Principle of Statistics

November 15, 2017

Contents

0 Introduction

The idea of the course is to present some mathematical principles of statistics.

The objective of statistics can be described as below: in probability we have a known distribution and we wish to describe what's going to happen with random variables with this distribution. In contrast, in statistics we observe the behaviour of random variables with some unknown distributions and we try to recover the distribution.

Formally, for a real-valued random variable X on a probability space Ω, we define its distribution

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

When X is discrete,

$$F(t) = \sum_{x \leq t} f(x)$$

and f is called the *probability mass function* (*pmf*).

When X is continuous,

$$F(t) = \int_{-\infty}^{t} f(s)ds$$

and we call f the *probability density function* (*pdf*).

In many problems in statistics, we have a *sample* $X_1, X_2, ..., X_n$, i.e. n independent copies of the same random variable X. We call n the *sample size*, and write $X_1, ..., X_n$ *i.i.d.*.

Often, the distribution is known to belong to a certain class.

Definition. A *statistical model*, or a *parametric model*, is any family of pmf/pdf or probability distribution, indexed by a parameter: $\{f(\theta, \cdot) : \theta \in \Theta\}$ or $\{P_\theta : \theta \in \Theta\}$.

Example. (1) Consider $N(\theta, 1); \theta \in \Theta = \mathbb{R}$, the normal distribution;
(2) $N(\mu, \sigma^2)$; in this case, $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$. This example is to show that θ doesn't have to be just one real sigmamber;
(3) $Exp(\theta)$; $\theta \in \Theta = (0, \infty)$;
(4) $N(\theta, 1)$, $\theta \in \Theta = [-1, 1]$.

Definition. For a random variable X with distribution P, we say that the model

$$\{P_\theta : \theta \in \Theta\}$$

is *correctly specified* if there exists a $\theta \in \Theta$ s.t. $P_\theta = P$.

Remark. If $X \sim N(2, 1)$, then in the above example, (1) is correctly specified but (4) is not.

When the model is correctly specified, we will write θ_0 for the value s.t. $P = P_{\theta_0}$. $theta_0$ is usually the unknown.

Here are some task/problems in statistics:
1. Estimation: construct $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$, i.e. a function of the observation, such that when $X_i \sim P_{\theta_0}$, we have that $\hat{\theta}$ is close to θ_0.

2. Testing hypotheses: we want to determine whether we are under a sigmall hypothesis $H_0 : \theta = \theta_0$ or the alternative $H_1 : \theta \neq \theta_0$.

3. Inference: to find intervals or sets $\varphi_n = \varphi_n(X_1, ..., X_n)$ such that for some $0 < \alpha < 1$, we have

$$P_\theta(\theta \in \varphi_n) = 1 - \alpha$$

(sometimes $\geq$ or $\to 1 - \alpha$).

1 The Likelihood Principle

1.1 Basic ideas and concepts

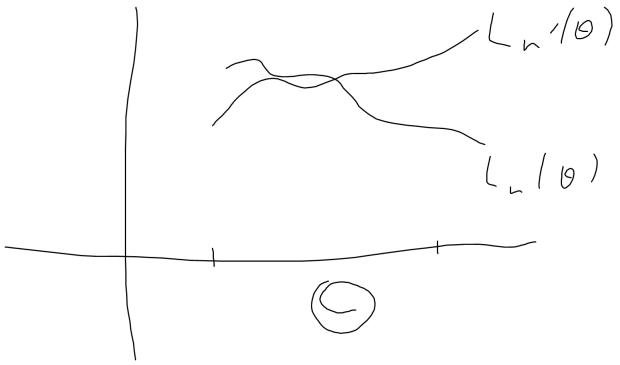
Let $X_1, ..., X_n$ be i.i.d. from $\{Poi(\theta) : \theta \geq 0\}$ with numerical values $X_i = x_i$ for $1 \leq i \leq n$.

The joint distribution is

$$\begin{aligned} f(x_1, ..., x_n; \theta) &= P_\theta(X_1 = x_1, ..., X_n = x_n) \\ &= \prod_{i=1}^{n} P_\theta(X_i = x_i) \\ &= \prod_{i=1}^{n} f(x_i, \theta) \\ &= \prod_{i=1}^{n} \left(e^{-\theta} \frac{\theta^{x_i}}{x_i!} \right) \\ &= e^{-n\theta} \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} \\ &= L_n(\theta) \end{aligned}$$

where $L_n(\theta)$ is the probability of this particular sample under P_θ. We will be interested in the value(s) of θ that maximize this likelihood.

Mathematical principle: it is helpful to think of $L_n(\theta)$ as a random function on Θ, with the randomness coming from the observations. Different observations could lead to different L_n.



To maximize $L_n(\theta)$, it is often more practical to maximize $l_n(\theta) = \log(L_n(\theta))$, which is an increasing function and more practical for calculus. We have

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

take the derivative, we have

$$l_n'(\theta) = 0 \iff -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i = 0$$

if $\sum_{i=1}^{n} x_i = 0$, $\hat{\theta} = 0$ directly. Otherwise we have

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

we can also check $l_n''(\theta)$ that this is indeed a maximum.

Definition. Let $\{f(\cdot, \theta) : \theta \in \Theta\}$ be a statistical model of pmf/pdf for the distribution P of n i.i.d. observations $X_1, ..., X_n$ of $X \sim P$ with realisations $x_i 1 \leq i \leq n$ (???).

The *likelihood function* L_n is

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

where f is the probability or density of the realization with iid assumption). The *log-likelihood function* is

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

and the *normalized log-likelihood function* is

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

Definition. The *maximum likelihood estimator(MLE)* is any element $\hat{\theta} \in \Theta$ such that

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

Remark. By definition of all these functions and of the estimators, we should see that it's equivalent to maximize any of the L_n, l_n or $\bar{l}_n$. In particular, we can use any of these in the definition of the maximum likelihood estimator. We can think of the MLE as a function of the observations, i.e. $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$.

The definitions above can be relaxed if the variables are not i.i.d but a joint distribution can be specified.

Example. • For $X_i \sim Poi(\theta)$, $\theta \geq 0$; $\hat{\theta} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
• For $X_i \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$, we've seen from previous courses (but it's also good to check again) that

$$l_n(\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

To maximize this, we write $\tau = \sigma^2$, then

$$\frac{\partial l_n}{\partial \mu}(\theta) = (-2) \times \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu) \right)$$
$$\frac{\partial l_n}{\partial \tau}(\theta) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (X_i - \mu)^2$$

We need $\nabla l_n = 0$, so the first equation gives $\hat{\mu} = \bar{X}_n$. Using this in the second equation we get

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

By checking the Hessian we could see that these indeed maximizes l_n.

In many examples, the MLE is found as a root of the gradient of the log-likelihood (as we've been used to in Part IB statistics). So for convenience we might as well give it a name.

Definition. For $\Theta \subseteq \mathbb{R}^p$, the *score function* is

$$S_n(\theta) = \nabla_\theta l_n(\theta) = \left[\frac{\partial}{\partial \theta_1} l_n(\theta), ..., \frac{\partial}{\partial \theta_p} l_n(\theta) \right].$$

Remark. One of the main uses of this function is that we have $S_n(\hat{\theta}) = 0$, under '*nice*' conditions (e.g. MLE is not achieved on the boundaries, the likelihood function is smooth, etc.).

It's important to note (again) that l_n and S_n are random functions of a variable θ, where the randomness comes from the observations $X_1, ..., X_n$ i.i.d. with $X_i \sim P_{\theta_0}$ (the true θ). Therefore it only makes sense to take derivative with respect to the variable θ but not the observations, because they are what determine the function l_n and S_n themselves and should really be considered *fixed*.

1.2 Information geometry and likelihood function

Definition. We recall that for a variable X with distribution P_θ on $\mathcal{X} \subseteq \mathbb{R}^d$ and $g : \mathcal{X} \subset \mathbb{R}$, we have

$$\mathbb{E}_\theta[g(X)] = \int_{\mathcal{X}} g(x) dP_\theta(x) = \int_{\mathcal{X}} g(x) f(x, \theta) dx$$

for the discrete case we just change from integrals to sums.

Theorem. For a model $\{f(\cdot, \theta), \theta \in \Theta\}$ and a variable $X \sim P$ such that $\mathbb{E}[|\log f(X, \theta)|] < \infty$, if the model is well-specified (i.e. $\theta_0 \in \Theta$) with $f(x, \theta_0)$ the pdf of P, the function l defined by

$$l(\theta) = \mathbb{E}_{\theta_0}[\log f(X, \theta)]$$

is maximized at θ_0. In other words, if we somehow had access to the function l (which is impossible in reality because we don't know θ_0), then we can maximize it and find the *true* value of θ_0. We usually do the next best thing and take instead a *finite sample approximation*

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

We use it as a proxy for $l(\theta)$. The idea is that if $\bar{l}_n(\theta)$ is a good approximation of $l(\theta)$, then by maximizing it, the MLE $\hat{\theta}$ we find would be a good approximation for the true value θ_0 as well.

Proof. For all $\theta \in \Theta$,

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

For a concave function φ, Jenson's inequality states that

$$\mathbb{E}[\varphi(Z)] \leq \varphi(\mathbb{E}[Z])$$

And we know log is concave. So

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

because $f(x, \theta)$ is a distribution. So we see that θ_0 maximizes l. □

If we satisfy *strict identifiability*, i.e. $f(\cdot, \theta) = f(\cdot, \theta') \iff \theta = \theta'$, then the strict version of Jenson's inequality holds and θ_0 is the unique value that maximizes l.

In information theory, we have the notion of divergence between distributions

$$KL(P_{\theta_0}, P_\theta) := \int_{\mathcal{X}} f(x, \theta_0) \log \frac{f(x,\theta_0)}{f(x,\theta)} dx$$

i.e. the expectation of the log under the *true* (θ_0) distribution.

It is equal to $l(\theta_0) - l(\theta)$, and can be thought of as a *distance* between the distributions.

So maximizing $l(\theta) = l(\theta_0) - KL(P_{\theta_0}, P_\theta)$ can be thought of as minimizing a distance to θ_0.

—-lecture 3—-

We had $\bar{l}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i, \theta)$, and $E_{\theta_0}[\bar{l}_n(\theta)] = l(\theta)$.

We saw that the MLE $\hat{\theta}$ is often solution to $S_n(\hat{\theta}) = \nabla_\theta \bar{l}_n(\hat{\theta}) = 0$. by exchanging sum and gradient, i.e. $\frac{1}{n}\sum_{i=1}^n \nabla_\theta \log f(X_i, \theta) = 0$.

Theorem. For a parametric model $\{f(\cdot, \theta) : \theta \in \Theta\}$ that is "regular enough", we have for all θ in the interior of Θ,

$$\mathbb{E}_\theta[\nabla \log f(X, \theta)] = 0$$

Note that the θ involved in the subscript of $\mathbb{E}$ and the argument of f are the same θ.

Proof.

$$\begin{aligned}
\mathbb{E}_\theta[\nabla \log f(X, \theta)] &= \int_{\mathcal{X}} (\nabla \log f(x, \theta)) f(x, \theta) dx \\
&= \int_{\mathcal{X}} (\nabla f(x, \theta) \cdot \frac{1}{f(x, \theta)}) f(x, \theta) dx \\
&= \int_{\mathcal{X}} \nabla f(x, \theta) dx \\
&= \nabla \int_{\mathcal{X}} f(x, \theta) dx \\
&= \nabla_\theta(1) = 0
\end{aligned}$$

Note that we interchanged ∇ and $\int_{\mathcal{X}}$ by assuming enough regularity. □

In particular, for the 'true' θ_0, we have

$$\mathbb{E}_{\theta_0}[\nabla_\theta \log f(X, \theta_0)] = 0$$

Definition. For a parameter space $\Theta \subseteq \mathbb{R}^p$, we define the Fisher information matrix as the $p \times p$ covariance matrix,

$$I(\theta) = \mathbb{E}_\theta[\nabla \log f(X, \theta) \cdot \nabla \log f(X, \theta)^T]$$

coefficient-wise,

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

Remark. In particular, if we are in the one-dimensional case, we have that

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

Theorem. With the same regularity assumptions, for all θ in the interior of Θ we have

$$I(\theta) = -\mathbb{E}[\nabla_\theta^2 \log f(X, \theta)]$$

Proof.

$$\begin{aligned}\nabla_\theta^2 \log f(x,\theta) &= \nabla_\theta \left(\frac{1}{f(x,\theta)} \cdot \nabla_\theta f(x,\theta) \right) \\ &= \frac{1}{f(x,\theta)} \nabla_\theta^2 f(x,\theta) - \frac{1}{f(x,\theta)^2} \cdot \nabla_\theta f(x,\theta) \cdot \nabla_\theta f(x,\theta)^T\end{aligned}$$

Now

$$\begin{aligned}-\mathbb{E}_\theta[\nabla_\theta^2 \log f(X,\theta)] &= -\int_{\mathcal{X}} \frac{1}{f(x,\theta)} \nabla_\theta^2 f(x,\theta) \cdot f(x,\theta) dx \\ &\quad + \mathbb{E}_\theta \left[\left(\frac{1}{f(X,\theta)} \cdot \nabla_\theta f(X,\theta) \right) \left(\frac{1}{f(X,\theta)} \cdot \nabla_\theta f(X,\theta) \right)^T \right] \\ &\quad + \mathbb{E}_\theta \left[\nabla_\theta \log f(X,\theta) \cdot \nabla_\theta \log f(X,\theta)^T \right] \\ &= I(\theta)\end{aligned}$$

As the first two terms are 0 by $\int_{\mathcal{X}} f(x,\theta)dx = 1$. □

Remark. In dimension $p = 1$, we have

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

We shall mention here that the "regularity assumptions" will be specified later although they can just be stated as here. It is not the main focus of the course, nor examinable.

Definition. For a random vector $X = (X_1, ..., X_n) \in \mathbb{R}^n$, the Fisher information is naturally defined as

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

Proposition. When $X = (X_1, ..., X_n)$ consists of n i.i.d. observations of a random variable from $\{f(x,\theta) : \theta \in \Theta\}$,

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

Proof. Using that

$$f(X_1, ..., X_n, \theta) = \prod_{i=1}^n f(X, \theta)$$

by independence, we have

$$\nabla_\theta \log f(X_1, ..., X_n, \theta) = \sum_{i=1}^n \nabla_\theta \log f(X_i, \theta)$$

So

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

For $i = j$, each term of the sum is equal to $I(\theta)$ and there are n of them. For $i \neq j$, $\mathbb{E}_\theta[\nabla_\theta \log f(X_i,\theta)] = 0$ (we've proved that before), so all those terms are 0. □

Theorem. (Cramer-Rao lower bound)
Let $\{f(\cdot,\theta) : \theta \in \Theta\}$ be a 'regular' parametric model with $p = 1$, $\Theta \subseteq \mathbb{R}$. Also let $\tilde{\theta} = \tilde{\theta}(X_1, ..., X_n)$ be an unbiased estimator of θ_0 where $X_i \sim P_{\theta_0}$ are i.i.d. For all θ_0 in interior of Θ, we have

$$Var_{\theta_0}(\tilde{\theta}) = \mathbb{E}_{\theta_0}[(\tilde{\theta} - \theta_0)^2] \geq \frac{1}{nI(\theta_0)}$$

Proof. For $Var_{\theta_0}[\tilde{\theta}] < \infty$, we first treat the case $n = 1$. We recall the Cauchy-Schwarz inequality: for Y, Z random variables,

$$Cov(Y,Z)^2 \leq Var(Y) \cdot Var(Z)$$

Now take $Y = \tilde{\theta}$, and

$$Z = \frac{d}{d\theta} \log f(X,\theta)$$

we have

$$Var_{\theta_0}(\tilde{\theta}) \geq \frac{Cov^2_{\theta_0}(\tilde{\theta}, Z)}{Var_{\theta_0}(Z)}$$

but we know $Var_{\theta_0}(Z) = I(\theta_0)$. We recall that $\mathbb{E}_{\theta_0}[Z] = 0$, so $Cov_{\theta_0}(\tilde{\theta}, Z) = \mathbb{E}_{\theta_0}[\tilde{\theta} \cdot Z]$ i.e.

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\tilde{\theta}\frac{d}{d\theta}\log f(X,\theta_0)\right] &= \int_{\mathcal{X}} \tilde{\theta}(x)\frac{1}{f(x,\theta_0}\frac{d}{d\theta}f(x,\theta_0) \cdot f(x,\theta_0)dx \\
&= \frac{d}{d\theta}\int_{\mathcal{X}} \tilde{\theta}(x) \cdot f(x,\theta_0)dx \\
&= \frac{d}{d\theta}\mathbb{E}_{\theta_0}[\tilde{\theta}] \\
&= 1
\end{aligned}$$

for $n = 1$.

□

Proposition. For a differentiable function $\Phi : \Theta \to \mathbb{R}$ and an unbiased estimator $\tilde{\Phi}$ of $\Phi(\theta)$, we have for all θ in the interior of Θ,

$$Var_{\theta_0}(\tilde{\Phi}) \geq \frac{1}{n}\nabla_\theta\Phi(\theta_0)^T I^{-1}(\theta_0)\nabla_\theta\Phi(\theta_0).$$

For example, consider $\Phi(\theta) = \alpha^T\theta = \sum_{i=1}^{p} \alpha_i\theta_i$, so $\nabla\Phi = \alpha$. So the lower bound is

$$Var_{\theta_0}(\tilde{\Phi}) \geq \frac{1}{n}\alpha^T I^{-1}(\theta_0)\alpha.$$

Another example: let $(X_1, X_2)^T \sim N(\theta, \Sigma)$ where Σ is a known covariance matrix. There are several cases:
• case 1: estimating θ_1 when θ_2 is known, the model is 1-dimensional and one must compute $I(\theta_1)$ (a scalar, one dimensional model);
• case 2: estimating θ_1 when θ_2 is unknown. We can consider $\Phi(\theta) = \theta_1$ when θ_2 is unknown: we can consider $\Phi(\theta) = \theta_1$ and must compute $I(\theta)$, which is a 2*2 matrix.
(Discussion: what if Σ is diagonal, i.e. we have independence between variables?)

2 Asymptotic theory for the MLE

Many estimates (in particular the MLE) are not unbiased, but reasonable estimators should satisfy

$$\mathbb{E}_\theta[\tilde{\theta}] \to \theta$$

as $n \to \infty$. A stronger statement is of the form

$$\tilde{\theta} \xrightarrow{(?)} \theta$$

in some sense (that we wish to define) when $n \to \infty$ when sampling from P_θ.

2.1 Stochastic convergence

Definition. (Convergence *a.s.* and *in probability*)
Let $(X_n)_{n\geq 0}$, X be random vectors in $\mathbb{R}^k$ defined in a probability space.
(i) We say X_n converges to X *almost surely (a.s.)*, written as $X_n \xrightarrow{a.s.} X$ if

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

(ii) We say X_n converges to X *in probability*, written as $X_n \xrightarrow{P} X$ if

$$\forall \varepsilon > 0 \ \mathbb{P}(||X_n - X|| > \varepsilon) \to 0.$$

Definition. (Convergence *in distribution*)
We say $(X_n)_{n\geq 0}$ converges to X *in distribution* if for all $t \in \mathbb{R}^k$,

$$\mathbb{P}(X_n \preceq t) \to \mathbb{P}(X \leq t)$$

where LHS means $\mathbb{P}(X_{(1)} \leq t_1, X_{(2)} \leq t_2, ..., X_{(k)} \leq t_k)$, i.e. $\leq$ in each component.

Usually we just have

$$\mathbb{P}(X_n \preceq t) = F(t).$$

We write this as $X_n \xrightarrow{d} X$.

Proposition. $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{d} X$, as $n \to \infty$.

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

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

respectively.

Proposition. (Slutsky's lemma)
Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, where c is deterministic and non-stochastic, i.e. $\mathbb{P}(C = c) = 1$, then:
(a) $Y_n \xrightarrow{P} c$; this only works because c is deterministic – the same doesn't apply

for X_n;
(b) $X_n + Y_n \xrightarrow{d} X + c$;
(c) $(k = 1)$ $X_n Y_n \xrightarrow{d} cX$, and $X_n / Y_n \xrightarrow{d} X/c$ if $c \neq 0$;
(d) If $(A_n)_{n\geq 0}$ are random matrices such that $(A_n)_{ij} \xrightarrow{P} A_{ij}$ where A is a similarly non-stochastic matrix, then $A_n X_n \xrightarrow{d} AX$.

Proposition. If $X_n \xrightarrow{d} X$ as $n \to \infty$, then $(X_n)_{n\geq 0}$ is bounded *in probaility* or $X_n = O_P(1)$, i.e.

$$\forall \varepsilon > 0 \ \exists M(\varepsilon) < \infty$$

s.t. for all $n \geq 0$, $\mathbb{P}(||X_n|| \geq M(\varepsilon)) < \varepsilon$.

2.2 Law of Large numbers and Central Limit Theorem

Proposition. (Weak law of large numbers)
Let $X_1, ..., X_n$ be i.i.d. copies of X with $Var(X) < \infty$. Then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}[X]$ as $n \to \infty$.

Proof. We consider the random variable $Z_n = \frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X])$,

$$\begin{aligned}\mathbb{P}(|\bar{X}_n - E[X]| > \varepsilon) &= \mathbb{P}(\left|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X])\right| > \varepsilon) \\ &\leq \frac{Var(Z_n)}{\varepsilon^2}\end{aligned}$$

by Chebyshev's inequality. Then

$$Var(Z_n) = Var(X)/n$$

by direct computation using independence,

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}[X]| > \varepsilon) \leq \frac{Var(X)}{\varepsilon^2} \cdot \frac{1}{n} \to 0$$

as $n \to \infty$. □

Proposition. (Strong law of large numbers)
Let $X_1, ..., X_n$ be i.i.d. copies of $X \sim P$ in $\mathbb{R}^k$, and $\mathbb{E}[||X||] < \infty$. Then as $n \to \infty$,

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

Theorem. (Central limit theorem)
Let $X_1, ..., X_n$ be i.i.d. copies of $X \sim P$ on $\mathbb{R}$ $(p = 1)$ and $Var(X) = \sigma^2 < \infty$. We have, as $n \to \infty$,

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

Definition. (Multivariate normal)
A random variable X on $\mathbb{R}^k$ has normal distribution with mean μ and covariance Σ, denoted by $N_k(\mu, \Sigma)$ (sometimes the dimension is omitted), if either
• its (joint) density is

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

or
• it is the unique random variable such that for all linear functions $\alpha^T X \sim N(\mu^t \alpha, \alpha^T \Sigma \alpha)$ (in dimension 1).

Note that the second definition works even if Σ is not positive definite.

Proposition. • For a $d \times k$ matrix A and $b \in \mathbb{R}^k$, $AX + b \sim N(A\mu + b, A\Sigma A^T)$;
• If $A_n \xrightarrow{P} A$ are random matrices and $X_n \xrightarrow{d} N(\mu, \Sigma)$, then

$$A_n X_n \xrightarrow{d} N(A\mu, A\Sigma A^T)$$

which is a consequence of Slutsky's lemma;
• If Σ is diagonal, all the coefficients $X_{(j)}$ are independent.

Theorem. (Multivariate CLT)
Let $X_1, ..., X_n$ be i.i.d. copies of a random variable $X \sim P$ on $\mathbb{R}^k$ with $Cov(X) = \Sigma$ and is positive definite. Then as $n \to \infty$ we have

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

Proof of this will be exactly the same as the univariate case.

Corollary. Under the condition of the theorem above,

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X] = O_P\left(\frac{1}{\sqrt{n}}\right)$$

i.e. the deviation of $\bar{X}_n$ around $\mathbb{E}[X]$ are 'of order' $\frac{1}{\sqrt{n}}$ 'in probability'.

In light of the results and of the Cramer-Rao lower bound, we can expect that for 'optimal estimation' $\tilde{\theta}_n$,

$$nCov_\theta(\tilde{\theta}_n) \to I^{-1}(\theta)$$

as $n \to \infty$, sampling from P_θ. In dimension one we have

$$nVar_\theta(\tilde{\theta}_n) \to I^{-1}(\theta)$$

and we can also expect results "of the type"

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$$

If we have a result of this type for an estimator, it is also useful for inference.

Example. (Confidence interval)
Let $X_1, ..., X_n$ be a sequence of i.i.d. copies of $X \sim P$, real-valued random variable, with unknown mean μ_0 and known variance σ^2. For any $\alpha \in (0,1)$, we 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 taken s.t. $\mathbb{P}(|Z| \leq z_\alpha) = 1 - \alpha$, for $Z \sim N(0,1)$.

To show that $\mathcal{C}_n$ is a good confidence region, we compute (note that μ_0 is fixed but $\mathcal{C}_n$ is random)

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

where $\tilde{X}_i = \frac{X_i - \mu_0}{\sigma}$, so $\mathbb{E}[\tilde{X}_i] = 0$ and $Var\tilde{X}_i = 1$. The above is then equal to

$$= \mathbb{P}\left(\sqrt{n} \left| \frac{1}{n} \sum_{i=1}^n (\tilde{X}_i - E[\tilde{X}_i]) \right| \leq z_\alpha \right)$$

by CLT, $\sqrt{n}\left(\frac{1}{n} \sum_{i=1}^n \tilde{X}_i - \mathbb{E}[\tilde{X}_i] \right) \xrightarrow{d} N(0,1)$. Then by the continuous mapping theorem, the absolute value of above converges in distribution to $|Z|$ wher $Z \sim N(0,1)$. Finally, by the definition of z_α, we know that the above probability converges to $1 - \alpha$ as $n \to \infty$.

Remark. Here, we assumed that the variance σ^2 was known.
Note that it can be substituted by an estimate of the variance as well (see example sheet).

Here, we have shown that an estimator based on $\bar{X}_n$ converges to the true value and has deviations that are asymptotically normal (by LLN and CLT respectively). In next lecture, we would show that the same results (under assumptions) using mainly the same theorem will apply to the MLE.

Definition. (Consistency)
Consider $X_1, ..., X_n$ i.i.d. from a model $\{P_\theta : \theta \in \Theta\}$. An estimator $\tilde{\theta}_n = \tilde{\theta}_n(X_1, ..., X_n)$ is *consistent* whenever $\tilde{\theta}_n \to \theta_0$ in probability, when sampling from P_{θ_0} ($X_i \sim P_{\theta_0}$).

Remark. We will simply write this as $\tilde{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$.

Assumptions: Let $\{f(\cdot, \theta) : \theta \in \Theta\}$ be a model of pdf/pmf on $\mathcal{X} \subseteq \mathbb{R}^d$ such that
• $f(x, \theta) > 0$ for all $x \in \mathcal{X}, \theta \in \Theta$;

• $\int_{\mathcal{X}} f(x,\theta)dx = 1$ for all $\theta \in \Theta$; in the discrete case, the sum is 1;
• The function $f(x,\cdot) : \theta \to f(x,\theta)$ is continuous for all $x \in \mathcal{X}$;
• The parameter space $\Theta \subseteq \mathbb{R}^p$ is compact;
• For all $\theta, \theta' \in \Theta$, we have $f(\cdot,\theta) = f(\cdot,\theta') \iff \theta = \theta'$, i.e. θ uniquely determines a distribution;
• $\mathbb{E}_\theta \sup_{\theta'} |\log f(X,\theta')| < \infty$.

These things can be stated as usual regular assumptions. The idea is that in the most models in the real world obey these assumptions.

Remark. Assumption 1,2,5,6 guarantee that l has a unique maximum at θ_0. With these hypothesis (particularly 6) guarantee that $l(\theta) = \mathbb{E}_\theta[\log f(X,\theta)]$ is continuous.
These subtleties are not examinable; just refer to "usual regularity assumptions".

Theorem. (Consistency of MLE)
Let $X_1, ..., X_n$ be i.i.d. from the mode $\{f(\cdot,\theta) : \theta \in \Theta\}$ satisfying the assumption above. Then the MLE exists, and any MLE is consistent.

Proof. Proof of existence: The mapping $\theta \to \bar{l}_n\theta = \frac{1}{n}\sum_{i=1}^n \log f(X_i,\theta)$ is continuous and defined on a compact set Θ. Therefore, a maximum $\hat{\theta}_{MLE} = \hat{\theta}_n$ exists (maximum value theorem, see Analysis II).
Note: The idea is that for all $\theta \in \Theta$, $\bar{l}_n(\theta)$ converges to $l(\theta)$ since

$$\frac{1}{n}\sum_{i=1}^n \log f(X_i,\theta) \xrightarrow{P_{\theta_0}} \mathbb{E}_{\theta_0}[\log f(X,\theta)]$$

by LLN. We need to have a stronger fact: as $n \to \infty$,

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

Proof of consistency: define $\Theta_\varepsilon = \{\theta \in \Theta : ||\theta - \theta_0|| \geq \varepsilon\}$ for $\varepsilon > 0$. Θ_ε is a closed subset of a compact set Θ, so is also compact. The function l is continuous, so it has a maximum value on Θ_ε, with $l(\theta_\varepsilon) = \sup_{\theta\in\Theta_\varepsilon} l(\theta) = c(\varepsilon) < l(\theta_0)$. As a consequence, there exists $\delta(\varepsilon) > 0$ such that

$$c(\varepsilon) + \delta(\varepsilon) < l(\theta_0) - \delta(\varepsilon)$$

By the triangle inequality, we have

$$\begin{aligned}\sup_{\theta\in\Theta_\varepsilon} \bar{l}_n(\theta) &= \sup_{\theta\in\Theta_\varepsilon} [\bar{l}_n(\theta) - l(\theta) + l(\theta)] \\ &\leq \underbrace{\sup_{\theta\in\Theta_\varepsilon} l(\theta)}_{c(\varepsilon)} + \underbrace{\sup_{\theta\in\Theta} |\bar{l}_n(\theta) - l(\theta)|}_{<\delta(\varepsilon) \text{ on } A_n(\varepsilon)}\end{aligned}$$

We consider the sequence of events $A_n(\varepsilon) = \{\sup_{\theta\in\Theta} |\bar{l}_n(\theta) - l(\theta)| < \delta(\varepsilon)\}$.

On these events,

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

We also have that $l(\theta_0) - \bar{l}_n(\theta_0) \leq \delta(\varepsilon)$, i.e. the function $\bar{l}_n$ has a greater value at θ_0 than anywhere on Θ_ε. As a consequence, we have that

$$\sup_{\theta \in \Theta_\varepsilon} \bar{l}_n(\theta) \leq \bar{l}_n(\theta_0)$$

on $A_n(\varepsilon)$. On $A_n(\varepsilon)$, $\hat{\theta}_n$ cannot lie in Θ_ε as this would lead to a contradiction

$$\bar{l}(\hat{\theta}_n) \leq \bar{l}_n(\theta_0)$$

We therefore have that $A_n(\varepsilon) \subseteq \{||\hat{\theta}_n - \theta_0|| < \varepsilon\}$. Since $\mathbb{P}(A_n(\varepsilon)) \to 1$ by the uniform law of large numbers, we have $\mathbb{P}(||\hat{\theta}_n - \theta_0|| < \varepsilon) \to 1$. □

Remark. This proof can be simplified if the likelihood function has additional properties, such as differentiablility. This can be applied to examples where Θ is not compact.

We digress a bit on the uniform law of large numbers. The point is to prove that $\sup_{\theta \in \Theta} |\bar{l}_n(\theta) - \underbrace{\mathbb{E}_{\theta_0}[\bar{l}_n(\theta)]}_{l(\theta)}| \xrightarrow{P} 0$.

We state an observation for the finite case: Let $X_1, ..., X_n$ be i.i.d. in $\mathcal{X} \subseteq \mathbb{R}^d$ and $h_j : \mathcal{X} \to \mathbb{R}$ for $1 \leq j \leq M$. As $n \to \infty$, we have

$$\max_{1 \leq j \leq M} \left| \frac{1}{n} \sum_{i=1}^{n} h_j(X_i) - \mathbb{E}[h_j(X)] \right| \xrightarrow{P} 0$$

Consistency of the MLE: $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$, as $n \to \infty$. One of the results in the proof:

$$\sup_{\theta \in \Theta} |\bar{l}_n(\theta) - \underbrace{\mathbb{E}_{\theta_0}[\bar{l}_n(\theta)]}_{l(\theta)}| \xrightarrow{a.s.} 0$$

Observation (LLN, from 1 to finite case): Let $X_1, ..., X_n$ i.i.d. in $\mathcal{X} \subseteq \mathbb{R}^d$ and $h : \mathcal{X} \to \mathbb{R}$ a function. The variables $h(X_i)$ are also *i.i.d.* in $\mathbb{R}$, so if $\mathbb{E}[|h(X)|] < \infty$,

$$\frac{1}{n} \sum_{i=1}^{n} \underbrace{h(X_i)}_{Y_i} - \mathbb{E}[\underbrace{h(x)}_{\mathbb{E}[Y]}] \xrightarrow{a.s.} 0$$

by LLN. Consider $h_1, ..., h_M$ a finite class of functions such that $\mathbb{E}[h_j(X)] < \infty$. On events A_j s.t. $\mathbb{P}(A_j^c) = 0$, for all $1 \leq j \leq M$, we have (for every $\omega \in A_j$),

$$\frac{1}{n} \sum_{i=1}^{n} h_j(X_i(\omega)) - \mathbb{E}[h_j(X(\omega))] \to 0$$

as $n \to \infty$. Hence $A = \bigcap_{j=1}^{M} A_j$.We have

$$\max_{1 \leq j \leq M} \left| \frac{1}{n} \sum_{i=1}^{n} h_j(X_i) - \mathbb{E}[h_j(X)] \right| \to 0$$

in the reals, with implicit ω for X (i.e. $X(\omega)$) as well.

Furthermore, we have $\mathbb{P}(A^c) - \mathbb{P}(\bigcup_{j=1}^M A_j^c) \leq \sum_{j=1}^M \mathbb{P}(A_j^c) = 0$ (finite union of events with probability 0).

$$\max_{1\leq j\leq M} \left| \frac{1}{n}\sum_{i=1}^n h_\theta(X_i) - \mathbb{E}[h_j(X)] \right| \xrightarrow{a.s.} 0$$

as $n \to \infty$.

The uniform law of large number therefore holds over a finite set. In order to extend this result to $\log f(X_i, \theta) = h_\theta(X_i)$ (j becomes θ), we have to use some properties of h_θ, and of Θ.

The main idea is to use continuous analogue to finiteness: compactness of Θ + continuity of the function at hand.

(maybe missing a bit here)

Theorem. (Uniform LLN) Let Θ be a compact set in $\mathbb{R}^p$ and $q : \mathcal{X} \times \Theta \to \mathbb{R}$ be continuous in θ for all $x \in \mathcal{X}$ such that $\mathbb{E}[\sup_{\theta\in\Theta} |q(X, \theta)|] < \infty$. Then, as $n \to \infty$, we have

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

Proof. Idea: since Θ is compact, there exists Θ' finite that covers Θ up to precision δ. The uniform LLN holds on Θ'. By continuity in θ of q, it therefore extends to the whole Θ. □

2.3 Asymptotic normality of the MLE

We have that $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$ using an extension of the LLN (just like in the special case). We can similarly obtain that $(\hat{\theta}_n - \theta_0) \cdot \sqrt{n}$ averages in distribution, $X_n = \frac{1}{n}\sum_{i=1}^n X_i$ using an extension of the CLT.

Assumptions:
0. All the assumptions from consistency;
1. The true θ_0 belongs to the interior of Θ;
2. There exists $U \subseteq \Theta$ open such that $\theta \in f(x, \theta)$ is twice differntiably continuous, for all $x \in \mathcal{X}$.
3. The $p \times p$ Fisher information matrix $I(\theta_0)$ is non singular... (more technicalities, see lecture notes).

Theorem. Let the statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$ satisfy all the assumptions above, and $\hat{\theta}_n$ be the MLE based on $X_1, ..., X_n$ i.i.d .from P_{θ_0} (with pdf/pmf $f(\cdot, \theta_0)$). We have, as $n \to \infty$,

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

Proof. By assumptions, and consistency of the MLE, $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$, so $\hat{\theta}_n$ is in the interior of Θ on events of probability $\to 1$. By regularity assumptions, this implies that $\nabla_\theta \bar{l}_n(\hat{\theta}_n) = 0$. By regularity of $\theta \to \nabla_\theta \bar{l}_n(\theta)$, we can apply the mean value theorem from each coordinate, between θ_0 and $\hat{\theta}_n$,

$$0 = \nabla_\theta \bar{l}_n(\hat{\theta}_n) = \nabla_\theta \bar{l}_n(\theta_0) + \underbrace{A_n}_{\approx \nabla^2_\theta \bar{l}_n(\theta_0)} \cdot(\hat{\theta}_n - \theta_0)$$

Assuming (for this lecture) that A_n converges in probability to $\mathbb{E}_{\theta_0} \nabla^2_\theta \bar{l}_n(\theta_0)] = -I(\theta_0)$, this yields that

$$\sqrt{n}(\hat{\theta_n} - \theta_0) = \underbrace{(-A_n^{-1})}_{\xrightarrow{P} I(\theta_0)^{-1}} \cdot \sqrt{n} \nabla_\theta \bar{l}_n(\theta_0)$$

By definition of $\bar{l}_n$, we have

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

As a consequence, by CLT we have

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

Applying Slutsky's to the formula above, we get

$$\sqrt{n}(\hat{\theta_n} - \theta_0) \xrightarrow{d} N(0, \underbrace{I^{-1}(\theta_0) I(\theta_0) I(\theta_0)^{-1}}_{A \Sigma A})$$

□

A result from asymptotic normality of the MLE is that, if $\hat{\theta}_n$ is the MLE, then

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

Definition. (Asymptotic efficiency)
In a parametric model $\{f(\cdot, \theta), \theta \in \Theta\}$, a consistent estimator $\hat{\theta}_n$ is called asymptotically efficient if $nVar_{\theta_0}(\hat{\theta}_n) \to I^{-1}(\theta_0)$ **(missing a few lines here)**

Remark. (about theorem/assumptions/definitions) • at the exposure of more complicated proofs, one can reduce the regularity assumptions in the theorem (eg: laplace distribution).
• Same regularity is required: example of the uniform distribution

$$f(x, \theta) = \frac{1}{\theta} 1_{[0,\theta)}(x)$$

the MLE is not asymptotically normal in this case.
• For θ at the boundary of the parameter space, the result might not hold. For

a model $N(0,1)$ with $\theta \in \Theta = [0,\infty)$, in the case where $\theta_0 = 0$(boundary),
• This result (asymptotic normal) reinforces the intuition given by the Cramer-Rao lower bound, that $nVar_{\theta_0}(\hat{\theta}_n) \sim I^{-1}(\theta_0)$ is optimal. The lower bound holds only for unbiased estimators. The Hodge estimator $\hat{\theta}_n$ defined as the modification of any estimator $\hat{\theta}_n$ on $\mathbb{R}$ with $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0,\sigma^2)$ by

$$\tilde{\theta}_n = \begin{cases} \hat{\theta_n} & |\hat{\theta}_n| > n^{-4} \\ 0 & otherwise \end{cases}$$

In order to 'solve' these paradoxes, we will consider different ways of comparing estimators, in particular, in terms of their worst performance over $\theta_0 \in \Theta$.

2.4 Plug-in MLE and Delta method

It is often practical to consider estimation problems with $\{f(\cdot,\theta) : \theta \in \Theta\}$ and $\Phi : \Theta \to \mathbb{R}^k$. (ex: in $Bin(n,p)$ with $\theta = p$ and $Var(X) = np(1-p) = \Phi(p)$).

Definition. For $\Theta = \Theta_1 \times \Theta_2$ with $\theta = (\theta_1,\theta_2)^T, \theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$, we define the *profile likelihood* for $\Phi(\theta) = \theta_1$, by

$$L^{(p)}(\theta_1) = \sup_{\theta_2 \in \Theta_2} L((\theta_1,\theta_2)^T)$$

Remark. We note that it is equivalent to maximize $L^{(p)}$ in θ_1 or to maximize in θ and to take the first coefficient.

More generally, we shows that the MLE in a new parametrization

$$\{f(\cdot,\phi) : \phi = \Phi(\theta) \text{ for some } \theta \in \Theta\}$$

is equal to $\Phi(\hat{\theta}_{MLE})$ (see example sheet).

Definition. For a statistical model $\{f(\cdot,\theta) : \theta \in \Theta\}$ and $\Phi : \Theta \to \mathbb{R}^k$, the plug-in MLE of $\Phi(\theta,0)$ is the estimator $\Phi(\hat{\theta}_{MLE})$.

Using the limiting distribution of an estimator and the regularity properties of Φ to derive properties of $\Phi(\hat{\theta}_{MLE})$ is known as the Delta method.

Theorem. (Delta method)
Let $\Phi : \Theta \to \mathbb{R}$ $(k = 1)$ with gradient satisfying $\nabla_\theta \Phi(\theta_0) \neq 0$. Let $\hat{\theta}_n$ be a sequence of continuously differentiable random variables (estimators) satisfying $\sqrt{n}(\hat{\theta}_n \cdot \theta_0) \xrightarrow{d} Z$, where Z is a random variable in $\mathbb{R}$. Then we have, as $n \to \infty$,

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

Proof. We have, by mean value theorem, for some $\tilde{\theta}_n$ in the segment $[\theta_0, \hat{\theta}_n]$,

$$\sqrt{n}(\Phi(\hat{\theta_n} - \Phi(\theta_0)) = \nabla_\theta \Phi(\tilde{\theta}_n) \cdot \sqrt{n}(\hat{\theta_n} - \theta_0)$$

Since $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z$, it is bounded in probability, and $||\hat{\theta}_n - \theta_0|| = O_p(\frac{1}{\sqrt{n}})$. As a consequence, $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$, so $\tilde{\theta}_n \xrightarrow{P_\theta} \theta_0$, by continuous mapping theorem, $\nabla_\theta \Phi(\tilde{\theta}_n) \xrightarrow{P_{\theta_0}} \nabla_\theta \Phi(\theta_0)$. Applying Slutsky's lemma we get the final result. □

Remark. We can generalize this result to other estimators with r_n instead of $\sqrt{n}$, provided that $r_n \to \infty$ (???).
For the MLE under regularity assumptions, this yields for the plug-in MLE,

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

(plug-in MLE is asymptotically efficient). If $\Theta \subseteq \mathbb{R}$, then

$$LHS \xrightarrow{d} N(0, \Phi'(\theta_0)^2 I^{-1}(\theta_0))$$

which is the 1-d version of the result.

2.5 Asymptotic inference with the MLE

In previous example about hte mean of a random variable, we were able to construct confidence region and to control $\mathbb{P}(\mu \in \mathcal{C}_n)$ as $n \to \infty$, using the CLT.

Generalization: Consider a statistical model with usual regularity assumption. If we are interested in confidence regions for $\theta_{0,j}$ (the j-th coefficient of θ_0), we can use

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

By continuous mapping theorem, where e_j is the j-th vector of the canonical basis.
Using the same logic as in the previous example, we take

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

for $\mathbb{P}(|Z| \leq z_\alpha) = 1 - \alpha$ when $Z \sim N(0,1)$.

We compute

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

the variable $\sqrt{n}(I^{-1}(\theta_0))_{jj}^{-1/2}(\hat{\theta}_{n,j} - \theta_{0,j}) \xrightarrow{d} Z \to \mathbb{P}(|Z| \leq z_\alpha) = 1 - \alpha$ by remark above (asymptotic normality).

Remark. To construct this region $\mathcal{C}_n$, one requires $I(\theta_0)$, which is unknown in general.

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

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

Remark. • This estimate allows us to have a proxy for the function $\theta \to I(\theta)$. It is defined as

$$\mathbb{E}_{\theta_0}\left[\nabla_\theta \log f(X,\theta) \cdot \nabla_\theta \log f(X,\theta)^T\right]$$

• We also need to know where to evaluate it, we commonly take $\hat{i}_n = i_n(\hat{\theta}_{MLE})$ as an estimator for $I(\theta_0)$.

Proposition. Under the usual regularity assumption, as $n \to \infty$,

$$\hat{i}_n \xrightarrow{P_{\theta_0}} I(\theta_0)$$

Proof. We have, noting $g(X,\theta) = \nabla_\theta \log f(X,\theta) \cdot \nabla_\theta \log f(X,\theta)^T$, that $i_n(\theta) = \frac{1}{n}\sum_{i=1}^n g(X_i,\theta)$, and $I(\theta) = \mathbb{E}_{\theta_0}[g(X,\theta)]$. We therefore have

$$\hat{i}_n - I(\theta_0) = \left[\underbrace{i_n(\hat{\theta}_{MLE}) - I(\hat{\theta}_{MLE})}_{approximating\ the\ function}\right] + \left[\underbrace{I(\hat{\theta}_{MLE}) - I(\theta_0)}_{approximating\ the\ evaluation\ point}\right]$$

• The first term satisfies

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

by uniform law of large numbers.

• The second term satisfies

$$\left|I(\hat{\theta}_{MLE}) - I(\theta_0)\right| \xrightarrow{P_{\theta_0}} 0$$

by consistency of $\hat{\theta}_{MLE}$ and continuity of $I(\cdot)$.
By triangle inequality, the result holds. □

Remark. It is also possible to use $j_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \nabla_\theta^2 \log f(X_i,\theta)$ and $\hat{j}_n = j_n(\hat{\theta}_{MLE})$. The same result will hold with essentially the same proof.

Definition. (The Wald Statistic)
For all $\theta \in \Theta$, the Wald statistic is defined as

$$W_n(\theta) = n(\hat{\theta}_{MLE} - \theta)^T \hat{i}_n (\hat{\theta}_{MLE} - \theta_0)$$

(we can really think of the n as the product of two $\sqrt{n}$, one from each bracket.) This is a quadratic form, with positive semidefinite $\hat{i}_n$. Its level sets (contours?) are ellipsoids.

Proposition. (Confidence ellipsoids)
Under the same regularity assumptions, the confidence region defined by $\mathcal{C}_n = \{\theta : W_n(\theta) \le \xi_\alpha\}$ for ξ_α satisfying $\mathbb{P}(|\chi_p^2| \le \xi_\alpha) = 1-\alpha$, is an α-level asymptotic confidence region.

Proof. We compute $\mathbb{P}(W_n(\theta_0) \leq \xi_\alpha) = \mathbb{P}(\theta_0 \in \mathcal{C}_n)$. We have

$$\begin{aligned} W_n(\theta_0) &= \sqrt{n}(\hat{\theta}_n - \theta_0)^T \cdot \hat{i}_n \cdot \sqrt{n}(\hat{\theta}_n - \theta_0) \\ &= \sqrt{n}(\hat{\theta}_n - \theta_0)^T I(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) + \sqrt{n}(\hat{\theta}_n - \theta_0)^T(\hat{i}_n - I(\theta_0)) \cdot \sqrt{n}(\hat{\theta}_n - \theta_0) \end{aligned}$$

Since $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$, the 'original' first term converges in distribution to $v^T v$, where $v \sim N(0, I_p) = v_1^2 + ... + v_p^2 \sim \chi_p^2$.

The second term is a product of $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$ and
$\underbrace{\sqrt{n}(\hat{\theta}_n - \theta_0)^T}_{\xrightarrow{d} Z} \underbrace{(\hat{i}_n - I(\theta_0))}_{\xrightarrow{P_{\theta_0}} 0}$.

Applying Slutsky's lemma gives the second term converges to 0 in distribution, the product of thes two terms converges to 0 in distribution , and hence the sum with the original first term converges to χ_p^2 in distribution. □

Remark. This statistic can therefore be used in hypothesis testing problems, to distinguish $H_0 : \theta = \theta_0 \in \Theta$ from $H_1 : \Theta \setminus \{\theta_0\}$. (?) Having $\mathbb{P}(W_n(\theta_0) > \xi_\alpha) \to \alpha$ controls the error.

Last lecture we see how $W_n(\theta_0)$ is useful in testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta \setminus \{\theta_0\}$.

Generalization to $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta \setminus \Theta_0$: we want to construct a *test* ψ_n, function of the data (observations) with values in $\{0, 1\}$. Its objective is to output 0 under H_0, and 1 under H_1.

Measure of performance of a test:
• Type I error (false positive): $\mathbb{P}_\theta(\psi_n = 1) = \mathbb{E}_\theta[\psi_n], \theta \in \Theta_0$;
• Type II error (false negative): $\mathbb{P}_\theta(\psi_n = 0) = \mathbb{E}_\theta[(1 - \psi_n)], \theta \in \Theta \setminus \Theta_0$.

Definition. (likelihood ratio)
We define the *likelihood ratio statistic* as

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

where $\hat{\theta}_{MLE,0}$ denotes the MLE on Θ_0.

Theorem. (Wilks Theorem)
Let $\{f(\cdot, \theta) : \theta \in \Theta\}$ be a model satisfying the usual regularity assumptions (everything's smooth, continuous, regular etc.), and a hypothesis testing problem value $\Theta_0 = \{\theta_0\}$ for some fixed θ_0 in the interior of Θ. We have, under H_0 ($\theta \geq \theta_0$), as $n \to \infty$,

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

(p=dimension – i.e. $\Theta \subseteq \mathbb{R}^p$)

Proof. Considering events of probability $\to 1$, where $\hat{\theta}_n$ is in the interior of Θ (here $\hat{\theta}_n = \hat{\theta}_{MLE}$). We have, by definition of this statistic

$$\begin{aligned}\Lambda_n(\Theta, \Theta_0) &= 2l_n(\hat{\theta}_n) - 2l_n(\theta_0) \\ &= (-2l_n(\theta_0)) - (-2l_n(\hat{\theta}_n)) \\ &= -2\nabla_\theta l_n(\hat{\theta}_n)^T(\theta_0 - \hat{\theta}_n) + \sqrt{n}(\theta_0 - \hat{\theta}_n)^T A_n \sqrt{n}(\theta_0 - \hat{\theta}_n);\end{aligned}$$

(seems there's some expansion here) where A_n is the matrix of second order derivatives, i.e.

$$(A_n)_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \bar{l}_n(\theta^{(j)})$$

with $\theta^{(j)} \in [\theta_0, \hat{\theta}_n]$, as in the proof of asymptotic normality of the MLE, similarly to consistency of $\hat{i}_n$ (to $I(\theta_0)$). By definition/property of MLE,

$$\nabla_\theta l_n(\hat{\theta}_n) = 0$$

and $A_n \xrightarrow{P_{\theta_0}} (-I(\theta_0))$ and $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$, we have, similar to the proof on the convergence of Wald's statistics, $\Lambda_n \xrightarrow{d} \chi^2_p$. □

Remark. • As a consequence $\psi_n = 1\{\Lambda(\Theta, \Theta_0) \geq \xi_\alpha\}$, we can control the probability of type 1 error. (Reminder: $\mathbb{P}(|\chi^2_p| \leq \xi_\alpha) = 1 - \alpha$).
• For $\Theta \subseteq \mathbb{R}^p$, if Θ_0 has dimension $p_0 < p$ under the same assumptions, the limit is $\chi^2_{p-p_0}$.

3 Bayesian Inference

For a given $\{f(\cdot,\theta) : \theta \in \Theta\}$, in many applications, it can be practical to consider that θ is a random variable with some known distributions.

Example. Consider a finite parameter space $\Theta = \{\theta_1, ..., \theta_k\}$, and possible hypothesis $H_i : \theta = \theta_i$, with prior beliefs $\pi_i = \mathbb{P}(H_i)$. If the true hypothesis is H_i, the distribution of the observation X is $f_i(x)$, i.e. $\mathbb{P}(X = x|\underbrace{H_i}_{\theta=\theta_i}) = f_i(x)$.

By Bayes rule, when observing $X = x$,

$$\begin{aligned}\mathbb{P}(H_i|X = x) &= \frac{\mathbb{P}(X = x, H_i)}{\mathbb{P}(X = x)} \\ &= \frac{\pi_i f_i(x)}{\sum_j \pi_j f_j(x)}\end{aligned}$$

We will consider H_i more likely than H_h if

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

If the π_i are all equal (i.e. no prior information), this is the usual likelihood ratio rule based on $f_i(x)/f_j(x)$. Otherwise, the priors are her to update these priors.(?)

Definition. For a sample space $\mathcal{X}$, and a parameter space Θ we can consider the product measure Q over $\mathcal{X} \times \Theta$ such that

$$Q(x, \theta) = f(x, \theta) \cdot \pi(\theta)$$

The distribution π is the prior distribution of θ. As expected, the distribution

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

since the integral is 1.

The *posterior distribution* is the law of θ given X,

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

Similarly we define $\pi(\theta|X_1, ..., X_n)$.

Remark. The posterior distribution (as a function of θ) is a renormalized and reweighted (by $\pi(\theta)$) version of the likelihood.

Note that the denominator doesn't depend on θ, and can usually be ignored in computation.

Example. Let $X|\theta \sim N(\theta, 1)$ with prior $\theta \sim N(0, 1)$. The posterior satisfies

$$\begin{aligned}\Pi(\theta|X_1, ..., X_n) &\propto \underbrace{e^{-\theta^2/2}}_{\propto\pi(\theta)} \cdot \prod_{i=1}^{n} \underbrace{\exp(-\frac{(X_i-\theta)^2}{2})}_{\propto f(X_i,\theta)} \\ &\propto \exp(-n\theta\bar{X}_n - \frac{n+1}{2}\theta^2) \\ &\propto \exp(-\frac{(\theta\sqrt{n+1} - n\bar{X}_n/\sqrt{n+1})^2}{2}) \\ &\propto \exp(-\frac{(\theta - n\bar{X}_n/(n+1))^2}{2/(n+1)})\end{aligned}$$

We recognize a normal distribution, with mean $\frac{1}{n+1}\sum_{i=1}^{n} X_i$, and variance $\frac{1}{n+1}$, i.e.

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

(recall $\theta \sim N(0, 1)$).

Definition. In a statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$, when the prior $\pi(\theta)$ and the posterior $\Pi(\theta|X)$ belong to the same class of distribution, it is called a conjugate prior.

Example. Normal prior + normal sampling gives normal posterior;
Beta prior + binomial sampling gives beta posterior;
Gamma prior + poisson sampling gives gamma posterior.

Note that the definition of the posterior doesn't require Π to integrate to 1, i.e. to be a proper distribution. This can be useful when taking a prior with 'a simple formula' over a 'complicated set'.

Indeed, in the natural case where $\int_\Theta \pi(\theta)d\theta < \infty$, we can renormalize the prior without affecting the posterior.

Note that it is not even required that $\pi(\theta)$ has a finite integral to define a posterior, only that $f(x, \theta)\pi(\theta)$ does is enough.

Definition. A prior nonnegative function over Θ with infinite integral is an improper prior.

This can be useful when a prior is not available (from model or prior belief), but we have to construct it. This allows us to put constant mass $\pi(\theta) = \text{const}$ over any parameter space.

For example, taking $\pi(\theta) = 1$ for $N(0, 1)$ model and $N(\theta^3, 1)$ model will give different results.

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

In a $N(\mu, \tau)$ model with unknown $\theta = (\mu, \tau)^T \in \mathbb{R} \times (0, \infty)$, the Fisher information is

$$\begin{bmatrix} 1/\tau & 0 \\ 0 & \frac{1}{2\tau^2} \end{bmatrix}$$

As a consequence, the Jeffreys prior is given by $\pi(\mu, \tau) \propto \frac{1}{\tau^{3/2}}$.

However, in this case, the prior is improper, but the posterior is well-defined. In particular, the posterior marginal distribution for μ is $N(\bar{X}_n, \tau/n)$.

3.1 Statistical inference with the posterior distribution

The probability measure $\Pi(\cdot|X_1, ..., X_n)$ is a random probability measure that depends on observations. It can be used to answer question about the value of the parameter.

Definition. For a Bayesian model with posterior $\Pi(\cdot|X_1, ..., X_n)$,
• ESTIMATION: we can take, as an estimator of θ, the *posterior mean*

$$\bar{\theta}(X_1, ..., X_n) = \mathbb{E}_\Pi[\theta|X_1, ..., X_n]$$

we can also take the median and the mode.
• UNCERTAINTY ESTIMATION: Any subset $\mathcal{C}_n$ such that the posterior

$$\Pi(\mathcal{C}_n|X_1, ..., X_n) = 1 - \alpha$$

is a level $1 - \alpha$ *credible set* for θ.
• HYPOTHESIS TESTING: As in the motivating example, for $\Theta_0, \Theta_1 \subseteq \Theta$, the *Bayes factor* satisfies

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

Remark. Having different priors will yield in general different posterior distribution and different estimators. For example, consider a $N(0, 1)$ model. With $N(0, 1)$ prior we get $\frac{1}{n+1}\sum_{i=1}^n X_i$, as the posterior mean, while with Jeffreys prior we get $\frac{1}{n}\sum_{i=1}^n X_i$. In both cases (and in this example), the estimator $\bar{\theta}_n$ computes the 'true value'.

In the previous lecture we discussed the posterior mean $\bar{\theta} = \mathbb{E}_\Pi[\theta|X_1, ..., X_n]$ as an estimator of θ and behaviour of $\bar{\theta}$ (and more generally, of Π) when $X_i \overset{i.i.d.}{\sim} f(x, \theta_0)$.

Example. Sampling the $X_i \overset{i.i.d.}{\sim} N(0, 1)$, $\theta \in \mathbb{R}$ with prior $\theta \in N(0, 1)$. Posterior distribution

$$\Pi : \theta|X_1, ..., X_n \sim N(\frac{1}{n+1}\sum_{i=1}^n X_i, \frac{1}{n+1})$$

and mean

$$\frac{1}{n+1}\sum_{i=1}^n X_i = \frac{n}{n+1}\bar{X}_n$$

which is different from the MLE $\hat{\theta}_n = \bar{X}_n$.

Under the assumption $X_i \overset{i.i.d.}{\sim} N(0,1)$,

$$\bar{\theta}_n = \frac{n}{n+1}\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$$

as $n \to \infty$. For the deviation

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

but the second term $\xrightarrow{d} N(0,1)$ (applying CLT). The first term is $\sqrt{n}(\bar{\theta}_n - \hat{\theta}_n) = \sqrt{n}(\frac{1}{n+1} - \frac{1}{n})\sum_{i=1}^n X_i = -\frac{\sqrt{n}}{n+1}(\bar{X}_n - \theta_0 + \theta_0) \xrightarrow{P_{\theta_0}} 0$. By Slutsky's lemma, $\sqrt{n}(\bar{\theta}_n - \theta_0) \xrightarrow{d} N(0,1)$.

One of the consequences:

$$\mathcal{C}_n = \left\{ v : |v - \bar{\theta}_n| \leq \frac{I(\theta_0)^{-1/2} z_\alpha}{\sqrt{n}} \right\}$$

are 'good' confidence regions.

In a Bayesian formalism, the equivalent are credible sets, built not based on the limiting distribution of the estimator, but on the posterior:

$$\mathcal{C}_n = \{ v : |v - \hat{\theta}_n| \leq \frac{R_n}{\sqrt{n}}$$

or

$$\{ v : |v - \bar{\theta}_n| \leq \frac{R_n}{\sqrt{n}} \}$$

where R_n is chosen such that $\Pi(\mathcal{C}_n | X_1, ..., X_n) = 1 - \alpha$.

In order to prove that credible sets are confidence regions of level $1 - \alpha$, i.e. $\mathbb{P}_{\theta_0}(\theta_0 \in \mathcal{C}_n) \to 1 - \alpha$ as $n \to \infty$, we have to understand the behaviour of $\Pi(\cdot, |X_1, ..., X_n) = \Pi_n$.

Theorem. (Bernstein-von Mises theorem)
For a parametric model with the usual regularity assumptions, and a prior continuous at θ_0 such that $\pi(\theta_0) > 0$, consider the posterior $\Pi_n = \Pi(\cdot | X_1, ..., X_n)$ and ϕ_n the random distribution $N(\hat{\theta}_n, I^{-1}(\theta_0)/n)$. We have, for $\Theta \subseteq \mathbb{R}$ with $n \to \infty$, $||\Pi_n - \phi_n||_{TV}$, i.e. the total variation, is equal to

$$\int_\Theta |\Pi_n(\theta) - \phi_n(\theta)| d\theta \xrightarrow{a.s.} 0$$

Remark. This implies, for any event A, $\phi_n(A) - \Pi_n(A) \xrightarrow{a.s.} 0$. In particular, for credible sets $\mathcal{C}_n$, $\phi_n(\mathcal{C}_n) \to 1 - \alpha$ almost surely.

Recall the Bayesian formalism: given observation $x_1, ..., x_n$, we can build a posterior Π_n, treating $X_i \overset{i.i.d.}{\sim} f(x, \theta_0)$ makes Π_n, ϕ_n random objects.

Proof. (Informal)
$\phi_n, \mathbb{P}_n$ are probability distributions, so

$$\int_\Theta (\Pi_n(\theta) - \phi_n(\theta))d\theta = 1 - 1 = 0$$

This means

$$\int_\Theta |\Pi_n(\theta) - \phi_n(\theta)|d\theta = 2\int_\Theta (\Pi_n(\theta) - \phi_n(\theta))^+ d\theta$$

where we denote $x^+ = \max(x, 0)$.
For $\phi_n > 0$, we have the above is

$$2\int_\Theta \left(\frac{\Pi_n(\theta)}{\phi_n(\theta)} - 1\right)^+ \phi_n(\theta)d\theta$$

The first term in the product, $\frac{\Pi_n(\theta)}{\phi_n(\theta)} \to 1$ almost surely for all θ, $x \to (x-1)^+$ is bounded, so by dominated convergence theorem (see PM) we could obtain the conclusion. □

Intuition: we have that $\Pi_n(\theta) = \frac{\pi(\theta)\prod_{i=1}^n f(X_i,\theta)}{Z_n}$ where Z_n is a normalization factor independent of θ. The variation $V = \sqrt{n}(\theta - \hat{\theta}_n)$ (missing two lines?)

Taking logarithms, we have

$$\begin{aligned}
\log \Pi_{n,v}(v) &= \log \Pi_n(\hat{\theta}_n + \frac{v}{\sqrt{n}}) + \log \frac{1}{\sqrt{n}} \\
&= \log \pi(\hat{\theta}_n + \frac{v}{\sqrt{n}}) + l_n(\hat{\theta}_n + \frac{v}{\sqrt{n}}) - \log Z'_n \\
&\approx \log \pi(\theta_0) + l_n(\hat{\theta}_n) + l'(\hat{\theta}_n)\cdot\frac{v}{\sqrt{n}} + \frac{1}{2}l''_n(\hat{\theta}_n)\frac{v^2}{n} - \log Z'_n \\
&\approx -\frac{1}{2}I(\theta_0)v^2 - \log \tilde{Z}_n
\end{aligned}$$

(??)

Under ϕ_n, $\theta \sim N(\hat{\theta}_n, I(\theta_0)^{-1}/n)$, we have that $v \sim N(0, I^{-1}(\theta_0))$, so

$$\log \phi_{n,v} = -\frac{1}{2}I(\theta_0)v^2 - \log C(\theta_0)$$

(what's $C(\theta_0)$? or is it l)

Remark. A full proof is beyond the scope here, but special case can be done explicitly (see example sheet).
The message is that as $n \to \infty$, the impact/influence of the prior vanishes, and the behaviour of the posterior is dominated by the observations.
Laplace's method (not in this course) is concerned with the evaluation of $e^{nf(\theta)}$ for all θ such that $f(\theta) < f(\theta^*)$ we have $e^{nf(\theta)} << e^{nf(\theta^*)}$ (where $<<$ means exponentially smaller as n grows), and

$$\begin{aligned}
f(\theta^* + \frac{x}{\sqrt{n}} &\approx f(\theta^*) + \nabla_\theta f(\theta^*)\cdot\frac{x}{\sqrt{n}} - \frac{1}{2}\frac{x^+}{\sqrt{n}}\nabla^2_\theta f(\theta^*)\frac{x}{\sqrt{n}} \\
&\approx f(\theta^*) - \frac{1}{2}\frac{1}{n}x^T Q x
\end{aligned}$$

where $Q = -\nabla_\theta^2 f(\theta^*) \succeq 0$,

$$e^{nf(\theta^* + \frac{x}{\sqrt{n}})} \approx e^{nf(\theta^*)} \cdot e^{-\frac{1}{2}x^T Q x}$$

—- For the past lecture there is an error: we need to use $(1 - \frac{\Pi_n(\theta)}{\phi_n(\theta)})^+$ instead of the other way round, as $(1-x)^+$ is bounded for positive x but $(x-1)^+$ is not. —-

3.2 Credible sets

Let $\mathcal{C}_n = \{v : |v - \hat{\theta}_n| \leq \frac{R_n}{\sqrt{n}}\}$, where R_n is s.t. $\Pi_n(\mathcal{C}_n) = 1 - \alpha$.

Our objective is to let $\mathbb{P}_{\theta_0}(\theta_0 \in \mathcal{C}_n) \to 1 - \alpha$.

Remark: for confidence sets $\mathcal{C}'_n = \{v : |v - \hat{\theta}_n| \leq \frac{I(\theta_0)^{-1/2} z_\alpha}{\sqrt{n}}\}$, $\mathbb{P}(\theta_0 \in \mathcal{C}'_n) \to 1 - \alpha$,
• R_n converges almost surely to 'its frequentist equivalent';
• This implies that $\mathcal{C}_n$ (credible set) is a confidence set of level $1 - \alpha$ when $n \to \infty$.

Definition. (Notation)
For all $t > 0$, we define Φ_0, function defined by

$$\Phi_0(t) = \underbrace{\mathbb{P}(|Z_0| \leq t)}_{Z_0 \sim N(0, I^{-1}(\theta_0))} = \int_{-t}^{t} \phi_0(x) dx$$

It is an increasing continuous function, one-to-one from $[0, \infty)$ to $[0, 1)$. Its well-defined inverse Φ_0^{-1} is also continuous.

Lemma. Under the regularity assumption of the B-vM theorem, we have that R_n converges almost surely to $\Phi_0^{-1}(1 - \alpha)$ as $n \to \infty$.

Proof. We have

$$\begin{aligned}
\Phi_0(R_n) &= \int_{-R_n}^{R_n} \phi_0(v) dv \\
&= \int_{\hat{\theta_n} - R_n/\sqrt{n}}^{\hat{\theta}_n + R_n/\sqrt{n}} \phi_n(\theta) d\theta = \phi_n(\mathcal{C}_n) \\
&= \phi_n(\mathcal{C}_n) - \Pi_n(\mathcal{C}_n) + \Pi_n(\mathcal{C}_n)
\end{aligned}$$

where $v = \sqrt{n}(\theta - \hat{\theta}_n)$, and $\phi_n = N(\hat{\theta}_n, \frac{I(\theta_0)^{-1}}{n})$. By the B-vM theorem, $\phi_n(\mathcal{C}_n) - \Pi_n(\mathcal{C}_n) \xrightarrow{a.s.} 0$ as $n \to \infty$. Also, by definition of $\mathcal{C}_n$, $\Pi_n(\mathcal{C}_n) = 1 - \alpha$, so $\Phi_0(R_n) \xrightarrow{a.s.} 1 - \alpha$. So by CMT on Φ_0^{-1}, $\underbrace{\Phi_0^{-1}(\Phi_0(R_n))}_{R_n} \xrightarrow{a.s.} \Phi_0^{-1}(1 - \alpha)$. □

Theorem. Under the same assumption above, for $\alpha \in (0, 1)$, $n \to \infty$, we have $\mathbb{P}_{\theta_0}(\theta_0 \in \mathcal{C}_n) \to 1 - \alpha$.

Proof. We compute

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

we have that $\sqrt{n}(\hat{\theta}_n - \theta_0 \xrightarrow{d} N(0, I^{-1}(\theta_0))$. By the lemma above, we have

$$\frac{\Phi_0^{-1}(1-\alpha)}{R_n} \xrightarrow{a.s.} 1$$

So by Slutsky's lemma, the LHS term of the inequality converges in distribution to $N(0, I^{-1}(\theta_0))$. □

4 Decision theory

Given a statistical model $\{f(\cdot,\theta) : \theta \in \Theta\}$, an observation $X \in \mathcal{X}$, from this model that belongs to sample space $\mathcal{X}$, we can phrase statistical problems often as decision problems, with an action space $\mathcal{A}$ and decision rules $\delta : \mathcal{X} \to \mathcal{A}$.

Example. • In an hypothesis testing problem $\mathcal{A} = \{0,1\}$ and δ is a test. $\delta(X) \in \{0,1\}$;
• In an estimation problem, $\mathcal{A} = \Theta$ is a parameter space, and *delta* is an estimator: $\delta(X) \in \Theta$.
• Inference problems: $\mathcal{A}$ ="all subsets of Θ", and δ is a confidence region $\mathcal{C}_n$.

Definition. The performance of a decision δ is assessed by a *loss function* $L : \mathcal{A} \times \Theta \to [0,\infty)$. $L(a,\theta)$ describes the "loss" of action a when the true value is θ.

Example. • In a hypothesi testing problem, we can consider $L(a,\theta) = 1\{a \neq \theta\}$, where $\theta \in \{0,1\}$ is the index of the hypothesis (either we are right or wrong).
• In estimation problems, we care about the distance from the truth. We can consider

$$L(a,\theta) = |a-\theta|$$

which is the absolute deviation, or

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

which is the squared deviation.

Definition. For a loss function L, and a decision rule δ, we call the *risk* of δ, for $X \sim P_\theta$,

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

if P_θ has pdf $f(x,\theta)$.

Example. In a hypothesis testing

$$R(\delta,\theta) = \mathbb{E}_\theta[1\{\delta(X) \neq \theta\}] = \mathbb{P}_\theta(\delta(X) \neq \theta)$$

is the probability of error.

In an estimation problem we look at the squared risk

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

For $X \sim Bin(n,\theta)$ and $\theta \in [0,1]$, we can take $\delta(X) = \hat{\theta}(X) = \frac{X}{n}$,

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

a function of θ for fixed estimation. We can also take $\delta(X) = \hat{\eta}(X) = \frac{1}{2}$ (not very smart). In this case

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

So we see that we can't uniformly compare $\hat{\theta}$ and $\hat{n}$ in terms of their risk.

In the previous lecture we discussed about loss function L, decision rule δ, which results in risk $R(\delta, \theta) = \mathbb{E}_\theta[L(\delta(X), \theta)]$ where $X \sim P_\theta$.

4.1 Bayes rule for risk minimization

Definition. Given a loss function L, a decision rule δ, and a prior π over Θ, the (π-)Bayes risk of δ is

$$\begin{aligned} R_\pi(\delta) = \mathbb{E}_\pi[R(\delta, \theta)] &= \int_\Theta R(\delta, \theta)\pi(\theta)d\theta \\ &= \int_\Theta \int_\mathcal{X} L(\delta(x), \theta)\pi(\theta)f(x, \theta)dxd\theta \end{aligned}$$

from definition of pdf.

Example. In a binomial model $Bin(n\theta)$ with uniform prior on θ, we have for the quadratic risk, $R(X/n, \theta) = \theta(1-\theta)/n$ and $R(1/2, \theta) = (\theta - 1/2)^2$. Then

$$\begin{aligned} R_\pi(X/n) = \mathbb{E}_\pi[&\frac{\theta(1-\theta)}{n} \\ &= \frac{1}{n}\int_0^1 \theta(1-\theta)d\theta = \frac{1}{6n}, \\ R_\pi(1/2) = &\int_0^1 (\theta - 1/2)^2 d\theta = \frac{1}{12} \end{aligned}$$

Definition. For a Bayesian model (i.e. statistical model + prior), the corresponding posterior risk is defined for all observation $x \in \mathcal{X}$ as

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

Remark. The expectation is taken with respect to $\theta \sim \Pi(\cdot|x)$. In any model with a quadratic loss,

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

Proposition. An estimator δ that minimizes the Π-posterior risk for all $x \in X$ also minimizes the Bayes risk.

Proof. The π-Bayes risk (for a model with pdf)

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

Now let δ_Π be a decision rule that minimizes the posterior risk, i.e.

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

for all $x \in \mathcal{X}$. Multiplying on both sides by $m(x) \geq 0$ and integrating gives the result. □

Example. For the quadratic risk, for all x, minimizing this function in δ gives

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

As a consequence, in a Bayesian model with quadratic risk, the posterior mean is a rule that minimizes the Bayesian risk.

Definition. A decision rule that minimizes the Bayes risk is called a (π-) Bayes rule, denoted by δ_π.

Proposition. Let $|delta$ be an unbiased decision rule, 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 $\mathbb{E}_Q[(\delta(X)-\theta)^2] = \int_\Theta \mathbb{E}_\theta[(\delta(X)-\theta)^2]\pi(\theta)d\theta = 0$, where $\mathbb{E}_Q$ is taken with respect to the join distribution of (X,θ), $Q(x,\theta) = f(x,\theta)\pi(\theta)$. In particular, $\delta(X) = \theta$ with Q-probability 1.

Proof. We recall that for any random variable $Z(X,\theta)$, we can apply the 'tower rule'

$$\begin{aligned}\mathbb{E}_Q[Z(X,\theta)] &= \mathbb{E}_Q[\mathbb{E}_\Pi[Z(X,\theta)|X]] \\ &= \mathbb{E}_Q[\mathbb{E}_\theta[Z(X,\theta)]]\end{aligned}$$

□

For a π-decision rule, for the quadratic risk, we have that $\delta(x) = \mathbb{E}_\Pi[\theta|X]$. As a consequence, taking $Z(X,\theta) = \theta\delta(X)$ gives

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

and

$$\begin{aligned}\mathbb{E}_Q[\theta\delta(X)] &= \mathbb{E}_Q[\theta\mathbb{E}_\theta[\delta(x)]] \\ &= \mathbb{E}_Q[\theta^2]\end{aligned}$$

by unbiasedness. So

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

Remark. The result is counter-intuitive (estimator exactly correct with probability 1). The consequence of this idea is that 'unbiasedness' and minimization of the Bayes risk are usually disjoint properties.

In a normal model $N(\theta,1)$, the MLE $\bar{X}_n$ is unbiased. It is not a Bayes rule for any prior π.

In a $Bin(n,\theta)$ model, the MLE X/n is only a Bayes rule in very degenerate cases (see example sheet).

In the previous lecture we discussed the risk $R(\delta,\theta) = \mathbb{E}_\theta[L(\delta(X),\theta)]$, Bayes risk $R_\pi(\delta) = \mathbb{E}_\pi[R(\delta,\theta)]$.

4.2 Mimimax risk

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

Proposition. For any prior π and decision rule δ, we have

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

Proof. We have that

$$R_\pi(\delta) = \mathbb{E}_\pi[R(\delta, \theta)] \leq \sup_{\theta \in \Theta} R(\delta, \theta) = R_m(\delta, \Theta)$$

□

Definition. (Minimax risk)
The infimum over all decision rules of the maximal risk is known as the minimax risk: $\inf_\delta \sup_{\theta \in \Theta} R(\delta, \theta)$.

Taking the maximal risk as a way to evaluate estimates is a conservative approach. We want to mitigate the risk of the worst possible θ.

Definition. A decision rule that attains the minimax risk (as its maximal risk) is known as minimax.

Definition. A prior λ is called *least favourable* if for every prior λ', $R_\lambda(\delta_\lambda) \geq R_{\lambda'}(\delta_{\lambda'})$ corresponding to the worst case Bayes estimator.

Proposition. 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 it holds
1. The rule δ is minimax;
2. If δ_λ is unique Bayes, then it is unique minimax;
3. The prior λ is least favourable.

Proof. 1. Let δ be any decision rule. We have

$$\underbrace{\sup_{\theta \in \Theta} R(\delta, \theta)}_{R_m(\delta, \theta)} \geq \mathbb{E}_\lambda[R(\delta, \theta)] \geq \mathbb{E}_\lambda[R(\delta_\lambda, \theta)]$$

the first inequality is by the proposition above, and the second is by definition of δ_λ. The third term is then equal to $\underbrace{\sup_{\theta \in \Theta} R(\delta_\lambda, \theta)}_{R_m(\delta, \lambda, \theta)}$.

2. If δ_λ is unique, the second inequality is strict for $\delta \neq \delta_\lambda$. So $R_m(\delta, \theta) > R_m(\delta, \theta)$ for all $\delta \neq \delta_\lambda$.

3. For any prior λ', we have

$$\begin{aligned}\underbrace{R_{\lambda'}(\delta_{\lambda'})}_{\lambda'-Bayes\ risk} &= \mathbb{E}_{\lambda'}[R(\delta_{\lambda'},\theta)] \\ &\leq \mathbb{E}_{\lambda'}[R(\delta_\lambda,\theta)] \\ &\leq \sup_{\theta\in\Theta} R(\delta_\lambda,\theta) \\ &= \mathbb{E}_\lambda[R(\delta_\lambda,\theta)]\end{aligned}$$

Where the first inequality is by definition of $\delta_{\lambda'}$ and the second is by proposition above. □

Corollary. If a (unique) Bayes rule δ_λ has constant risk in θ (i.e. if $R(\delta_\lambda,\theta)$ is a constant function), then it is (the unique) minimax.

Proof. If δ_λ has constant risk, then

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

□

Example. If the maximal risk of δ_λ-Bayes risk, λ is leat favourable and δ_λ is minimax,
• In a $Bin(n,\theta)$, let $\pi_{a,b}$ be a $Beta(a,b)$ prior for $\theta\in[0,1]$. Then the unique Bayes rule for the quadratic risk is the posterior mean. $\delta_{a,b} = \bar{\theta}_{a,b} = \mathbb{E}_{\Pi_{a,b}}[\theta|x]$. Trying to find (a,b) such that $R(\delta_{a,b},\theta) = \text{const}$ will give us a^*, b^* and corresponding prior π_{a^*,b^*} with δ_{a^*,b^*} of constant risk, it is minimax, but not MLE (see example sheet).
• In a $N(0,1)$ model, with $\theta\in\mathbb{R}$, the MLE is minimax.

4.3 Admissibility

Definition. A decision rule δ is *inadmissible* if there exists δ' s.t.
• $R(\delta',\theta)\leq R(\delta,\theta)$ for all $\theta\in\Theta$, and
• $R(\delta',\theta) < R(\delta,\theta)$ for some $\theta\in\Theta$.
We say that δ' dominates δ. If δ is not inadmissible, it is admissible.

Remark. The intuition is that if δ is not admissible, we should always prefer an estimator that dominates δ.

It is not, however, the only criterion for a 'reasonable estimator'. In a lot of models, a constant estimator $\delta(X)=\theta_1$ is admissible.

Recall the previous example that $Bin(n,\theta)$, $\theta\in[0,1]$, $\pi_{a,b} = Beta[a,b]$.

Theorem. Let $X_1,...,X_n$ i.i.d. $N(\theta,\sigma^2)$, σ^2 known, and $\theta\in\Theta=\mathbb{R}$. Then $\hat{\theta}_{MLE} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. The MLE is adimissible and minimax in the quadratic risk (square loss $L(\delta,\theta) = (\delta-\theta)^2$).

Remark. 1-dimension Gaussian model.

Proof. WLOG let $\sigma^2 = 1$. The risk of the MLE:

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

which is constant in θ. For any decision rule δ we have

$$\begin{aligned} R(\delta, \theta) &= \mathbb{E}_\theta[(\delta(X) - \theta)^2] \\ &= \mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta[\delta] + \mathbb{E}_\theta[\delta] - \theta)^2] \\ &= \mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta[\delta])^2] + B(\theta)^2 \end{aligned}$$

where $B(\theta) = \mathbb{E}_\theta[\delta] - \theta$.

If we look at the proof of the CR inequality, we know that for biased estimators

$$Var_\theta[\delta] \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta[\delta]\right)^2}{I(\theta)} = \frac{(1 + B'(\theta))^2}{I(\theta)}$$

(bias is differentiable by smoothness, regularity assumption.

We have for decision rule δ,

$$R(\delta, \theta) \geq \underbrace{B^2(\theta)}_{bias\ squared} + \frac{(1 + B'(\theta))^2}{n}$$

If δ dominates the MLE, we have for all $\theta \in \mathbb{R}$, $R(\delta, \theta) \leq \underbrace{\frac{1}{n}}_{R(\bar{X}_n, \theta)}$ and $B^2(\theta) + \frac{(1+B'(\theta))^2}{n} \leq \frac{1}{n}$.

This inequality implies that $|B(\theta)| \leq \frac{1}{\sqrt{n}}$ and also $B'(\theta) \leq 0$, so B is non-increasing. We cannot have $B'(\theta) \leq -\varepsilon$ for some $\varepsilon > 0$, for values of θ going to either $+\infty$ or $-\infty$, otherwise the function would be unbounded. So we can take sequences $(\theta_n)_{n\geq 0}$, one going to $+\infty$, the other to $-\infty$, such that

$$\lim_{n\to\infty} B'(\theta_n) = 0.$$

As a consequence of $B^2(\theta) + (1 + B'(\theta))^2/n \leq \frac{1}{n}$, we therefore have that $\lim_n B(\theta_n) = 0$ for both sequences. But $B(\theta)$ is non increasing over $\mathbb{R}$, so $B(\theta) = 0$ for all $\theta \in \mathbb{R}$. So $|delta$ is unbiased and the CR inequality applies, $R(\delta, \theta) = \frac{1}{n}$ so δ does not dominate $\bar{X}_n$ (contradiction) so $\bar{X}_n$ is admissible. Recall that it has constant risk, so it is minimax. □

Remark. • $\bar{X}_n$ is not a Bayes rule for any prior in this model. It is however the "limit" of the Bayes rule $\delta_{\pi_{v^2}}$ with a $N(0, v^2)$ and $v \to \infty$. In general, all minimax rules can be obtained in this manner.
• The result still holds in a $N(\theta, I_2)$ model where $\theta \in \mathbb{R}^2$ ($p = 2$), but not for $p \geq 3$. In a $N(\theta, I_p)$ model with one observation, the MLE for θ is just X.

We consider the James-Stein estimator

$$\delta^{JS} = (\delta_1^{JS}...\delta_p^{JS})^T$$

where $\delta_i^{JS} = (1 - \frac{p-2}{||X||^2})X_i$. We will show that $R(\delta^{JS}, \theta) = \mathbb{E}_\theta||\delta_{JS} - \theta||^2$ dominates $R(X, \theta) = \mathbb{E}_\theta||X - \theta||^2 = \mathbb{E}[\sum_{i=1}^n (X_i - \theta)^2] = p$, therefore the latter is not admissible.

Lemma. (Stein's formula)
Let $X \sim N(\theta, 1)$ and $g : \mathbb{R} \to \mathbb{R}$ that is bounded, differentiable, with $\mathbb{E}[|g'(X)|] < \infty$. Then

$$\mathbb{E}_\theta[g'(X)] = \mathbb{E}_\theta[(X - \theta)g(X)]$$

Remark. Think of this a "Gaussian integration by parts".

Proof.

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

□

Remark. • If this result holds for some r.v. X for all such functions g, then $X \sim N(\theta, 1)$.
• If this result "almost holds" for some r.v. X for all such functions g, then X is "close to" $N(\theta, 1)$.

Proposition. For all $\theta \in \mathbb{R}^p$, for $p \geq 3$

$$R(\delta^{JS}, \theta) < R(X, \theta) = p.$$

Proof. Apply definition, $R(\delta_{JS}, \theta) = \mathbb{E}_\theta||(1 - \frac{p-2}{||X||^2})X - \theta||^2 = \mathbb{E}_\theta||X_\theta - \frac{p-2}{||X||^2}X||^2$. That is equal to

$$\begin{aligned}
&= \mathbb{E}_\theta||X - \theta||^2 + (p-2)^2\mathbb{E}_\theta[\frac{||X||^2}{||X||^4}] - 2(p-2)\mathbb{E}_\theta\frac{(X-\theta)^T X}{||X||^2} \\
&= p + (p-2)^2\mathbb{E}_\theta[\frac{1}{||X||^2}] - 2(p-2)\mathbb{E}_\theta\frac{(X-\theta)^T X}{||X||^2}(*)
\end{aligned}$$

Study the last term,

$$\begin{aligned}
\mathbb{E}_\theta\left[\frac{(X-\theta)^T X}{||X||^2}\right] &= \sum_{j=1}^p \mathbb{E}_\theta\left[\frac{(X_j - \theta_j)X_j}{||X||^2}\right] \\
&= \sum_{j=1}^p \mathbb{E}_\theta\left[\mathbb{E}_j[\frac{(X_j - \theta_j)X_j}{||X||^2} \mid X_{(-j)}]\right]
\end{aligned}$$

where $\mid X_{(-j)}$ means condition on all the values $X_1, ..., X_{j-1}, X_{j+1}, ..., X_p$ other than X_j.

We evaluate

$$\mathbb{E}_j \left[\frac{(X_j - \theta_j)X_j}{||X||^2} \mid X_{(-j)} \right] = \mathbb{E}\left[(X_j - \theta_j) \cdot g(X_j)\right]$$

where $X_j \sim N(\theta_j, 1)$ and $g(X_j) = \frac{X_j}{X_j^2 + \sum_{i \neq j} X_i^2}$. g is bounded, differentiate we get

$$g'(X_j) = \frac{(X_j^2 + \sum_{i \neq j} X_i^2) - 2X_j^2}{(\sum_{i=1}^p X_i^2)^2}$$

Applying Stein's formula (see page 39, a formula for expectation of derivative),

$$\begin{aligned} \mathbb{E}[(X_j - \theta_j)g(X_j)] &= \mathbb{E}[g'(X_j)] \\ &= \mathbb{E}\left[\frac{\sum_{i=1}^p X_i^2 - 2X_j^2}{||X||^4} \right] \end{aligned}$$

So

$$\sum_{j=1}^p \mathbb{E}[\mathbb{E}[(X_j - \theta_j)g(X_j)]] = \mathbb{E}\left[\frac{1}{||X||^4}(p||X||^2 - 2||X||^2) \right]$$

Putting the result of this computation in (*)(this was in page 39 at the start of proof), we get

$$\begin{aligned} R(\delta^{JS}, \theta) &= p + (p-2)^2 \mathbb{E}_\theta \left[\frac{1}{||X||^2} \right] - 2(p-2)\mathbb{E}\left[(p-2)\frac{||X||^2}{||X||^4} \right] \\ &= p - (p-2)^2 \mathbb{E}_\theta [\frac{1}{||X||^2}] \end{aligned}$$

We can show explicitly

$$\begin{aligned} \mathbb{E}_\theta \left[\frac{1}{||X|^2} \right] &= \int_{\mathbb{R}^p} \frac{1}{||x||^2} \phi(x - \theta) dx \\ &\leq \int_{\mathbb{R}^p} \frac{1}{||x||^2} \phi(x - \theta) dx \geq \frac{1}{c_i^2} \mathbb{P}_\theta(||x|| \in [c_1, c_2]) > 0 \end{aligned}$$

where ϕ is the pdf of the standard Gaussian in $\mathbb{R}^p$. □

Remark. • One proves that the supremum of the risk for this estimator is still p: they have the same maximial risk.
• While δ^{JS} dominates X, it is itself not admissible: the estimator $\delta^{JS+} = (1 - \frac{(p-2)}{||X||^2})^+ X$.
• Because it's not clear what would be a 'best estimator' in this model and of the reasons above, we might often prefer $\hat{\theta}_{MLE} = X$ (X_n with several observations). We know the limiting distribution of $\hat{\theta}_{MLE}$ and it is easier to work with.

4.4 Classification problems

This is a finite decision problem of practical importance. For example, we are given a vector X of clinical measurements $X = (X_1, ..., X_p)^T$ of an individual, and we want to know whether X belongs to a certain group.

Formally, we have two fixed populations in $\mathcal{X}$ with distribution $X \sim f_1$ or $X \sim f_2$ (e.g. distribution of X among people who don't or who do have the flu). This is similar in principle to a hypothesis testing problem, but only for one observation, not several.

In order to determine whether X belongs to population 1 or 2, we use a decision rule δ, with associated region $R \subseteq X$,

$$\delta = \delta_R(x) = \begin{cases} 1 & x \in R \\ 2 & x \in R^c \end{cases}$$

The probability of misclassifying $X \sim f_1$ is

$$P(2|1, R) = \int_{R^c} f_1(x)dx = P_1(X \in R^c)$$

and for $X \sim f_2$ it is

$$P(1|2, R) = \int_R f_2(x)dx = P_2(x \in R)$$

If $\pi = (q_1, q_2)$ $(q_2 = 1 - q_1)$ for $q_1 \in [0, 1]$ is a prior probability of the index of the population, the Bayes classification risk is

$$\mathbb{E}_\pi[R(\delta_R, \theta)] = q_1 P(2|1, R) + q_2 P(1|2, R)$$

Remark. Consider the joint distribution Q on $\mathcal{X} \times \{1, 2\}$, $Q(x, y) = f(x, y)\pi(y)$ where $\pi(1) = q_1$, $\pi(2) = q_2$ and $f(x, 1) = f_1(x)$, $f(x, 2) = f_2(x)$. Two interpretation for $(X, Y) \sim Q$.
• Y is drawn randomly from π. X is drawn from $f(x, y)$, conditionally on $Y = y$.
• X is drawn from P_X, the marginal distribution of X in the model

$$P_X(x) = \sum_y Q(x, y) = q_1 f_1(x) + q_2 f_2(x)$$

Y is drawn conditionally on $X = x$ with distribution

$$\Pi(1|X = x) = \frac{q_1 f_1(x)}{q_1 f_1(x) + q_2 f_2(x)},$$
$$\Pi(2|X = x) = \frac{q_2 f_2(x)}{q_1 f_1(x) + q_2 f_2(x)}$$

$(\eta(x), 1 - \eta(x))$ is a frequent notation.

These two interpretations give the same joint distribution, because

$$Q(x, y) = P_X(x)\Pi(y|X = x)$$

Proposition. The classification error, or Bayes classification risk, is given by

$$R_\pi(\delta) = \mathbb{P}_Q(\delta(X) \neq Y)$$

Proof. (informal)
$\mathbb{P}_Q(\delta(X) \neq Y) = \mathbb{P}_Q(Y = 1, \delta(X) \neq 1) + \mathbb{P}_Q(Y = 2, \delta(X) \neq 2)$ (first interpretation), and $\mathbb{P}_Q(\delta(X) \neq Y) = \mathbb{E}_Q[1\{\delta(X) \neq Y\}] = \int_X \Pi(\delta(x))dP_X(x)$, where $\delta(x)$ is the local probability of making a mistake. □

Definition. For a prior $\pi = (q_1, q_2)$, $0 < q_1 < 1$, the Bayes classifier is given by

$$\delta_\pi = \delta_R = \begin{cases} 1 & x \in R \\ 2 & x \in R^c \end{cases}$$

where $R\{x \in X : \frac{q_1 f_1(x)}{(1-q_1)f_2(x)} \geq 1\}$.

$$\delta_\pi = \delta_R = \begin{cases} 1 & \eta(x) \geq 1/2 \\ 2 & \eta(x) < 1/2 \end{cases}$$

Proposition. δ_π is a Bayes rule, it minimizes the Bayes classification risk. If

$$P_{f_i}(\frac{f_1(X)}{f_2(X)} = \frac{1-q_1}{q_1}) = 0$$

then δ_π is the unique Bayes rule.

Proof. Let J be some classification region. Its classification error is

$$q_1 \int_{J^c} f_1(x)dx + (1-q_1)\int_J f_2(x)dx$$
$$= \int_{J^c} [q_1 f_1(x) - (1-q_1)f_2(x)]dx + (1-q_1)\int_{X=J\cup J^c} f_2(x)dx$$

the first integral is minimized when taking

$$J = \{x \in X : \frac{f_1(x)}{f_2(x)} \geq \frac{1-q_1}{q_1}\}$$

and the second is independent of J. □

Remark. The intuition is that we minimize the probability of error by picking the most likely (locally) outcome, according to $\Pi(y|X = x)$ $(= y(x))$.
• Since a unique Bayes risk is admissible, the rule δ_Π is admissible.
• To find a minimax decision rule, we look for a prior π such that

$$qP(2|1, R_q) = (1-q)P(1|2, R_q)$$

$\pi_q = \{q, 1-q\}$ associated decision rule δ_{π_q}, with region R_q.

Example. Consider $f_1 = N(\mu_1, \Sigma)$, and $f_2 = N(\mu_2, \Sigma)$, where $\mu_i \in \mathbb{R}^p$, Σ is a $p \times p$ covariance matrix. One can show that hte Bayes classification $m|y$ on the discriminant function

$$D(X) = X^T\Sigma(\mu_1 - \mu_2)$$

Remark. In practice, f_1, f_2, π are not known in advance. In modern applications, we have $(X_1, Y_1), ..., (X_n, Y_n)$ from this distribution (e.g. a collection of images, with associated label "digit", "letter" etc). The objective is to minimize $\mathbb{P}(\delta(X) \neq Y)$ in δ. We minimize instead

$$\frac{1}{n}\sum_{i=1}^{n} 1\{\delta(X_i) \neq Y_i\}$$

If we can choose any function δ, the risk is to "overfit", i.e. taking $\delta(x_i) = y_i$ and whatever else for the rest.

To avoid this, we usually consider a class of decision function of the type $\delta_\beta(x) = 1\{f_\beta(x) \geq 1/2\}$, ($Y \in \{0, 1\}$ here), where f_β is smooth, or otherwise regular, parametrized by β.

In order to have a "tractable" optimization problem, instead of minimizing

$$\frac{1}{n}\sum_{i=1}^{m} 1\{\delta_\beta(X_i) \neq Y_i\} = \frac{1}{n}\sum_{i=1}^{n} 1\{|f_\beta(X_i) - Y_i| > 1/2\}$$

we minimize

$$\frac{1}{n}\sum l(f_\beta(X_i), Y_i)$$

i.e. the loss.