}\right)$$

Proof. Note that $\tilde{T}_{\text{JACK}} = nT_n - \frac{n-1}{n}\sum_{i=1}^n T_{(-i)}$. Thus

$$\begin{aligned}
&\mathbb{E}[\tilde{T}_{\text{JACK}}] \\
&= n(B_n(\theta) + \theta) - (n-1)(B_{n-1}(\theta) + \theta) \\
&= \theta + n\left(\frac{a}{n} + \frac{b}{n^2} + \mathcal{O}\left(\frac{1}{n^3}\right)\right) - (n-1)\left(\frac{a}{n-1} + \frac{b}{(n-1)^2} + \mathcal{O}\left(\frac{1}{(n-1)^3}\right)\right) \\
&= \theta + b\left(\frac{1}{n} - \frac{1}{n-1}\right) + \mathcal{O}\left(\frac{1}{n^2}\right) \\
&= \theta + \mathcal{O}\left(\frac{1}{n^2}\right)
\end{aligned}$$

□

Remark: the bias expansion in the theorem can be shown to hold in many cases of interest, e.g $T_n = g(\bar{X}_n)$, where g satisfies certain regularity assumptions. If we write $\mu = \mathbb{E}(X_1)$ and $\theta = g(\mu)$, we have

$$T_n - \theta = \nabla g(\mu)^T(\bar{X}_n - \mu) + \frac{1}{2}(\bar{X}_n - \mu)^T\nabla^2 g(\mu)(\bar{X}_n - \mu) + E_n$$

Taking an expectation, the first term is 0 and the second term is on the order of $\text{Var}(\bar{X}_n)$, which is $\frac{a}{n}$.

Example. Consider $X_i \sim^{\text{iid}} \mathcal{N}(\mu, 1)$, and the estimator $T_n = (\bar{X}_n)^2$ of $\theta = \mu^2$. We can compute the biases of T_n and $\tilde{T}_{\text{JACK}}$:

$$\mathbb{E}_\theta[T_n] - \mu^2 = \mathbb{E}_\theta[(\bar{X}_n)^2 - \mu^2] = \text{Var}_\theta(\bar{X}_n) = \frac{1}{n}$$

$$\begin{aligned}\mathbb{E}_\theta[\tilde{T}_{\text{JACK}}] - \mu^2 &= (\mathbb{E}_\theta[T_n] - \mu^2) - (n-1)\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n T_{(-i)} - T_n\right] \\ &= \frac{1}{n} - (n-1)\left(\mathbb{E}_\theta[T_{(-i)}] - \mathbb{E}_\theta[T_n]\right) \\ &= \frac{1}{n} - (n-1)\left(\frac{1}{n-1} + \mu^2 - \left(\frac{1}{n} + \mu^2\right)\right) \\ &= 0\end{aligned}$$

Bootstrap

We will now show how to use the bootstrap for inference. Previously, we showed how to construct confidence intervals based on the sample mean of iid observations:

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

using theory about the asymptotic distribution of $\bar{X}_n$. Also, we need to assume σ is known (or estimated). A completely different method uses the bootstrap.

Definition. For fixed observations $X_1, \ldots, X_n$, we define a discrete probability distribution

$$P_n = P_n(\cdot | X_1, \ldots, X_n)$$

that generates $\{X_{n,i}^n\}_{i=1}^n$, n independent copies of X_n^b with law

$$P_n(X_n^b = X_i) = \frac{1}{n}, \quad \text{for } 1 \leq i \leq n$$

In other words, we sample n values uniformly at random, independently, with replacement, from the X_i's. This is known as the *bootstrap.*

Note that the bootstrap is unbiased for the sample mean:

$$\mathbb{E}_{P_n}(X_n^b) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n$$

Thus, the *bootstrap sample mean* $\bar{X}_n^b = \frac{1}{n}\sum_{i=1}^n X_{n,i}^b$ "estimates" $\bar{X}_n$.

To build a confidence interval, we use variability of the bootstrap samples to estimate the variability of $\bar{X}_n$.

Definition. Let $R_n^b = R_n^b(X_1, \dots, X_n)$ satisfy

$$P_n\left(|\bar{X}_n^b - \bar{X}_n| \leq \frac{R_n^b}{\sqrt{n}}|X_1, \dots, X_n\right) = 1 - \alpha$$

THe *bootstrap confidence set* $\mathcal{C}_n^b$ is defined by

$$\mathcal{C}_n^b = \left\{\nu \in \mathbb{R} : |\nu - \bar{X}_n| \leq \frac{R_n^b}{\sqrt{n}}\right\}$$

The distribution of $\bar{X}_n^b$ is computed approximately from repeated bootstrap sampling.

Remark: note that $\mathcal{C}_n^b$ can be computed without estimating σ or knowing the asymptotic distribution of $\bar{X}_n$. Instead, for fixed $X_1, \dots, X_n$, the distribution P_n is known, and its quantiles can be determined approximately from simulations (or exactly).

Validity of the bootstrap

When will this procedure be valid? We know that $P_n(\bar{X}_n \in \mathcal{C}_n^b) = 1 - \alpha$, by definition. Our goal is to show that $\mathbb{P}(\mu \in \mathcal{C}_n^b) \to 1 - \alpha$ as $n \to \infty$, to show that it is a proper confidence interval.

Theorem 1.40. *Let $X_1, \dots, X_n$ be drawn iid from P with mean μ and finite variance σ^2. As $n \to \infty$, we have*

$$\sup_{t \in \mathbb{R}} |P_n(\sqrt{n}(\bar{X}_n^b - \bar{X}_n) \leq t|X_1, \dots, X_n) - \Phi(t)| \xrightarrow{a.s} 0 \tag{Kolmogorov Smirnov distance}$$

where Φ is the cdf of a $\mathcal{N}(0, \sigma^2)$ distribution.

Remark: as in the case of the Bernstein-von Mises theorem, this convergence result can be used to show that $\mathbb{P}(\mu \in \mathcal{C}_n^b) \to 1 - \alpha$ (Example sheet).

The following auxiliary result shows that convergence in distribution implies uniform convergence of cdf's:

Theorem 1.41. *Suppose $A_n \sim f_n$ with cdf's F_n, and $A \sim f$ with continuous cdf F. As $n \to \infty$, we have*

$$A_n \xrightarrow{d} A \implies \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \to 0$$

Proof. Since F is monotonic and continuous, for all k and $1 \leq i \leq k-1$, there exists $x_i \in \mathbb{R}$ such that $F(x_i) = \frac{i}{k}$, and $-\infty = x_0 < x_1 < \dots, x_k = \infty$. Hence, for all $x \in [x_{i-1}, x_i]$, with $1 \leq i \leq k$ we have

$$F_n(x) - F(x) \leq F_n(x_i) - F(x_{i-1}) = F_n(x_i) - F(x_i) + \frac{1}{k}$$

$$F_n(x) - F(x) \geq F_n(x_{i-1}) - F(x_i) = F_n(x_{i-1}) - F(x_{i-1}) - \frac{1}{k}$$

So

$$|F_n(x) - F(x)| \leq \frac{1}{k} + \max_{0 \leq i \leq k} |F_n(x_i) - F(x_i)|$$

Now fix $\varepsilon > 0$. Taking k large enough so $\frac{1}{k} < \frac{\varepsilon}{2}$ and taking n large enough so $|F_n(x_i) - F(x_i)| < \frac{\varepsilon}{2}$ for all $0 < i < k$, we see that

$$\sup_{x \in \mathbb{R}} \leq \frac{1}{k} + \frac{\varepsilon}{2} < \varepsilon$$

□

Definition. The sequence $(Z_{n,i} : 1 \leq i \leq n)_{n \geq 1}$ is a *triangular array* of iid random variables if for all $n \geq 1$, $\{Z_{n,1}, \ldots, Z_{n,n}\}$ are n iid draws from a distribution Q_n.

Theorem 1.42 (CLT for triangular arrays)**.** *Let $(Z_{n,i} : 1 \leq i \leq n)_{n \geq 1}$ be a triangular array of iid (in each row) random variables, all with finite variance, such that* $\mathrm{Var}(Z_{n,i}) = \sigma_n^2 \to \sigma^2$ *as $n \to \infty$. Suppose*

1. *For all $\delta > 0$, we have $nP_n(|Z_{n,1}| > \delta\sqrt{n}) \to 0$ as $n \to \infty$*
2. $\mathrm{Var}(Z_{n,1} 1\{|Z_{n,1}| \leq \sqrt{n}\}) \to \sigma^2$ *as $n \to \infty$*
3. $\sqrt{n}\mathbb{E}[Z_{n,1} 1\{|Z_{n,1}| > \sqrt{n}\}] \to 0$

Then

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

The conditions in the CLT for triangular arrays are provided for rigour and are ***non-examinable***.

Proof. Not given. □

Recall the theorem we wanted to prove:

Theorem. *Let $X_1, \ldots, X_n$ be drawn iid from P with mean μ and finite variance σ^2. As $n \to \infty$, we have*

$$\sup_{t \in \mathbb{R}} |P_n(\sqrt{n}(\bar{X}_n^b - \bar{X}_n) \leq t | X_1, \ldots, X_n) - \Phi(t)| \xrightarrow{a.s} 0$$

(Kolmogorov Smirnov distance)

where Φ is the cdf of a $\mathcal{N}(0, \sigma^2)$ distribution.

Proof. Under the distribution $P_n(\cdot | X_1, \ldots, X_n)$, with $Z_{n,i} = X_{n,i}^b$, we have

$$\mathbb{E}_n[Z_{n,i}] = \mathbb{E}_n[X_{n,i}^b] = \bar{X}_n$$

Note that the sequence $(X_{n,i}^b : 1 \leq i \leq n)_{n \geq 1}$ is a triangular array of iid variables, and

$$\mathrm{Var}(X_{n,i}^b) = \mathbb{E}_n[(X_{n,i}^b)^2] - (\mathbb{E}_n[X_{n,i}^b])^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 := \sigma_n^2$$

By the law of large numbers, we have $\sigma_n^2 \xrightarrow{\text{a.s}} \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = \sigma^2$. Thus for all $(X_1, \ldots, X_n)$ such that $\sigma_n^2 \to \sigma^2$ (which happens with probability 1), if in addition the conditions (1)-(3) of the above CLT for triangular arrays are satisfied, we have $\sqrt{n}(\bar{X}_n^b - \bar{X}_n) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. By the theorem from before, convergence in distribution implies uniform convergence of cdf's, giving the result.

Non-examinable: to verify the assumptions of the CLT for triangular arrays hold almost surely, the main step is to show that for $0 \leq p \leq 2$ and $\delta > 0$, we have

$$\frac{1}{n} \sum_{i=1}^n |X_i|^p 1\{|X_i| > \delta \sqrt{n}\} \xrightarrow{\text{a.s}} 0$$

□

Other variants/extensions

We have examined carefully the bootstrap for confidence intervals based on the sample mean. In more generality, this procedure is known as the *empirical bootstrap*: for a statistic $T_n = T_n(X_1, \ldots, X_n)$, consider the distribution of $T_n^b(X_{n,1}^b, \ldots, X_{n,n}^b)$. The bootstrap is "consistent" if

$$\sup_{t \in \mathbb{R}} |P_n(T_n^b \leq t | X_1, \ldots, X_n) - F(t)| \xrightarrow{\text{a,s}} 0$$

where F is an assumed limiting distribution of T_n. In a parametric model $\{P_\theta : \theta \in \Theta\}$, one can employ one of the following procedures for building a confidence interval for θ:

1. Special case of empirical bootstrap: let $\hat{\theta}_n$ be the MLE and let $\hat{\theta}_n^b$ be the MLE computed from bootstrap samples $(X_{n,1}^b, \ldots, X_{n,n}^b)$. Find values R_n such that $P_n(||\hat{\theta}_n^b - \hat{\theta}_n||_2 \leq \frac{R_n}{\sqrt{n}}|X_1, \ldots, X_n) = 1-\alpha$. It is possible to show that under appropriate regularity conditions, the bootstrap confidence set $\mathcal{C}_n^b = \{\nu \in \mathbb{R} : ||\nu - \hat{\theta}_n||_2 \leq \frac{R_n}{\sqrt{n}}\}$ satisfies $\mathbb{P}_{\theta_0}(\mu \in \mathcal{C}_n^b) \to 1-\alpha$. (Note that it is not necessary to estimate the Fisher information or know the asymptotic distribution of $\hat{\theta}_n$.) This is known as the *nonparametric bootstrap*.
2. In contrast, the *parametric bootstrap* refers to resampling from the distribution $P_{\hat{\theta}_n}$, using the MLE $\hat{\theta}_n$, and then using the same procedure as in (1) to construct confidence intervals.
3. One may also analyse an *m out of n bootstrap*, based on m resampled data points.

Monte Carlo methods

To apply many of the statistical techniques discussed in these lectures, one must be able to generate samples from a fixed, known distribution. For example, computing quantiles of a posterior distribution or bootstrap distribution.

In some cases, there are specific formulae one can apply; in other cases, one must resort to approximations via numerical simulations.

We will assume that it is possible to generate iid samples from a uniform $[0, 1]$ distribution.

Definition. A *pseudo-random uniform sample* is a collection of variables $U_1^*, \ldots, U_N^*$ such that for all $u_1, \ldots, u_N \in [0, 1]$, we have

$$\mathbb{P}(U_1^* \leq u_1, \ldots, U_N^* \leq u_n) \approx \mathbb{P}(U_1 \leq u_1, \ldots, U_N \leq u_N) = u_1 \ldots u_N$$

Where the "≈" is up to machine precision.

Theorem 1.43. *Let $U_1, \ldots, U_N \sim^{iid} U[0,1]$, and define $X_i = \sum_{j=1}^{n} x_j 1\{U_i \in (\frac{j-1}{n}, \frac{j}{n}]\}$, for all $1 \leq i \leq N$. Then the X_i's are iid uniform over the set $\{x_1, \ldots, x_n\}$.*

Proof. Independence of the X_i's follows directly from independence of the U_i's. Can check that

$$\mathbb{P}(X_i = x_j) = \mathbb{P}\left(U_i \in \left(\frac{j-1}{n}, \frac{j}{n}\right]\right) = \frac{1}{n}$$

□

The theorem shows how to sample from a discrete distribution, e.g, bootstrap resampling. To generate data from a continuous distribution, we can do the following:

Definition. Let F be the cdf of a distribution on $\mathbb{R}$. The *generalised inverse F^-* of F is defined for $u \in (0,1)$ as $F^-(u) = \inf\{x \in \mathbb{R} : u \leq F(x)\}$ (for continuous F, this is the functional inverse and gives quantiles of the distribution).

Theorem 1.44. *For any distribution P with cdf F and generalised inverse F^-, if U is uniform over $[0,1]$, we have*

$$X = F^-(U) \sim P$$

Proof. See example sheet. □

Remarks:

1. The usefulness of the theorem depends on if F^- can be computed, exactly or approximately.

2. For a uniform random sample $(U_1, \ldots, U_N)$, the theorem implies $(X_1, \ldots, X_N) = (F^-(U_1), \ldots, F^-(U_N))$ is an iid sample from P. This can be used to approximate integrals or expectations; as a consequence of the law of large numbers,
$$\frac{1}{N}\sum_{i=1}^{N} g(X_i) \xrightarrow{\text{a.s}} \mathbb{E}_P[g(X)].$$

We now explore some alternatives.

Importance sampling

Let P have density f, from which it is hard to simulate samples, and let h be a density, whose support includes that of f, from which it is easier to simulate samples.

Observe that

$$\mathbb{E}_h\left[\frac{g(X)}{h(X)}f(X)\right] = \int_{\mathcal{X}} \frac{g(x)}{h(x)} f(x)h(x)\mathrm{d}x = \int_{\mathcal{X}} g(x)f(x)\mathrm{d}x = \mathbb{E}_f[g(X)]$$

As a consequence, for $(X_1, \dots, X_n)$ generated from the distribution with density h, we have

$$\frac{1}{N}\sum_{i=1}^{N} \frac{g(X_i)}{h(X_i)} f(X_i) \xrightarrow{\text{a.s}} \mathbb{E}_f[g(X)].$$

The usefulness of this result depends on the rate of convergence, which depends on the functions involved.

Accept/reject Algorithm

In a similar setup, with densities f and h satisfying $f \leq Mh$ for some $M > 0$, consider the following algorithm:

1. Generate $X \sim h$ and $U \sim U(0,1)$.
2. If $U \leq \frac{f(X)}{Mh(X)}$, take $Y = X$. Otherwise, return to step 1.

One can show that Y has a distribution with density f, by calculating cdfs (Example sheet). However, the time it takes to generate a sample is random, depending on f, h and M.

Gibbs sampler

When dealing with joint distributions, the Gibbs sampler can be used to generate approximate samples. In the bivariate case (X, Y), one starts with some $X_0 = x$ and applies the following sampling steps for $t \geq 1$:

1. Calculate $Y_t \sim f_{Y|X}(\cdot|X = X_{t-1})$.
2. Calculate $X_t \sim f_{X|Y}(\cdot|Y = Y_t)$.

This algorithm generates a sequence $\{(X_t, Y_t) : t \geq 1\}$, assuming it is easy to sample from the conditional distributions. One can show that the sequences $\{(X_t, Y_t)\}_{t\geq 1}, \{X_t\}_{t\geq 1}$ and $\{Y_t\}_{t\geq 1}$ are Markov processes with invariant distributions $f_{X,Y}, f_X$ and f_Y respectively. As a consequence of the Ergodic theorem (see IB Markov Chains), one can show that as $N \to \infty$

$$\frac{1}{N}\sum_{t=1}^{N} g(X_t, Y_t) \xrightarrow{\text{a.s}} \mathbb{E}[g(X, Y)]$$

This method can be generalised to larger numbers of variables, e.g, for a multivariate vector $(X_1, \dots, X_d)$, by cycling through (blocks of) variables.

Discrete-time Markov chains & invariant measures

We now show how to generate samples $X_1, \dots, X_M$ from a Markov chain taking values in $\mathcal{X} \subseteq \mathbb{R}^p$, with prescribed invariant measure μ (this somewhat generalises the Gibbs sampler construction).

Recall that the distribution of a Markov chain is described by an initial condition X_0 and transition probabilities

$$\mathbb{P}(X_m \in B | X_{m-1} = t), \ t \in \mathcal{X} \text{ and } m \in \mathbb{N}$$

where B is any measurable subset of $\mathcal{X}$. By the Markov property, the preceding transition probabilities are the same for all m. We will assume that they are described by a *Markov kernel* $K(t, \cdot)$, defining a probability distribution on $\mathcal{X}$ for each $t \in \mathcal{X}$ fixed:

$$\mathbb{P}(X_1 \in B | X_0 = t) = K(t, B), \ \forall t \in \mathcal{X} \text{ and } B \subseteq \mathcal{X} \text{ measurable.}$$

(Later we will encounter a case where K is a mixture of a discrete and continuous probability distribution.)

Definition. A pdf μ on $\mathcal{X}$ is *invariant* or *stationary* for K if

$$\int_{\mathcal{X}} \mathbb{P}(X_1 \in B | X_0 = t) \mu(t) \mathrm{d}t = \int_{\mathcal{X}} K(t, B) \mu(t) \mathrm{d}t = \mu(B) \ \forall B \subseteq \mathcal{X} \text{ measurable.}$$

Results in Ergodic theory (see Part II Probability & Measure) imply that under certain conditions on the Markov chain, the distribution of X_M converges to the unique invariant measure μ. The idea will be to use samples $\{X_m\}$ from the chain to approximate μ.

Metropolis-Hastings algorithm

This is one of the most popular Markov Chain Monte Carlo (MCMC) algorithms.

Consider a situation where the ratio $\frac{\mu(s)}{q(s|t)}$ can be evaluated for some conditional pdf $q(\cdot|t)$ on $\mathcal{X}$ from which we can generate samples.

Then we generate a Markov chain $\{X_m : m \in \mathbb{N}\}$ as follows:

1. For $m \in \mathbb{N}$ and given X_m, generate a new draw $s_m \sim q(\cdot|X_m)$.
2. Define
$$X_{m+1} = \begin{cases} s_m & \text{with probability } \rho(X_m, s_m) \\ X_m & \text{with probability } 1 - \rho(X_m, s_m) \end{cases}$$
where $\rho(t,s) = \min\left\{\frac{\mu(s)}{\mu(t)}\frac{q(t|s)}{q(s|t)}, 1\right\}$.

Theorem 1.45. *Let the Markov chain $\{X_m : m \in \mathbb{N}\}$ be generated as above, and suppose μ and $q(\cdot|t)$ are strictly positive throughout $\mathcal{X}$ for each $t \in \mathcal{X}$. Then μ is an invariant measure for the Markov chain.*

Proof. The transition kernel K of the chain has "density"

$$k(t,s) = \rho(t,s)q(s|t) + (1 - r(t))\mathrm{d}\delta_t(s), \ \forall s \in \mathcal{X}$$

where $\delta_t(s)$ is the point mass probability measure at $t \in \mathcal{X}$ (i.e $\delta_t(A) = 1$ if $1 \in A$, $\delta_t(A) = 0$ otherwise), and $r(t) = \int_{\mathcal{X}} \rho(t,\tau)q(\tau|t)\mathrm{d}\tau$, i.e the probability of accepting the sample s_m. In other words, $K(t,B) = \int_B k(t,s)\mathrm{d}s$ for all $t \in \mathcal{X}$ and $B \subseteq \mathcal{X}$ measurable.

Now note that

$$\begin{aligned} \rho(t,s)q(s|t)\mu(t) &= \min\{q(t|s)\mu(s), \mu(t)q(s|t)\} \\ &= \min\left\{q(t|s)\mu(s), \frac{\mu(t)}{\mu(s)}\frac{q(s|t)}{q(t|s)}q(t|s)\mu(s)\right\} \\ &= \rho(s,t)q(t|s)\mu(s) \end{aligned}$$

Now interchanging the order of integration (see Fubini's theorem from Probability & Measure):

$$
\begin{aligned}
&\int_{\mathcal{X}} \mathbb{P}(X_1 \in B|X_0 = t)\mu(t)\mathrm{d}t \\
&= \int_{\mathcal{X}} \int_B k(s,t)\mu(t)\mathrm{d}s\mathrm{d}t \\
&= \int_{\mathcal{X}} \int_B \rho(t,s)q(s|t)\mu(t)\mathrm{d}s\mathrm{d}t + \int_{\mathcal{X}} \int_B (1 - r(t))\mathrm{d}\delta_t(s)\mu(t)\mathrm{d}s\mathrm{d}t \\
&= \int_B \int_{\mathcal{X}} \rho(s,t)q(t|s)\mu(s)\mathrm{d}t\mathrm{d}s + \int_{\mathcal{X}} 1\{t \in B\}(1 - r(t))\mu(t)\mathrm{d}t \\
&= \int_B (r(s) + (1 - r(s)))\mu(s)\mathrm{d}s \\
&= \mu(B)
\end{aligned}
$$

Implying that μ is invariant for the Markov chain. □

Remarks:

1. Conditions for convergence to μ, starting from any initial point, can be justified.
2. The choice of q affects the probability of acceptance/rate of convergence.
3. MCMC allows us to sample from all dimensions of a multivariate distribution simultaneously (unlike Gibbs sampling), but finding an appropriate proposal distribution $q(\cdot|t)$ might be hard in high dimensions.

The preconditional Crank-Nicolson (pCN) proposal

We now study a case of Metropolis-Hastings where the target distribution μ is the posterior of a general likelihood model with prior equal to a $\mathcal{N}(0,\Sigma)$ distribution on $\Theta = \mathbb{R}^p$, where Σ is a non-singular covariance matrix. More precisely, if $l(\theta) = \log L_n(\theta) = \log\left(\prod_{i=1}^n f(x_i,\theta)\right)$ is the log-likelihood function, we wish to sample from

$$\mu(\theta) = L_n(\theta)\pi(\theta) \propto \exp\left(l(\theta) - \frac{\theta^T\Sigma^{-1}\theta}{2}\right),\ \forall\theta \in \mathbb{R}^p$$

(reweighting of a Gaussian reference measure).

A simple algorithm to consider is the *Gaussian random walk* method where we take $q(\cdot|t) \sim \mathcal{N}(t,\sigma^2 I)$, for a parameter $\sigma > 0$. For σ too small, the state space is not explored quickly. For σ too big, the acceptance probability is small. In both cases, the convergence is slow. For large dimensions, this method is inefficient: for any fixed $\sigma > 0$, the acceptance probability will converge to 0 as $p \to \infty$.

The alternative pCN algorithm uses the proposal distribution $q(\cdot|t) \sim \mathcal{N}(t\sqrt{1-2\delta}, 2\delta\Sigma)$ on $\Theta = \mathbb{R}^p$, $t \in \Theta$, whee $\delta > 0$ is some stepsize. The formula is motivated by a discretised solution to a stochastic PDE (Crank-Nicolson is a finite-difference method for solving PDEs).

The point is that the convergence rate of pCN to the stationary distribution does not depend on p. To simplify algebra, consider $\Sigma = I_p$, so the proposal densities are

$$q(s|t) \propto \exp\left(-\frac{1}{4\delta}\left\|s - t\sqrt{1-2\delta}\right\|_2^2\right), \ \forall s, t \in \mathbb{R}^p$$

So conditional sampling is just from a Gaussian. To compute acceptance probabilities, one can compute $\frac{\mu(s)}{\mu(t)}\frac{q(t|s)}{q(s|t)} = \exp(l(s) - l(t))$ (see Example sheet). Thus, the acceptance probability amounts to computing a likelihood ratio test (to compute $\rho(X_m, s_m)$) for each $m \in \mathbb{N}$.

Introduction to nonparametric statistics

So far, we have mostly looked at situations where data cames from a distribution P_θ, where $\theta \in \Theta$ (usually a subset of $\mathbb{R}^d$). Estimating the distribution amounts to estimating θ (e.g, MLE). However, it is also possible to estimate a distribution directly (via the cdf). Suppose $X_1, \dots, X_n$ are iid samples from some distribution with cdf F:

$$F(p) = \mathbb{P}(X \leq t) = \mathbb{E}[1_{(-\infty,t)}(X)]$$

Definition. The *empirical distribution function* F_n of a sample $X_1, \dots, X_n$ is given, for all $t \in \mathbb{R}$, by

$$F_n(t) = \frac{1}{n}\sum_{i=1}^n 1_{(-\infty,t)}(X_i) = \frac{|\{i : X_i \leq t\}|}{n}$$

The law of large numbers guarantees that for any $t \in \mathbb{R}$, we have $F_n(t) \xrightarrow{\text{a.s}} F(t)$. However, it is possible to prove uniform convergence.

Theorem 1.46 (Glivenko-Cantelli)**.** *For any cdf F, as $n \to \infty$, we have*

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow{a.s} 0$$

Proof. We will only consider the case when F is continuous. We can write

$$|F_n(t) - F(t)| = \left|\frac{1}{n}\sum_{i=1}^n 1_{(-\infty,t)}(X_i) - \mathbb{E}[1_{(-\infty,t)(X)}]\right|$$

Now we use the uniform law of large numbers (theorem 1.18) with $\mathcal{H} = \{1_{(-\infty,t)} : t \in \mathbb{R}\}$.

Since F is continuous, we can choose $\{t_i\}_{i=1}^N(\varepsilon)$ such that $t_0 = -\infty, t_{N(\varepsilon)} = +\infty$ and $F(t_{i+1}) - F(t_i) < \varepsilon$ by partitioning $(0,1)$ into $\frac{2}{\varepsilon}$ pieces. The brackets are then $\{1_{(-\infty,t_i]}(X), 1_{(-\infty,t_{i+1})(X)}\}_{i=0}^{N(\varepsilon)-1}$. It is easy to check these satisfy the conditions of theorem 1.18, since $\mathbb{E}[1_{(-\infty,t_{i+1})}(X) - 1_{(-\infty,t_i)}(X)] = F(t_{i+1}) - F(t_i) < \varepsilon$. So we have

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

as desired. □

Remark: in lecture 20, we showed that convergence in distribution implies uniform convergence. This result is different since the F_n's are themselves random.

It is possible to further refine the covergence statement about cdfs into a distributional result (like the central limit theorem).

Definition. A *Brownian motion* or *Wiener process* is a continuous process $\{W_t\}_{t\geq 0}$ with independent increments such that $W_0 = 0$ and $W_t - W_s \sim \mathcal{N}(0, t-s)$ for $s < t$.

Remark: a more formal description and proof of existence of this process is beyond the scope of this course. Informally, it can be thought of the limit of a random walk with independent steps, as the timestep tends to 0: $W_n(t) = \frac{1}{\sqrt{n}} \sum_{1\leq k\leq \lfloor nt \rfloor} \xi_k$, where the ξ_k's are standard normal random variables.

Definition. A *Brownian bridge* is a continuous process $\{B_t\}_{0\leq t\leq 1}$ equal to a Brownian motion conditioned on $B_1 = 0$. It satisfies $B_0 = B_1 = 0$; $B_t \sim \mathcal{N}(0, t(1-t))$ and $\mathrm{Cov}(B_s, B_t) = s(1-t)$ for $s \leq t$. A Brownian bridge can be constructed from a Brownian motion by taking $B_t = W_t - tW_1$.

Theorem 1.47 (Donsker-Kolmogorov-Doob)**.** *As $n \to \infty$, we have $\sqrt{n}(F_n - F) \xrightarrow{d} \mathfrak{G}_F$, where $\mathfrak{G}_F$ is a Gaussian process such that $\mathfrak{G}_F(t) = B_{F(t)}$.*

This process satisfies $\mathrm{Cov}(\mathfrak{G}_F(s), \mathfrak{G}_F(t)) = F(s)(1 - F(t))$.

Remark: the precise definition of convergence in distribution of random processes is beyond the scope of this course.

Theorem 1.48 (Kolmogorov-Smirnov)**.** *As $n \to \infty$, we have $\sqrt{n}||F_n - F||_\infty \xrightarrow{d} ||\mathfrak{G}_F||_\infty = \sup_{t\in[0,1]} |B_t|$.*

Remark: note that the limiting distribution does not depend on F. The limiting distribution is called the *Kolmogorov distribution.*

Important applications of the theorem:

1. Nonparametric hypothesis testing: test $H_0 : F = F_0$ vs $H_1 : F \neq F_0$. The statistic $\sqrt{n}||F_n - F_0||_\infty$ can be compares to quantiles of $||B||_\infty$.

2. Goodness of fit testing: if $\hat{\theta}_n$ is the MLE of a parametric model, one can take the cdf $F_{\hat{\theta}_n}$ and compute $\sqrt{n}||F_n - F_{\hat{\theta}_n}||_\infty$, which can be compared to quantiles of $||B||_\infty$.

3. Confidence bands for F (see Example sheet).