Principles of Statistics

Rajen D. Shah r.shah@statslab.cam.ac.uk

Contents

0 Introduction

In IB Statistics, you were introduced to the concept of an estimator of a parameter in a statistical model, a key example of such being the maximum likelihood estimator (MLE). But this study will have left several questions unanswered.

1. Is the MLE a good estimator? Is it the best estimator?

2. Or even before considering the above: what does it mean for an estimator to be good?

3. Often we want more than just a point estimate of a parameter; we want to quantify uncertainty in the form of confidence intervals (sets) or perform hypothesis tests. Clearly understanding the distributions of estimators would be helpful to achieving these goals: how can we do this in general?

These are questions of enormous practical significance. However, they are very difficult, and in fact the ongoing task of Statistics to answer these is as much generality as possible. This course will introduce some of the most important mathematical ideas involved their study. More specifically, this course divides into 5 chapters:

Likelihood inference. Here we will study the distribution of maximum likelihood estimators, and prove that they enjoy certain optimality properties. One very powerful idea that we will use to do this, and which pervades much of the course, is that of an asymptotic analysis of an estimator.

Bayesian inference. While the optimality of maximum likelihood estimators does provide some formal justification for their use, we shall see that there is nevertheless room to improve on them in finite samples. The Bayesian approach provides a way of leveraging prior information to potentially realise such an improvement.

Decision theory. We consider question 2 above in a general way, and derive perhaps the most surprising result of the course concerning the optimality of estimating the mean of Gaussian data, a result with far reaching consequences that has since shaped much of the direction of modern statistics for the last 50 years.

Multivariate analysis. While toy examples where all data points are one-dimensional scalars are useful to illustrate certain concepts, we typically have to handle data that has multiple variables. These settings bring new questions about how to understand relationships between variables or reduce their dimension to aid interpretability. Here we will examine some methods for performing such tasks.

Nonparametric inference and Monte Carlo techniques. As datasets have become larger in recent times, simple parametric models have become harder to defend (why should the data be Gaussian?). Nonparametric statistics is the field devoted to analysing data without making such restrictive assumptions on the data-generating process. A second major change in the way we analyse data has been brought about through improvements in computing power: computationally intensive simulation-based methods have revolutionised Bayesian statistics and indeed inference more generally. We will introduce some of the most important ideas involved in these developments.

1 Likelihood inference

1.1 Introduction

Recall that a random vector $X \in \mathbb{R}^d$ is a (measurable[1]) function $X : \Omega \to \mathbb{R}^d$ where Ω is a probability space. The distribution P of X gives the probability that X lies in any (measurable) set:

$$P(B) := \mathbb{P}(X \in B) = \mathbb{P}(\{\omega : X(\omega) \in B\}) \qquad \text{where } B \subseteq \mathbb{R}^d.$$

We write $X \sim P$ to indicate it has distribution P. The distribution is completely determined by the multivariate *distribution function* $F : \mathbb{R}^d \to [0, 1]$ given by

$$F(t) = \mathbb{P}(X \leq t),$$

where the inequality is to be understood elementwise. When X is a discrete random vector with probability mass function (pmf) f, we have

$$P(B) = \sum_{x \in B} f(x).$$

If X is continuous with probability density function (pdf)[2] f, then

$$P(B) = \int_B f(x)\, dx.$$

[1] The formal definition is covered in *Probability and Measure*; we will typically not refer to the measurability of functions or sets, though sometimes we will make connections to the material in that course

[2] Often we will phrase results in terms of densities, with the understanding that the same result would hold with pmfs after replacing associated integrals with sums.

Statistics is largely concerned with the following problem: given (data) X, infer something about its distribution P. We often assume that the distribution comes from some family of distributions, a so-called statistical model:

Definition 1. A *statistical model* is a family of distributions $\{P_\theta : \theta \in \Theta\}$ where Θ is a *parameter space*. When these distributions have densities or pmfs $f(\cdot, \theta)$, we may write this as $\{f(\cdot, \theta) : \theta \in \Theta\}$.

Example 1. (i) $N(\mu, \sigma^2) : (\mu, \sigma) \in \mathbb{R} \times [0, \infty)$.

(ii) $\text{Pois}(\theta) : \theta \in [0, \infty)$.

(iii) $N(\theta, 1) : \theta \in [-1, 1]$.

(iv) The (fixed design) normal linear model is the family of distributions for the random vector $Y = Z\beta + \varepsilon$ where $Z \in \mathbb{R}^{n \times p}$ is a deterministic matrix of predictors, $\beta \in \mathbb{R}^p$ is an unknown vector of coefficients and $\varepsilon \sim N_n(0, \sigma^2 I)$ with $\sigma \geq 0$. Thus the statistical model is formally $\{N_n(Z\beta, \sigma^2 I) : \beta \in \mathbb{R}^p, \sigma^2 \in [0, \infty)\}$.

The model is *well-specified* for X if our assumption $X \sim P_\theta$ for some $\theta \in \Theta$ holds, and we will typically assume this is the case unless we mention otherwise. Thus if $X \sim N(1, 2)$, then model (i) is well-specified, but (ii) and (iii) are *misspecified*. We denote expectations and variances etc. under such a model by adding a subscript θ e.g. $\mathbb{E}_\theta(X)$. In the case of a correctly specified model, we will often write θ_0 to denote the “true value” of θ where $X \sim P_{\theta_0}$, to distinguish it from other values of θ. This notation implicitly assumes that the following holds.

Definition 2. We say θ_0 is *identifiable* if whenever $\theta \in \Theta$ satisfies $P_\theta = P_{\theta_0}$, we have $\theta = \theta_0$.

Given a statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$, *maximum likelihood estimation* gives a recipe for constructing estimators of the unknown parameter θ. It works by regarding the joint density (or pmf) of the data under the postulated model as a function of θ known as the *likelihood*. Suppose x is a realisation of data $X \sim f(\cdot, \theta)$. Then the likelihood is given by

$$L(\theta) := L(\theta; x) = c(x) f(x, \theta),$$

where $c(x)$ is an arbitrary constant of proportionality. The *maximum likelihood estimator* (MLE) maximises this, or equivalently the log-likelihood ℓ, over $\theta \in \Theta$.

We will mainly work in the setting where our data consist of i.i.d. random vectors $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(\cdot, \theta)$ (though Example 1 (iv) is an important case that falls outside this scenario). In this context we often add a subscript n to the quantities involved. The log-likelihood then takes the form

$$\ell_n(\theta) = \ell_n(\theta; x_1, \ldots, x_n) := \log c(x_1, \ldots, x_n) + \sum_{i=1}^n \log f(x_i, \theta).$$

We often regard $\ell(\theta) = \ell(\theta; X)$ and $L(\theta) = L(\theta; X)$ as random functions of θ.

Example 2. Consider the model $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\theta)$ where $\theta \in (0, \infty)$. Then

$$L_n(\theta) = \prod_{i=1}^{n} e^{-\theta} \theta^{X_i},$$

so

$$\ell_n(\theta) = -n\theta + \log(\theta) \sum_{i=1}^{n} X_i.$$

Thus $\ell_n'(\theta; X_1, \dots, X_n) = 0$ when $\theta = \widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i =: \bar{X}$, which one may check is a maximiser and hence the MLE.

Example 3. Consider the fixed design normal linear model of Example 1(iv). We have, writing z_i for the ith row of $Z \in \mathbb{R}^{n \times p}$,

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (Y_i - z_i^\top \beta)^2 \right),$$

so

$$\ell(\beta, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \|Y - Z\beta\|_2^2.$$

Thus the MLE $\widehat{\beta}$ for β minimises the least squares term $\|Y - Z\beta\|_2^2$ and hence is given by the ordinary least squares estimator: when Z has full column rank, $\widehat{\beta} = (Z^\top Z)^{-1} Z^\top Y$.

Note that the MLEs of β and θ above have the attractive property of being *unbiased*: $\mathbb{E}_\theta \widehat{\theta} = \theta$. This need not be true of MLEs in general: recall that the MLE of σ^2 in Example 3 is

$$\widehat{\sigma}^2 := \frac{1}{n} \|Y - Z\widehat{\beta}\|_2^2,$$

which has $\mathbb{E}_{\beta,\sigma^2}(\widehat{\sigma}^2) = \sigma^2(n-p)/n$ and is thus (slightly) biased.

Maximum likelihood estimators are of course not the only possible estimators one could consider. Recalling that if $X_i \sim \text{Pois}(\theta)$ then $\text{Var}(X_i) = \theta$, another (somewhat) natural estimator for θ is the unbiased sample variance

$$\widehat{\theta}_v := \frac{1}{n-1} \sum_{i=1}^{n} (X_i^2 - \bar{X}^2).$$

Indeed, regardless of the distribution of the X_i, we have

$$\mathbb{E}X_i^2 - \mathbb{E}\bar{X}^2 = \text{Var}(X_i) + (\mathbb{E}X_i)^2 - \big(\text{Var}(\bar{X}) + (\mathbb{E}\bar{X})^2\big) = \frac{n-1}{n} \text{Var}(X_i).$$

Thus in the Poisson model, $\widehat{\theta}_v$ is also an unbiased estimator of θ. This observation raises the question of which of the two estimators is the best unbiased estimator. To answer this, we can for example compare the variances of each of these estimators. We will instead however pursue a much more ambitious goal of understanding this sort of problem in generality.

1.2 Information geometry and the Cramér–Rao lower bound

In order to determine the maximum likelihood estimator in Example 2 we differentiated the log-likelihood and set it equal to zero. It turns out that this derivative plays an even more fundamental role, so we give it a special name.

Definition 3. Suppose $\ell = \ell(\cdot, X) : \Theta \to \mathbb{R}$ is differentiable on $\operatorname{int}\Theta \subseteq \mathbb{R}^p$. The random function $S : \operatorname{int}\Theta \to \mathbb{R}^p$ given by

$$S(\theta) := \nabla_\theta \ell(\theta) = \left(\frac{\partial}{\partial \theta_1}\ell(\theta), \dots, \frac{\partial}{\partial \theta_p}\ell(\theta) \right)^\top$$

is known as the *score function* or *score.* In the case where we have i.i.d. data $X_1, \dots, X_n$, we will notate this as S_n.

The key property of the score is the following.

Lemma 1. *Let $g : \mathbb{R}^d \to \mathbb{R}$. Under appropriate 'regularity conditions' allowing integration and differentiation to be exchanged (see* Probability and Measure*), we have that for $\theta \in \operatorname{int}\Theta$,*

$$\mathbb{E}_\theta[S(\theta)g(X)] = \nabla_\theta \mathbb{E}_\theta g(X).$$

Proof. We have,

$$\begin{aligned} \mathbb{E}_\theta[S(\theta)g(X)] &= \int \nabla_\theta \log(f(x,\theta))g(x)f(x,\theta)\,dx \\ &= \int \nabla_\theta f(x,\theta)g(x)\,dx \\ &= \nabla_\theta \int f(x,\theta)g(x)\,dx \\ &= \nabla_\theta \mathbb{E}_\theta g(X). \end{aligned}$$

□

Remark 1. Taking $g \equiv 1$ in the above, we see in particular that $\mathbb{E}_\theta[S(\theta)] = 0$.

The property of the score[3] derived in Lemma 1 allows us to derive a fundamental lower bound on the variance of estimators in parametric models. To introduce this remarkable result, we first define the following.

Definition 4. The *Fisher information matrix* $I(\theta) \in \mathbb{R}^{p\times p}$ for $\theta \in \operatorname{int}\Theta$ is given by the variance of the score: $I(\theta) := \operatorname{Cov}_\theta(S(\theta)) = \mathbb{E}(S(\theta)S(\theta)^\top)$.

[3] There are some deep ideas at play here. The score can be thought of as the Reisz representer for the linear functional $g \mapsto \nabla_\theta \mathbb{E}_\theta g(X)$ (see *Linear Analysis*). This perspective shows (a) why we should expect a function with this crucial property satisfied by the score function to exist and (b) that the score function is unique in this regard.

Note that when $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(\cdot, \theta)$, the corresponding Fisher information *tensorises*: $I_n(\theta) = nI_1(\theta)$. Indeed,

$$\mathbb{E}_\theta(S_n(\theta)S_n(\theta)^\top) = \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}_\theta(\nabla_\theta \log f(X_i, \theta) \nabla_\theta \log f(X_j, \theta)^\top),$$

but the quantities $(\nabla_\theta \log f(X_i, \theta))_{i=1}^n$ are are i.i.d. and mean-zero (recall Remark 1), so we obtain

$$\mathbb{E}_\theta(S_n(\theta)S_n(\theta)^\top) = \sum_{i=1}^n \mathbb{E}_\theta(\nabla_\theta \log f(X_i, \theta) \nabla_\theta \log f(X_i, \theta)^\top) = nI_1(\theta).$$

Theorem 2 (Cramér–Rao lower bound)**.** *Suppose the model $\{P_\theta : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^p$ is sufficiently 'regular' (such that appropriate integration and differentiation operations may be interchanged). For a function $\phi : \Theta \to \mathbb{R}$, consider estimating $\phi(\theta)$ with an estimator $\widehat{\phi}$. Then for any $\theta \in \operatorname{int}\Theta$ where $I(\theta)$ is invertible,*

$$\operatorname{Var}_\theta(\widehat{\phi}) \geq \nabla_\theta \mathbb{E}_\theta(\widehat{\phi})^\top I(\theta)^{-1} \nabla_\theta \mathbb{E}_\theta(\widehat{\phi}).$$

Remark 2. Suppose $p = 1$, ϕ is the identity function and $\widehat{\phi} = \widehat{\theta}$ is a (potentially biased) estimator of θ. Then we have following the bound

$$\operatorname{Var}_\theta(\widehat{\theta}) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta \widehat{\theta}\right)^2}{I(\theta)}.$$

Remark 3. Take ϕ to be the function $\phi(\theta) = v^\top \theta$ for some vector $v \in \mathbb{R}^p$, and take $\widehat{\phi} = v^\top \widehat{\theta}$ for some unbiased estimator $\widehat{\theta}$ of θ. Then $\nabla_\theta \mathbb{E}_\theta(\widehat{\phi}) = \nabla_\theta(\mathbb{E}_\theta(\widehat{\theta})^\top v) = \nabla_\theta(\theta^\top v) = v$ and we thus obtain

$$v^\top \operatorname{Cov}_\theta(\widehat{\theta}) v \geq v^\top I(\theta)^{-1} v.$$

Since v above was arbitrary, we see that

$$\operatorname{Cov}_\theta(\widehat{\theta}) - I(\theta)^{-1}$$

is positive semi-definite. Thus, for example, $\operatorname{Var}_\theta(\widehat{\theta}_j) \geq (I(\theta)^{-1})_{jj}$.

Proof of Theorem 2. Let $v = I(\theta)^{-1/2}u$ for[4] an arbitrary unit vector u. We have

$$\begin{aligned}
|v^\top \nabla_\theta \mathbb{E}_\theta(\widehat{\phi})| &= |v^\top \mathbb{E}_\theta(S(\theta)\widehat{\phi})| && \text{applying Lemma 1} \\
&= |\mathbb{E}_\theta\{v^\top S(\theta)(\widehat{\phi} - \mathbb{E}_\theta(\widehat{\phi}))\}| && \text{by Remark 1} \\
&\leq \left(\mathbb{E}_\theta\{(v^\top S(\theta))^2\}\right)^{1/2} \left(\operatorname{Var}_\theta(\widehat{\phi})\right)^{1/2} && \text{by the Cauchy–Schwarz inequality.}
\end{aligned}$$

[4] $I(\theta)$ is symmetric so $I(\theta) = UDU^\top$ for orthogonal and diagonal matrices U and D respectively. Then $I(\theta)^{-1/2} := UD^{-1/2}U^\top$ where $D^{-1/2}$ is the diagonal matrix with jth diagonal entry $D_{jj}^{-1/2}$.

Rearranging, squaring and writing $b := \nabla_\theta \mathbb{E}_\theta(\widehat{\phi})$ for notational simplicity, we obtain

$$\mathrm{Var}_\theta(\widehat{\phi}) \geq \frac{(v^\top b)^2}{v^\top I(\theta) v}.$$

Substituting $v = I(\theta)^{-1/2}u$, we get

$$\mathrm{Var}_\theta(\widehat{\phi}) \geq (b^\top I(\theta)^{-1/2}u)^2.$$

This is true for all unit vectors u, so maximising over u by taking

$$u = \frac{I(\theta)^{-1/2}b}{\|I(\theta)^{-1/2}b\|_2} = \frac{I(\theta)^{-1/2}b}{\sqrt{b^\top I(\theta)^{-1}b}},$$

(if the denominator were zero, there would be nothing to prove), we finally arrive at

$$\mathrm{Var}_\theta(\widehat{\phi}) \geq b^\top I(\theta)^{-1} b,$$

as required. □

Example 2 continued. Recall that when $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Pois}(\theta)$, the MLE is $\widehat{\theta} = \bar{X}$ and is unbiased. Note that $\mathrm{Var}(\bar{X}) = \theta/n$. On the other hand, the Fisher information (for a single observation) is $I_1(\theta) = \mathrm{Var}(X/\theta) = 1/\theta$, so the Cramér–Rao lower bound is θ/n. We may thus conclude that (provided we are using the variance as our measure of quality), in this case, the MLE is the best unbiased estimator of θ!

The following result gives an alternative representation of the Fisher information that is often easier to work with.

Proposition 3. *Under regularity conditions, we have $I(\theta) = -\mathbb{E}_\theta[\nabla^2_\theta \log(f(X, \theta))]$ for $\theta \in \operatorname{int}\Theta$, i.e.*

$$I_{jk}(\theta) = -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta_j \partial\theta_k} \log(f(X,\theta))\right).$$

Proof. We have

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log(f(x,\theta)) = \frac{\partial}{\partial\theta_j}\left(\frac{\frac{\partial}{\partial\theta_k}f(x,\theta)}{f(x,\theta)}\right) = \frac{\frac{\partial^2}{\partial\theta_j\partial\theta_k}f(x,\theta)}{f(x,\theta)} - \frac{\frac{\partial}{\partial\theta_j}f(x,\theta)\frac{\partial}{\partial\theta_k}f(x,\theta)}{f(x,\theta)^2}.$$

Now, interchanging differentiation and integration,

$$\int \frac{\partial^2}{\partial\theta_j\partial\theta_k} f(x,\theta)\,dx = \frac{\partial^2}{\partial\theta_j\partial\theta_k}\int f(x,\theta)\,dx = 0.$$

Thus

$$\begin{aligned} -\mathbb{E}\left(\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log(f(X,\theta))\right) &= \mathbb{E}_\theta\left(\frac{\frac{\partial}{\partial\theta_j}f(X,\theta)}{f(X,\theta)}\frac{\frac{\partial}{\partial\theta_k}f(X,\theta)}{f(X,\theta)}\right) \\ &= \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta_j}\log f(X,\theta)\frac{\partial}{\partial\theta_k}\log f(X,\theta)\right), \end{aligned}$$

as required. □

Example 3 continued. In the normal linear model, the MLE $\widehat{\beta} = (Z^\top Z)^{-1} Z^\top Y$ for β satisfies $\mathrm{Var}_{\beta,\sigma^2}(\widehat{\beta}) = \sigma^2 (Z^\top Z)^{-1}$, which is the $p \times p$ submatrix of $I(\beta, \sigma^2)^{-1} \in \mathbb{R}^{(p+1)\times(p+1)}$ corresponding to β (see Example Sheet).

That the MLEs for β in the normal linear model and θ in the Poisson model are unbiased and achieve the Cramér–Rao lower bound is no accident: in fact, we shall see that such a relationship holds 'approximately' in wide generality. To gain some intuition about why this may be true, consider the simple case of Remark 2 where $p = 1$, but where additionally, we have i.i.d. data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(\cdot, \theta)$ and the estimator is unbiased. You will show on the Example Sheet that in this case, such an estimator $\widetilde{\theta}$ achieves the Cramér–Rao lower bound if and only if

$$\widetilde{\theta} = \theta + I_1^{-1}(\theta) \cdot \tfrac{1}{n} S_n(\theta).$$

Now typically, the MLE $\widehat{\theta}$ solves $S_n(\widehat{\theta}) = 0$. Performing a Taylor expansion around θ, we obtain

$$0 = \frac{1}{n} S_n(\widehat{\theta}) \approx \frac{1}{n} S_n(\theta) + (\widehat{\theta} - \theta) \cdot \frac{1}{n} \frac{d}{d\theta} S_n(\theta).$$

Provided $\frac{1}{n}\frac{d}{d\theta} S_n(\theta) = \frac{1}{n}\sum_{i=1}^n \frac{d^2}{d\theta} \log(f(X_i, \theta)) \approx -I_1(\theta)$, we have

$$\widehat{\theta} \approx \theta + I_1^{-1}(\theta) \cdot \tfrac{1}{n} S_n(\theta).$$

One requirement for this argument to go through is that the remainder in the Taylor expansion above is 'small'. This should occur provided $\widehat{\theta} - \theta$ is 'small' when θ is the true parameter θ_0.

As a first step towards arguing that this should hold, recall that an MLE $\widehat{\theta}$ maximises the (normalised) log-likelihood $\bar{\ell}_n(\cdot) := \frac{1}{n}\ell_n(\cdot)$. The result below shows that θ_0 maximises a population version of this quantity.

Theorem 4. *Consider a model $\{f(\cdot, \theta) : \theta \in \Theta\}$ and suppose $X \sim f(\cdot, \theta_0)$ where θ_0 is identifiable. Then θ_0 is the unique maximiser of*[5]

$$\theta \mapsto \mathbb{E}_{\theta_0}\left(\log f(X, \theta) - \log f(X, \theta_0)\right).$$

Proof. We make use of the fact that $\log u \leq u - 1$ for every $u \geq 0$ with equality if and only if $u = 1$, so

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

with equality if and only if $f(\cdot, \theta) = f(\cdot, \theta_0)$, which occurs if and only if $\theta = \theta_0$ by identifiability. □

[5] Minor technical point: The reason for subtracting $\log f(X, \theta_0)$ rather than just considering $\theta \mapsto \mathbb{E}_{\theta_0} \log f(X, \theta)$ is that the latter may be infinite.

Remark 4. The quantity $\mathbb{E}_{\theta_0}\{\log f(X,\theta_0)/\log f(X,\theta)\}$ is the *Kullback–Leibler* (KL) divergence $\mathrm{KL}(P_{\theta_0}, P_\theta)$ of P_θ from P_{θ_0}, where for distributions P, Q with densities p, q for a random variable X,

$$\mathrm{KL}(P,Q) := \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

The theorem shows that $\mathrm{KL}(p,q) \geq 0$, with equality only if p and q coincide.

A major requirement for the argument sketched out above for $\widehat{\theta}$ approximately achieving the Cramér–Rao lower bound is that certain empirical quantities are close to their population counterparts. Clearly if $n = 1$, this seems unrealistic, but for large n this might be more plausible. In the next section, we introduce some language and tools relating to convergence of random variables that will provide us with the means to justify this sort of claim formally.

1.3 Stochastic convergence

Recall that a random vector X is formally a function $X : \Omega \to \mathbb{R}^d$, where Ω is a probability space. The interpretation is that 'Chance picks an $\omega \in \Omega$ and we the see the realisation $X(\omega)$'. Formally we have that for any (measurable) set $B \subseteq \mathbb{R}^d$,

$$\mathbb{P}(X \in B) := \mathbb{P}(\{\omega : X(\omega) \in B\}) = \mathbb{P}(X^{-1}(B))$$

and $\mathbb{P}(\cdot)$ should be thought of as a sort of 'area measure' on Ω. Sets of the form $\{X \in B\} := \{\omega : X(\omega) \in B\} \subseteq \Omega$ are known as *events*. If an event has probability 1, we say it occurs *almost surely*.

Given that random vectors are functions, it is perhaps unsurprising that there are several notions of stochastic convergence. In the below, for a vector $x \in \mathbb{R}^d$, $\|x\|_\infty := \max_j |x_j|$.

Definition 5. Let $(X_n)_{n\in\mathbb{N}}$ and X be random vectors taking values in $\mathbb{R}^d$.

(i) We say X_n converges to X *almost surely* as $n \to \infty$, and write $X_n \xrightarrow{a.s.} X$, if

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

(ii) We say X_n converges to X *in probability*, and write $X_n \xrightarrow{p} X$, if for all $\epsilon > 0$,

$$\mathbb{P}(\|X_n - X\|_\infty > \epsilon) \to 0.$$

(iii) We say X_n converges *in distribution*, and write $X_n \xrightarrow{d} X$, if

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

at all points where $t \mapsto \mathbb{P}(X \leq t)$ is continuous.

Remark 5. Note that if $X_n \xrightarrow{d} X$ and X has a continuous distribution (most often we will have that X is normally distributed), then

$$\mathbb{P}(X_n \in B) \to \mathbb{P}(X \in B)$$

for 'most'[6] sets B. In particular, if X is real-valued,

$$\mathbb{P}(X_n \in [a,b]) \to \mathbb{P}(X \in [a,b])$$

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

The above definitions also apply to random matrices by concatenating their columns and regarding them as random vectors. Ultimately, it is convergence in distribution that is typically most useful to us. Nevertheless, the other forms of convergence are helpful, partly because they are stronger forms of convergence:

Remark 6.

$$X_n \xrightarrow{a.s.} X \Longrightarrow X_n \xrightarrow{p} X \Longrightarrow X_n \xrightarrow{d} X.$$

None of the reverse implications are true in general, but if $X_n \xrightarrow{d} c$ for some deterministic $c \in \mathbb{R}^d$, then $X_n \xrightarrow{p} c$ (see Example sheet).

The following two facts allow us to derive new convergences from old ones:

Remark 7. Given another sequence $(Y_n)_{n \in \mathbb{N}}$ of random vectors taking values in $\mathbb{R}^k$, we have that

$$(X_n, Y_n) \xrightarrow{p} (X, Y) \iff \begin{cases} X_n \xrightarrow{p} X \text{ and} \\ Y_n \xrightarrow{p} Y; \end{cases}$$

the same also holds with all convergences replaced by almost sure convergence, but does *not* hold for convergences in distribution (see Example Sheet).

Remark 8. We do however have that if $c \in \mathbb{R}^k$ is deterministic, then

$$(X_n, Y_n) \xrightarrow{d} (X, c) \iff \begin{cases} X_n \xrightarrow{d} X \text{ and} \\ Y_n \xrightarrow{p} c. \end{cases}$$

Theorem 5 (Continuous mapping theorem (CMT))**.** *Let $g : \mathbb{R}^d \to \mathbb{R}^m$ be continuous at every point of a set C such that $\mathbb{P}(X \in C) = 1$. Then*

$$X_n \xrightarrow{a.s./p/d} X \Longrightarrow g(X_n) \xrightarrow{a.s./p/d} g(X).$$

Combining this with Remark 8 yields the following useful result: if $g : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^m$ is continuous on the set $\mathbb{R}^d \times \{c\}$ and $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then $g(X_n, Y_n) \xrightarrow{d} g(X, c)$. Some common applications of this are known as Slutsky's lemma:

[6] *This holds for all measurable sets B for which $\mathbb{P}(X \in \delta B) = 0$, where $\delta B := \mathrm{cl}(B) \setminus \mathrm{int}(B)$ is the *boundary* of the set B.*

Lemma 6 (Slutsky's lemma)**.** *Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ where c is deterministic.*

(i) If X_n and Y_n are random vectors of the same dimension, $X_n + Y_n \xrightarrow{d} X + c$.

(ii) If Y_n is real-valued, $Y_n X_n \xrightarrow{d} cX$. Moreover, if $c \neq 0$ then $Y_n^{-1} X_n \xrightarrow{d} c^{-1}X$.

(iii) We also have a matrix version of (ii) above: if Y_n is a matrix of appropriate dimension, then $Y_n X_n \xrightarrow{d} cX$. Moreover, if c is invertible, then $Y_n^{-1} X_n \xrightarrow{d} c^{-1}X$.

It is natural to ask whether the limit of the expectations of a sequence X_n is equal to the expectation of the limiting random variable X. This is not the case in general, but we have the following result.

Theorem 7 (Dominated convergence theorem (DCT))**.** *Suppose a sequence of real-valued random variables $(W_n)_{n\in\mathbb{N}}$ satisfies $W_n \xrightarrow{p} W$, and there exists a random variable V that dominates the W_n in the sense that $|W_n(\omega)| \leq |V(\omega)|$ for all $\omega \in \Omega$. Then $\mathbb{E}|W| < \infty$ and $\mathbb{E}W_n \to \mathbb{E}W$.*

Example 4. Suppose we wish to establish continuity of $\theta \mapsto \mathbb{E}_{\theta_0} \log f(X, \theta)$ and we know that $\mathbb{E}_{\theta_0} V < \infty$ where $V = \sup_{\theta\in\Theta} |\log f(X, \theta)|$ and $\theta \mapsto f(x, \theta)$ is continuous for all x.

Take any sequence $(\theta_n)_{n\in\mathbb{N}} \subset \Theta$ with $\theta_n \to \theta \in \Theta$. Define $W_n := \log f(X, \theta_n)$. Then $W_n(\omega) = \log(f(X(\omega), \theta_n) \to \log(f(X(\omega), \theta)) =: W(\omega)$. (Note this convergence holds for all $X(\omega) \in \{x : f(x, \theta_0) > 0\}$: we cannot have $f(x, \theta) = 0$ when $f(x, \theta_0) > 0$ since $\mathbb{E}V < \infty$. Thus in particular we have this convergence $\mathbb{P}_{\theta_0}$-almost surely and hence in probability too.) As V is a dominating function, by the DCT, $\mathbb{E}W_n \to \mathbb{E}W$ and so $\theta \mapsto \mathbb{E}_{\theta_0} \log f(X, \theta)$ is continuous.

1.4 Laws of large numbers and the central limit theorem

Many results in Statistics have at their heart, convergences of averages of i.i.d. random variables.

Theorem 8 (Strong law of large numbers (SLLN))**.** *Let $X_1, X_2, \ldots$ be i.i.d. taking values in $\mathbb{R}^d$ with $\mathbb{E}\|X_1\|_\infty < \infty$. Then*[7]

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

[7] Note that the underlying probability space Ω needs to support not just any finite number $X_1, \ldots, X_n$ of independent random vectors, but the entire infinite sequence $X_1, X_2, \ldots$ of independent random vectors: writing out the conclusion explicitly, we have

$$\mathbb{P}(\{\omega \in \Omega : \bar{X}_n(\omega) \to \mathbb{E}(X)\}) = 1.$$

(Showing that such a probability space exists takes some care: see *Probability and Measure* for more details.)

We shall show a weaker result known as the *weak law of large numbers* that is easier to prove:

Theorem 9 (Weak law of large numbers)**.** *Let $X_1, \dots, X_n$ be i.i.d. real-valued random variables with* $\mathrm{Var}(X_1) < \infty$*. Then*

$$\bar{X}_n \xrightarrow{p} \mathbb{E}(X_1).$$

Note that the assumption $\mathrm{Var}(X_1) < \infty$ automatically includes the assumption $\mathbb{E}(|X_1|) < \infty$ so in this sense, the assumption of the WLLN is stronger than that of the SLLN.

Proof of Theorem 9. Applying Markov's inequality[8] to $\{\bar{X}_n - \mathbb{E}(X_1)\}^2$, we have

$$\mathbb{P}(\{\bar{X}_n - \mathbb{E}(X_1)\}^2 > \epsilon^2) \le \epsilon^{-2}\mathbb{E}\{\bar{X}_n - \mathbb{E}(X_1)\}^2.$$

But

$$\mathbb{E}\{\bar{X}_n - \mathbb{E}(X_1)\}^2 = \mathrm{Var}(\bar{X}_n) = \mathrm{Var}(X_1)/n,$$

so

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}(X_1)| > \epsilon) \le \epsilon^{-2}\mathrm{Var}(X_1)/n \to 0$$

as $n \to \infty$. □

Example 5. Suppose $X_1, \dots, X_n$ are i.i.d. with mean μ_0 and variance $\sigma_0^2 > 0$. We shall show that the sample variance satisfies

$$\widehat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \xrightarrow{p} \sigma^2.$$

First note that we may subtract μ_0 from each X_i and $\widehat{\sigma}_n^2$ is unchanged. Now

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2 = \underbrace{\frac{1}{n}\sum_{i=1}^n (X_i - \mu_0)^2}_{\xrightarrow{a.s.}\sigma_0^2 \text{ by SLLN}} - \underbrace{(\bar{X}_n - \mu_0)^2}_{\xrightarrow{a.s.} 0 \text{ by SLLN and CMT}}.$$

Thus by Slutsky, we have $\widehat{\sigma}_n^2 \xrightarrow{p} \sigma_0^2$.

In fact, we can characterise the limiting behaviour of the average of i.i.d. random variables much more precisely. This turns out to be crucial for deriving inference results for estimators.

[8] Recall that if Z is a non-negative random variable, then $Z \ge t\mathbb{1}_{\{Z \ge t\}}$, so taking expectations, $t^{-1}\mathbb{E}(Z) \ge \mathbb{P}(Z \ge t)$.

Theorem 10 (Central limit theorem (CLT)). *Let $X_1, \ldots, X_n$ be i.i.d. taking values in $\mathbb{R}^d$ with finite variance Σ. Then*[9]

$$\sqrt{n}\{\bar{X}_n - \mathbb{E}(X_1)\} \xrightarrow{d} N_d(0, \Sigma).$$

Example 5 continued The CLT can for example be used to construct confidence intervals for μ_0 in the setting of Example 5. We have by the CLT that

$$\sqrt{n}(\bar{X}_n - \mu_0) \xrightarrow{d} N(0, \sigma_0^2)$$

so by Slutsky,

$$\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\widehat{\sigma}_n} \xrightarrow{d} N(0, 1).$$

Write z_α for the upper $\alpha/2$ point of a $N(0,1)$ distribution, so if $Z \sim N(0,1)$, then $\mathbb{P}(Z \in [-z_\alpha, z_\alpha]) = 1 - \alpha$. Then

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

is an *asymptotically valid* $(1-\alpha)$-level confidence interval, in the sense that

$$\mathbb{P}(\mu_0 \in \widehat{C}_n) = \mathbb{P}\left(\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\widehat{\sigma}_n} \in [-z_\alpha, z_\alpha]\right) \to \mathbb{P}(Z \in [-z_\alpha, z_\alpha]) = 1 - \alpha.$$

In the example above, we had an explicit expression for the estimators of the mean and variance, and so we were able to apply the limit theorems above rather directly. Recall however that our objective is to study the behaviour of maximum likelihood estimators in generality, and the MLE may only be defined implicitly through a maximiser of the random function $\theta \mapsto \bar{\ell}_n(\theta)$. While for any given fixed θ, the SLLN can for example be used to conclude that $\bar{\ell}_n(\theta) \xrightarrow{a.s.} \mathbb{E}\bar{\ell}_n(\theta)$, it does not offer any conclusions about the convergence of the function $\bar{\ell}_n(\cdot)$ as a whole to its population counterpart. This is problematic since the MLE may be sensitive to the entire function. Fortunately, there exist uniform versions of the convergence results above. Known as *uniform laws of large numbers*, they can provide sufficient conditions such that

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i, \theta) - \mathbb{E}(g(X, \theta)) \right| \xrightarrow{p} 0, \tag{1.1}$$

[9] Recall that a random vector $X \in \mathbb{R}^d$ with mean μ and positive definite covariance matrix Σ has a normal distribution (and we write $X \sim N_d(\mu, \Sigma)$) if its pdf f is given by

$$f(x) = \frac{1}{(2\pi)^{d/2}} \frac{1}{(\det(\Sigma))^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

where $X_1, \dots, X_n$ are i.i.d. copies of a random variable X taking values in $\mathcal{X} \subseteq \mathbb{R}^d$, $\Theta \subset \mathbb{R}^p$ and $g : \mathcal{X} \times \Theta \to \mathbb{R}$ is a given function. The following is one example of such a result that we will make use of. Its precise statement and proof are **non-examinable**.

Theorem 11 (Uniform law of large numbers)**.** *In the setting above, suppose Θ is compact (i.e. closed and bounded) and that $\theta \mapsto g(x, \theta)$ is continuous for all $x \in \mathcal{X}$. Suppose further that there exists a function $G(x) \geq \sup_{\theta \in \Theta} |g(x, \theta)|$ satisfying $\mathbb{E}G(X) < \infty$. Then* (1.1) *holds.*

**Proof*.* Write $B(\theta, \delta)$ for the open ball with radius δ centred at $\theta \in \Theta$. Fix $\theta_0 \in \Theta$ and consider

$$\Delta_\delta(X, \theta_0) := \sup_{\theta \in B(\theta_0, \delta) \cap \Theta} \{g(X, \theta) - \mathbb{E}g(X, \theta)\}.$$

We claim that $\mathbb{E}\Delta_\delta(X, \theta_0) \to 0$ as $\delta \to 0$[10]. Indeed $|\Delta_\delta(X, \theta)| \leq G(X) + \mathbb{E}G(X)$, so by the DCT, this holds provided $\Delta_\delta(x, \delta_0) \to g(x, \theta_0) - \mathbb{E}g(X, \theta_0)$. This latter fact follows from continuity of $\theta \mapsto g(x, \theta)$ and the DCT (which shows that $\theta \mapsto \mathbb{E}g(X, \theta)$ is continuous—see Example 4).

Now fix $\epsilon > 0$. We know that for all $\theta \in \Theta$, there exists some $\delta(\theta) > 0$ such that $\mathbb{E}\Delta_{\delta(\theta)}(X, \theta) < \epsilon/2$. The set $\{B(\theta, \delta(\theta)) : \theta \in \Theta\}$ forms an open cover of the compact set Θ, so we can find a finite subcover $\Theta \subseteq \bigcup_{k=1}^K B(\theta_k, \delta(\theta_k))$. Let $B_k := B(\theta_k, \delta(\theta_k)) \cap \Theta$ and $\Delta_k(x) := \Delta_{\delta(\theta_k)}(x, \theta_k)$. Then

$$\begin{aligned}
\sup_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^n \{g(X_i, \theta) - \mathbb{E}g(X, \theta)\} \right) &= \max_{k=1,\dots,K} \sup_{\theta \in B_k} \left(\frac{1}{n} \sum_{i=1}^n \{g(X_i, \theta) - \mathbb{E}g(X, \theta)\} \right) \\
&\leq \max_{k=1,\dots,K} \frac{1}{n} \sum_{i=1}^n \sup_{\theta \in B_k} \{g(X_i, \theta) - \mathbb{E}g(X, \theta)\} \\
&= \max_{k=1,\dots,K} \frac{1}{n} \sum_{i=1}^n \Delta_k(X_i).
\end{aligned}$$

Thus

$$\begin{aligned}
\mathbb{P}\left\{ \sup_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^n \{g(X_i, \theta) - \mathbb{E}g(X, \theta)\} \right) \leq \epsilon \right\} &\geq \mathbb{P}\left(\max_{k=1,\dots,K} \frac{1}{n} \sum_{i=1}^n \Delta_k(X_i) \leq \epsilon \right) \\
&\geq \mathbb{P}\Bigg(\max_{k=1,\dots,K} \left| \frac{1}{n} \sum_{i=1}^n \Delta_k(X_i) - \underbrace{\mathbb{E}\Delta_k(X)}_{<\epsilon/2} \right| \leq \epsilon/2 \Bigg) \\
&\to 1
\end{aligned}$$

[10] For measure theory enthusiasts: $\Delta_\delta(X, \theta_0)$ is a supremum over an uncountable set, and it is not clear if it is measurable: technically this measurability should be an extra assumption in the statement.

as $n \to \infty$ by SLLN (and Remark 6). Applying a similar argument replacing g with $-g$,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\{g(X_i,\theta)-\mathbb{E}g(X,\theta)\}\right|>\epsilon\right)\\
&\leq \mathbb{P}\left\{\sup_{\theta\in\Theta}\left(\frac{1}{n}\sum_{i=1}^{n}\{g(X_i,\theta)-\mathbb{E}g(X,\theta)\}\right)>\epsilon\right\}+\mathbb{P}\left\{\sup_{\theta\in\Theta}\left(\frac{1}{n}\sum_{i=1}^{n}\{\mathbb{E}g(X,\theta)-g(X_i,\theta)\}\right)>\epsilon\right\}\\
&\to 0,
\end{aligned}
$$

as required. □

1.5 Consistency of the MLE

We now have in place all of the tools required to formalise the argument sketched out earlier for connecting the MLE to the Cramér–Rao lower bound. Recall that our basic strategy involved a Taylor expansion, and in order to make progress with this, a first requirement was that the MLE $\widehat{\theta}_n$ be 'close' to the true parameter θ_0 (for large n). The appropriate form of closeness here is convergence in probability. From herein, we will be working in the setting where our data consist of i.i.d. copies $X_1,\ldots,X_n$ of a random vector $X \in \mathbb{R}^d$.

Definition 6. We say an estimator $\widehat{\theta}_n = \widehat{\theta}_n(X_1,\ldots,X_n)$ (not necessarily the MLE) is *consistent* for estimating a parameter θ_0 (corresponding to the true distribution) if $\widehat{\theta}_n \xrightarrow{p} \theta_0$.

We can thus re-express the first conclusion of Example 5 as showing that the sample variance is a consistent estimator of the population variance. Consistency is a very basic requirement for an estimator: indeed, the strong law of large numbers shows that provided $\mathbb{E}\|X\|_\infty < \infty$, the sample average of the first $\log n$ data points (discarding all other data) $\frac{1}{\log n}\sum_{i=1}^{\log n} X_i$ is a consistent estimator of the mean. Nevertheless it is a good first start for studying the MLE, and we present this now.

In the following we assume that the X_i have a distribution from a statistical model $\{f(\cdot,\theta) : \theta \in \Theta\}$ where $\theta_0 \in \Theta$ denotes the true parameter, so $X_i \overset{\text{i.i.d.}}{\sim} f(\cdot,\theta_0)$, and θ_0 is identifiable. We will require several regularity conditions on our statistical model; note that the precise form of these is **non-examinable**: on the example sheet or exam, these will be referred to as the 'usual regularity conditions'.

Regularity assumptions (R1). Let the statistical model be such that:

(1) $\Theta \subset \mathbb{R}^p$ is closed and bounded (compact);

(2) Writing $\mathcal{X} := \{x : f(x,\theta_0) > 0\}$ for the *support* of $f(\cdot,\theta_0)$, $\theta \mapsto f(x,\theta)$ is continuous for all $x \in \mathcal{X}$;

(3) $\mathbb{E}_{\theta_0}\sup_{\theta\in\Theta}|\log(f(X,\theta))| < \infty$.

In particular, these assumptions are required for application of our ULLN. (3) and (2) imply that $\theta \mapsto \mathbb{E}_{\theta_0} \log(f(X, \theta)) =: \bar{\ell}(\theta)$ is continuous (see Example 4). Note that a necessary condition for (3) is that all the densities $\{f(\cdot, \theta) : \theta \in \Theta\}$ have support containing $\mathcal{X}$.

Theorem 12. *Suppose (R1) holds. An MLE exists (almost surely) and any MLE is consistent.*

Proof. First note that on the almost sure event $\{X_1, \ldots, X_n \in \mathcal{X}\}$, $\bar{\ell}_n(\theta)$ is continuous, so a maximiser on the compact set Θ exists. Also, the regularity assumptions accommodate the following ULLN:

$$\sup_{\theta \in \Theta} |\bar{\ell}_n(\theta) - \bar{\ell}(\theta)| \xrightarrow{p} 0.$$

Now fix $\epsilon > 0$ and let $\Theta_- := \{\theta \in \Theta : \|\theta - \theta_0\|_\infty \geq \epsilon\}$. Note that for any MLE $\widehat{\theta}_n$,

$$\{\bar{\ell}_n(\theta_0) > \sup_{\theta \in \Theta_-} \bar{\ell}_n(\theta)\} \subseteq \{\|\widehat{\theta}_n - \theta_0\|_\infty < \epsilon\},$$

so it suffices to show that the former event has probability converging to 1.

Now Θ_- is a closed and bounded (compact) set as the intersection of the compact set Θ and the closed set $\{\|\theta - \theta_0\|_\infty < \epsilon\}^c$. Recall that $\bar{\ell}$ is continuous, so there exists $\theta_- \in \Theta_-$ with $\bar{\ell}(\theta_-) = \sup_{\theta \in \Theta_-} \bar{\ell}(\theta)$. Let us write $\delta := \bar{\ell}(\theta_0) - \bar{\ell}(\theta_-) > 0$ (recall Theorem 4 which shows θ_0 is the unique maximiser of $\bar{\ell}$).

Also,

$$\sup_{\theta \in \Theta_-} \bar{\ell}_n(\theta) \leq \sup_{\theta \in \Theta_-} \bar{\ell}(\theta) + \sup_{\theta \in \Theta} |\bar{\ell}_n(\theta) - \bar{\ell}(\theta)|.$$

Now

$$\bar{\ell}_n(\theta_0) - \sup_{\theta \in \Theta_-} \bar{\ell}_n(\theta) = \bar{\ell}_n(\theta_0) - \bar{\ell}(\theta_0) + \underbrace{\bar{\ell}(\theta_0) - \bar{\ell}(\theta_-)}_{=\delta} + \underbrace{\bar{\ell}(\theta_-) - \sup_{\theta \in \Theta_-} \bar{\ell}_n(\theta)}_{\geq -\sup_{\theta \in \Theta} |\bar{\ell}_n(\theta) - \bar{\ell}(\theta)|}.$$

But by ULLN, $\mathbb{P}(\sup_{\theta \in \Theta} |\bar{\ell}_n(\theta) - \bar{\ell}(\theta)| < \delta/2) \to 1$, so $\mathbb{P}(\bar{\ell}_n(\theta_0) > \sup_{\theta \in \Theta_-} \bar{\ell}_n(\theta)) \to 1$ as required. □

Remark 9. The proof extends to the following more general setting where we replace the log-likelihood $\bar{\ell}_n$ maximised by the MLE, by another function

$$\theta \mapsto M_n(\theta) := \sum_{i=1}^n m(\theta, X_i).$$

Let $M(\theta) := \mathbb{E}m(\theta, X)$ and suppose θ_0 is a maximiser of M: for example we could take $m(\theta, X) = -|X - \theta|$, in which case θ_0 would be a population median. Provided the appropriate regularity conditions are met, we may conclude that a maximiser $\widehat{\theta}_n$ of $M_n(\theta)$ is consistent for estimating θ_0.

1.6 Asymptotic normality of the MLE

We are finally ready to formalise the Taylor series-based argument sketched out earlier concerning the MLE. We first re-state this argument for the case where p may be greater than one. Define the *observed information matrix* $J_n(\theta) \in \mathbb{R}^{p\times p}$ with entries

$$(J_n(\theta))_{jk} := -\frac{\partial}{\partial\theta_k}(\tfrac{1}{n}S_n(\theta))_j = -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(X_i,\theta).$$

(Recall that, under regularity conditions, $\mathbb{E}_\theta J_n(\theta) = I_1(\theta)$; see Proposition 3.) We have

$$0 = \frac{1}{\sqrt{n}}S_n(\widehat{\theta}_n) \approx \frac{1}{\sqrt{n}}S_n(\theta_0) - J_n(\theta_0)\sqrt{n}(\widehat{\theta}_n - \theta_0).$$

But by the CLT,

$$\frac{1}{\sqrt{n}}S_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \nabla_\theta \log f(X_i,\theta) \overset{d}{\to} N(0, I_1(\theta_0))$$

so if $J_n(\theta_0) \overset{p}{\to} I_1(\theta_0)$, by Slutsky, we should expect

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \overset{d}{\to} N_p(0, I_1(\theta_0)^{-1}).$$

That is, MLEs should not only enjoy a form of 'approximate' optimality by approximately achieving the Crameŕ–Rao lower bound, but also have an approximately Gaussian distribution, a fact which allows one to quantify uncertainty in the estimation and perform hypothesis tests. A sequence of estimators achieving this distributional convergence is said to be *asymptotically efficient* or simply *efficient*. Such a sequence of estimators 'asymptotically' is unbiased and achieves the Cramér–Rao lower bound (but for example we are not guaranteed that $\mathbb{E}_{\theta_0}\widehat{\theta}_n \to \theta_0$).

To prove the remarkable result hinted at above, we require some regularity conditions in addition to (R1), which, as before, are **non-examinable**.

Regularity assumptions (R2). Let the statistical model be such that:

(1) $\theta_0 \in \operatorname{int}\Theta$;

(2) there exists an open neighbourhood N of θ_0 on which $\theta \mapsto f(x,\theta)$ is twice continuously differentiable for all $x \in \mathcal{X}$;

(3) $I(\theta_0)$ exists and is invertible ;

(4)

$$\begin{aligned}
\int_{\mathcal{X}} \sup_{\theta\in\Theta} \|\nabla_\theta \log f(x,\theta)\|_2 \, dx &< \infty \\
\mathbb{E}_{\theta_0}\left(\sup_{\theta\in\Theta} \|\nabla^2_\theta \log f(X,\theta)\|_2 \right) &< \infty \\
\int_{\mathcal{X}} \sup_{\theta\in\Theta} \|\nabla^2_\theta \log f(x,\theta)\|_2 \, dx &< \infty.
\end{aligned}$$

(In fact the above can be weakened by replacing Θ above with any compact set K with non-empty interior that contains θ_0.)

Theorem 13. *Suppose regularity conditions (R1) and (R2) hold. Then any sequence of MLEs $\widehat{\theta}_n$ satisfies*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N_p(0, I_1(\theta_0)^{-1}).$$

Proof. As $\theta_0 \in \operatorname{int}\Theta$, there exists some $\epsilon > 0$ such that $\{\theta : \|\theta - \theta_0\|_\infty \leq \epsilon\} \in \operatorname{int}\Theta$. We already know $\widehat{\theta}_n \xrightarrow{p} \theta_0$ (Theorem 12), so writing $A_n := \{\widehat{\theta}_n \in \operatorname{int}\Theta\} \supseteq \{\|\widehat{\theta}_n - \theta_0\|_\infty \leq \epsilon\}$, we have $\mathbb{P}(A_n) \to 1$. We henceforth work on this sequence of events,[11] noting that it will not affect the distributional convergence result[12].

Now fix $j \in \{1, \ldots, p\}$ and define $q(t) := S_n(t\widehat{\theta}_n + (1-t)\theta_0)_j$. Then, by the mean value theorem,

$$q(1) - q(0) = q'(t)$$

for some $t \in [0,1]$. Thus there exists $\widetilde{\theta}_n^{(j)} \xrightarrow{p} \theta_0$ with[13]

$$S_n(\widehat{\theta}_n)_j - S_n(\theta_0)_j = \sum_{k=1}^p (\widehat{\theta}_n - \theta_0)_k \sum_{i=1}^n \frac{\partial^2}{\partial\theta_k \partial\theta_j} \log f(X_i, \theta)\big|_{\theta = \widetilde{\theta}_n^{(j)}}.$$

Then, defining $\widetilde{J}_n \in \mathbb{R}^{p\times p}$ with $(\widetilde{J}_n)_{jk} := (J_n(\widetilde{\theta}^{(j)}))_{jk}$, we have

$$\frac{1}{n} S_n(\widehat{\theta}_n) - \frac{1}{n} S_n(\theta_0) = -\widetilde{J}_n(\widehat{\theta}_n - \theta_0).$$

Now on A_n, $S_n(\widehat{\theta}_n) = 0$. Also $\widetilde{J}_n \xrightarrow{p} I_1(\theta_0)$ (see Lemma 14 below). Thus writing $B_n := \{\widetilde{J}_n \text{ is invertible}\}$, we have $\mathbb{P}(B_n) \to 1$ (see Example Sheet 1, Question 11(i)). We now work on $A_n \cap B_n$ (and note that $\mathbb{P}(A_n \cap B_n) \to 1$). We have

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \widetilde{J}_n^{-1} \frac{1}{\sqrt{n}} S_n(\theta_0).$$

[11] More explicitly, we understand every subsequent equation as being multiplied by $\mathbb{1}_{A_n}$ on the left and right, or in other words, the equations only apply to those $\omega \in A_n$.

[12] Indeed, if $W_n \mathbb{1}_{\Omega_n} \xrightarrow{d} W$ for some W and events Ω_n with $\mathbb{P}(\Omega_n) \to 1$, then $(1 - \mathbb{1}_{\Omega_n})W_n \xrightarrow{p} 0$, so $W_n = W_n \mathbb{1}_{\Omega_n} + (1 - \mathbb{1}_{\Omega_n})W_n \xrightarrow{d} W$.

[13] A technical difficulty arises with applications of the mean value theorem here and elsewhere as the intermediate values $\widetilde{\theta}_n^{(j)}$ are not guaranteed to be measurable and hence are formally not necessarily random variables to which the usual rules of stochastic convergence can be applied. See `http://www.statslab.cam.ac.uk/~nickl/Site/__files/stat2013.pdf` for example for how this issue can be circumvented.

By the CMT, $\widetilde{J}_n^{-1} \xrightarrow{p} I_1(\theta_0)^{-1}$ and by the CLT, $\frac{1}{\sqrt{n}}S_n(\theta_0) \xrightarrow{d} N_p(0, I(\theta_0))$. Thus by Slutsky, we get the result. □

Lemma 14. *Suppose that regularity conditions (R1) and (R2) hold. Suppose that $\widetilde{\theta}_n^{(j)} \xrightarrow{p} \theta_0$ for $j = 1, \ldots, p$. Then the matrix $\widetilde{J}_n \in \mathbb{R}^{p\times p}$ with $(\widetilde{J}_n)_{jk} := (J_n(\widetilde{\theta}^{(j)}))_{jk}$ satisfies $\widetilde{J}_n \xrightarrow{p} I_1(\theta_0)$.*

Proof. Fix $j, k \in \{1, \ldots, p\}$ and define $g(X_i, \theta) := -\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(X_i, \theta)$. The conditions ensure that $\theta \mapsto \mathbb{E}_{\theta_0} g(X, \theta) =: \bar{g}(\theta)$ is continuous at θ_0 and that we have a ULLN of the form

$$\sup_{\theta\in\Theta}\left|\underbrace{\frac{1}{n}\sum_{i=1}^n g(X_i,\theta)}_{=(J_n(\theta))_{jk}} - \bar{g}(\theta)\right| \xrightarrow{p} 0.$$

Now

$$(\widetilde{J}_n)_{jk} - I_1(\theta_0)_{jk} = \left(\frac{1}{n}\sum_{i=1}^n g(X_i, \widetilde{\theta}_n^{(j)}) - \bar{g}(\widetilde{\theta}_n^{(j)})\right) + \left(\bar{g}(\widetilde{\theta}_n^{(j)}) - \bar{g}(\theta_0)\right).$$

The first term converges to 0 in probability by the ULLN, and the second term converges to 0 in probability by the CMT. Thus $(\widetilde{J}_n)_{jk} - I_1(\theta_0)_{jk} \xrightarrow{p} 0$ by Slutsky. □

1.7 Wald confidence intervals and tests

We can leverage the asymptotic normality of MLEs to quantify uncertainty through confidence intervals (or regions). Although the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ involves the Fisher information $I_1(\theta_0)$, which may be unknown as θ_0 is unknown, we may estimate $I_1(\theta_0)$ by $I_1(\widehat{\theta}_n)$ or $J_n(\widehat{\theta}_n)$. Provided $\theta \mapsto I_1(\theta)$ is continuous at θ_0, since $\widehat{\theta}_n \xrightarrow{p} \theta_0$, by the CMT, we have $I_1(\widehat{\theta}_n) \xrightarrow{p} I_1(\theta_0)$, and Lemma 14 shows in particular that $J_n(\widehat{\theta}_n) \xrightarrow{p} I_1(\theta_0)$.

One consequence of this is that by Slutsky's lemma, for any given $j \in \{1, \ldots, p\}$,

$$\frac{\sqrt{n}(\widehat{\theta}_{n,j} - \theta_{0,j})}{\sqrt{(J_n(\widehat{\theta}_n)^{-1})_{jj}}} \xrightarrow{d} N(0,1).$$

This leads to the *Wald* confidence interval for $\theta_{0,j}$ given by

$$\widehat{C}_n := \left[\widehat{\theta}_{n,j} - \frac{z_\alpha\sqrt{(J_n(\widehat{\theta}_n)^{-1})_{jj}}}{\sqrt{n}}, \widehat{\theta}_{n,j} + \frac{z_\alpha\sqrt{(J_n(\widehat{\theta}_n)^{-1})_{jj}}}{\sqrt{n}}\right].$$

By an analagous argument to that of Example 5, we have $\mathbb{P}_{\theta_0}(\widehat{\theta}_n \in \widehat{C}_n) \to 1 - \alpha$.

If, alternatively, we want to conduct inference for the whole parameter θ_0, we can base this on the following result.

Theorem 15 (Wald confidence region)**.** *Under regularity conditions,*

$$W_n(\theta_0) := n(\widehat{\theta}_n - \theta_0)^\top J_n(\widehat{\theta}_n)(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \chi^2_p.$$

Proof. We have

$$W_n(\theta_0) = n(\widehat{\theta}_n - \theta_0)^\top I_1(\theta_0)(\widehat{\theta}_n - \theta_0) + \sqrt{n}(\widehat{\theta}_n - \theta_0)^\top \{(J_n(\widehat{\theta}_n) - I_1(\theta_0)\}\sqrt{n}(\widehat{\theta}_n - \theta_0).$$

But we know $J_n(\widehat{\theta}_n) - I_1(\theta_0) \xrightarrow{p} 0$, so by Slutsky, $\{(J_n(\widehat{\theta}_n) - I_1(\theta_0)\}\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{p} 0$, and hence by Slutsky again, the second term above converges to 0 in probability.

For the first term, note that $I_1(\theta_0)^{1/2}\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N_p(0, I)$ by the CMT, so by the CMT again, $\{I_1(\theta_0)^{1/2}\sqrt{n}(\widehat{\theta}_n - \theta_0)\}^2 \xrightarrow{d} \chi^2_p$, so the result follows by a final application of Slutsky. □

This leads to an elliptical Wald confidence region of the form

$$\widehat{C}_n := \{\theta \in \Theta : W_n(\theta) \leq \xi_\alpha\},$$

where ξ_α is such that when $Z \sim \chi^2_p$, we have $\mathbb{P}(Z \geq \xi_\alpha) = \alpha$.

Exploiting the duality of confidence regions and tests, we can also use this approach to test the null hypothesis $H_0 : \theta = \theta_0$ for a given θ_0: we reject when $\theta_0 \notin \widehat{C}_n$, that is when

$$n(\widehat{\theta}_n - \theta_0)^\top J_n(\widehat{\theta}_n)(\widehat{\theta}_n - \theta_0) > \xi_\alpha.$$

Then the *Type I error* or *size* satisfies

$$\mathbb{P}_{\theta_0}(n(\widehat{\theta}_n - \theta_0)^\top J_n(\widehat{\theta}_n)(\widehat{\theta}_n - \theta_0) > \xi_\alpha) \to \alpha.$$

Note that in all of the above, we may replace $J_n(\widehat{\theta}_n)$ with $I_1(\theta_0)$ to obtain alternative versions of Wald tests and confidence regions and the conclusions remain unchanged.

1.8 Generalised likelihood ratio tests and score tests

The Wald approach is not the only way to perform hypothesis tests. Consider more generally the problem of testing

$$H_0 : \theta \in \Theta_0 \ \text{ against } \ H_1 : \theta \in \Theta \setminus \Theta_0.$$

where Θ_0 is some subset of $\Theta \subset \mathbb{R}^p$. Recall that the *generalised likelihood ratio statistic* $\Lambda_n := \Lambda_n(\Theta, \Theta_0)$ is given by

$$\Lambda_n(\Theta, \Theta_0) := 2 \log \left(\frac{\sup_{\theta \in \Theta} L_n(\theta)}{\sup_{\theta \in \Theta_0} L_n(\theta)} \right)$$

where L_n is the likelihood $L_n(\theta) = \prod_{i=1}^n f(X_i, \theta)$. Note that $\Lambda_n \geq 0$, and large values should indicate deviation from the null.

Theorem 16 (Wilks' theorem: special case $\Theta_0 = \{\theta_0\}$.). *Consider the special case where $\Theta_0 = \{\theta_0\}$. Under regularity conditions, we have*

$$\Lambda_n \xrightarrow{d} \chi^2_p.$$

Proof. Let $\widehat{\theta}_n$ be the MLE. We have that

$$\Lambda_n = 2\{\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0)\}.$$

We now Taylor expand[14] $\ell_n(\theta_0)$ about $\widehat{\theta}_n$ using a mean-value form of the remainder. As in the proof of Theorem 13, we work on the events $A_n := \{\widehat{\theta}_n \in \operatorname{int}\Theta\}$, which have probability converging to 1. We have

$$\ell_n(\theta_0) = \ell_n(\widehat{\theta}_n) + (\theta_0 - \widehat{\theta}_n)^\top \underbrace{S_n(\widehat{\theta}_n)}_{=0} - \frac{n}{2}(\theta_0 - \widehat{\theta}_n)^\top J_n(\widetilde{\theta}_n)(\theta_0 - \widehat{\theta}_n)$$

where $\widetilde{\theta}_n$ is on the closed line segment between θ_0 and $\widehat{\theta}_n$, so in particular $\widetilde{\theta}_n \xrightarrow{p} \theta_0$. Thus

$$\Lambda_n = \underbrace{\sqrt{n}(\theta_0 - \widehat{\theta}_n)^\top}_{\xrightarrow{d} N_p(0, I_1(\theta_0)^{-1})} \times \underbrace{J_n(\widetilde{\theta}_n)}_{\xrightarrow{p} I_1(\theta_0) \text{ (Lem. 14)}} \times \sqrt{n}(\theta_0 - \widehat{\theta}_n).$$

Just as in the proof of Theorem 15, we see that this converges in distribution to a χ^2_p. □

The result shows that rejecting the null when $\Lambda_n \geq \xi_\alpha$ gives a test with asymptotic size α. Moreover, we obtain the following asymptotically valid $(1-\alpha)$-level confidence set for θ:

$$\widehat{C}_n := \{\theta \in \Theta : \Lambda_n(\Theta, \{\theta\}) \leq \xi_\alpha\}.$$

One advantage of this test compared to the Wald test is that it does not require computation of $J_n(\widehat{\theta}_n)$ or $I_1(\widehat{\theta}_n)$: instead the test only involves evaluation of the likelihood at θ_0 and computation of $\widehat{\theta}_n$. In fact, there is also a test that avoids computation of the MLE altogether, which can be helpful in particular when p is large. The *score test* is based on the simple observation that under the null,

$$\frac{1}{\sqrt{n}} S_n(\theta_0) \xrightarrow{d} N_p(0, I_1(\theta_0)),$$

so by the CMT,

$$\lambda_n := \frac{1}{n} S_n(\theta_0)^\top I_1(\theta_0)^{-1} S_n(\theta_0) \xrightarrow{d} \chi^2_p.$$

Both Wilks' theorem and the score test can be generalised to settings with composite null hypotheses. For the score test, we replace θ_0 above with the MLE $\widetilde{\theta}_n$ under the null i.e. maximising ℓ_n only over Θ_0. In this setting, the limiting distribution of both Λ_n and λ_n become $\chi^2_{p-p_0}$ where $p_0 \leq p$ is the "dimension" or "degrees of freedom" of Θ_0. For example, if Θ_0 fixes the values of $k \leq p$ coordinates of θ, we will have $p_0 = p - k$.

[14] Note this is different from Theorem 13 where we instead expanded $S_n(\widehat{\theta}_n)$ about $S_n(\theta_0)$.

Informal summary: For **simple nulls** $H_0 : \theta = \theta_0$, we have the following test statistics:

$$\begin{aligned} \text{Wald:} \quad & n(\widehat{\theta}_n - \theta_0)^\top I_1(\widehat{\theta}_n)(\widehat{\theta}_n - \theta_0) \\ \text{Generalised likelihood ratio:} \quad & 2\{\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0)\} \\ \text{Score:} \quad & \frac{1}{n} S_n(\theta_0)^\top I_1(\theta_0)^{-1} S_n(\theta_0). \end{aligned}$$

Replacing $I_1(\theta_0)$ and $I_1(\widehat{\theta}_n)$ with any of $J_n(\widetilde{\theta}_n)$ or $I_1(\widetilde{\theta}_n)$ where $\widetilde{\theta}_n \in \{\theta_0, \widehat{\theta}_n\}$ will all yield the same asymptotic distribution under the null and may thus be used in the tests.

For **composite nulls** $H_0 : \theta \in \Theta_0$, we have

$$\begin{aligned} \text{Generalised likelihood ratio:} \quad & 2\{\ell_n(\widehat{\theta}_n) - \ell_n(\text{restricted MLE})\} \\ \text{Score:} \quad & \frac{1}{n} S_n(\text{restricted MLE})^\top I_1(\text{restricted MLE})^{-1} S_n(\text{restricted MLE}). \end{aligned}$$

The restricted MLE is $\text{argmax}_{\theta \in \Theta_0} \ell_n(\theta)$. Again, for the score test, there are several options that will yield the same asymptotic distribution under the null. These can be obtained by replacing the argument of $I_1(\text{restricted MLE})$ with $\widetilde{\theta}_n \in \{\text{restricted MLE}, \widehat{\theta}_n\}$, or using $J_n(\widetilde{\theta}_n)$.

For the particular composite null $H_0 : \theta_j = \theta_{0,j}$, we can also use a α-level Wald test which rejects when

$$\sqrt{n} \frac{|\widehat{\theta}_{n,j} - \theta_{0,j}|}{\sqrt{(I_1(\widehat{\theta}_n)^{-1})_{jj}}} > \Phi^{-1}(1 - \alpha/2),$$

where as indicated above, $I_1(\widehat{\theta}_n)$ can be replaced by several other quantities to yield the same asymptotic distribution under the null.

1.9 The Delta method

Consider now the problem of estimating a certain function $\phi(\theta)$ of the parameter θ in the model $\{f(\cdot, \theta) : \theta \in \Theta\}$. We first look at the special case where $\phi(\theta) = \theta_1$ and $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2 = \Theta$. One option is to maximise the *profile likelihood*

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

More generally, one can maximise the *induced likelihood function* $L^*(\psi) := \sup_{\theta \in \Theta : \phi(\theta) = \psi} L(\theta)$ over ψ. A conceptually simpler approach is to compute the MLE $\widehat{\theta}$ and report the so-called *plug-in MLE* $\phi(\widehat{\theta})$. It turns out, these two approaches amount to the same thing.

Proposition 17. *Let $\widehat{\theta}$ be an MLE. Then $\phi(\widehat{\theta})$ maximises $L^*(\psi)$ over $\psi \in \phi(\Theta) := \{\psi : \psi = \phi(\theta) \text{ for some } \theta \in \Theta\}$.*

Proof. Suppose for a contradiction there exists $\widehat{\psi} \in \phi(\Theta)$ with $L^*(\widehat{\psi}) > L^*(\phi(\widehat{\theta})) + \epsilon$ for some $\epsilon > 0$. Then there exists $\widetilde{\theta}$ such that $L(\widetilde{\theta}) > L^*(\widehat{\psi}) - \epsilon > L^*(\phi(\widehat{\theta})) \geq L(\widehat{\theta})$, contradicting the optimality of $\widehat{\theta}$. □

For example, if we reparametrise the $\{N(\mu,\sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$ model as $\{N(\mu, 1/\gamma^2) : \mu \in \mathbb{R}, \gamma^2 > 0\}$, the MLE for the precision γ^2 will be the reciprocal of the MLE for σ^2.

The Delta method, which is a general procedure for finding the limiting distribution of a function of an estimator based on knowledge of the limiting distribution of the estimator itself, can allow us to conduct inference on $\phi(\theta)$.

Theorem 18 (Delta method)**.** *Let $\phi : \Theta \to \mathbb{R}^m$ be continuously differentiable at θ_0. Let $\widehat{\theta}_n$ be a sequence of random vectors (estimators—not necessarily MLEs) such that $r_n(\widehat{\theta}_n - \theta_0) \xrightarrow{d} Z$ where Z is some random vector and $r_n \to \infty$ is some deterministic scalar sequence (e.g. $r_n = \sqrt{n}$). Then*

$$r_n(\phi(\widehat{\theta}_n) - \phi(\theta_0)) \xrightarrow{d} \begin{pmatrix} \nabla_\theta \phi_1(\theta_0)^\top \\ \vdots \\ \nabla_\theta \phi_m(\theta_0)^\top \end{pmatrix} Z.$$

Proof. By the mean value theorem applied to each component,

$$r_n(\phi(\widehat{\theta}_n) - \phi(\theta_0)) = \underbrace{\begin{pmatrix} \nabla_\theta \phi_1(\widetilde{\theta}_n^{(1)})^\top \\ \vdots \\ \nabla_\theta \phi_m(\widetilde{\theta}_n^{(m)})^\top \end{pmatrix}}_{=:D_n} r_n(\widehat{\theta}_n - \theta_0),$$

for some $\widetilde{\theta}_n^{(k)}$, $k = 1, \dots, m$ in the closed line segment between θ_0 and $\widehat{\theta}_n$. Now $\widehat{\theta}_n \xrightarrow{p} \theta_0$ (see Example Sheet 1, Question 8(b)), so $\widetilde{\theta}_n^{(k)} \xrightarrow{p} \theta_0$ also. But then by the CMT, $\nabla_\theta \phi_k(\widetilde{\theta}_n^{(k)}) \xrightarrow{p} \nabla_\theta \phi_k(\theta_0)$ for each k so by Slutsky,

$$D_n \underbrace{r_n(\widehat{\theta}_n - \theta_0)}_{\xrightarrow{d} Z} \xrightarrow{d} \begin{pmatrix} \nabla_\theta \phi_1(\theta_0)^\top \\ \vdots \\ \nabla_\theta \phi_m(\theta_0)^\top \end{pmatrix} Z. \qquad \square$$

Considering the case where $\widehat{\theta}_n$ is the MLE and $m = 1$, we have

$$\sqrt{n}(\phi(\widehat{\theta}_n) - \phi(\theta_0)) \xrightarrow{d} N(0, \nabla_\theta \phi(\theta_0)^\top I_1(\theta_0)^{-1} \nabla_\theta \phi(\theta_0)).$$

Recall that the Cramér–Rao lower bound for an estimator $\widehat{\phi}$ with $\mathbb{E}_\theta \widehat{\phi} = \phi(\theta)$ is

$$n^{-1} \nabla_\theta \phi(\theta_0)^\top I_1(\theta_0)^{-1} \nabla_\theta \phi(\theta_0),$$

so this 'matches' what we see in the asymptotic distribution of the plug-in MLE.

Example 6. Suppose we have i.i.d. data $X_1, \dots, X_n$ with mean μ_0 and variance 1 and we wish to estimate μ_0^2. We have $\sqrt{n}(\bar{X}_n - \mu_0) \xrightarrow{d} N(0,1)$ by the CLT, so by the Delta method

$$\sqrt{n}(\bar{X}_n^2 - \mu_0^2) \xrightarrow{d} N(0, 4\mu_0^2).$$

What happens if $\mu_0 = 0$? Then we simply have convergence in probability to 0 above. To obtain a more informative result, we should consider the limiting distribution of

$$n\bar{X}_n^2 = n\left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\right)^2 \xrightarrow{d} \chi_1^2 \quad \text{(by the CMT).}$$

1.10 Beyond maximum likelihood

It may appear that maximum likelihood estimation has largely 'solved' the essential problem of learning from data that Statistics is concerned with: MLEs enjoy a form of optimality in the form of efficiency, and the fact that they are asymptotically Gaussian means that inferential questions can be answered with confidence statements and hypothesis tests that have formal asymptotic justifications. What remains to be done?

Of course there are settings where the regularity conditions we have employed may fail (such as when θ_0 is at the boundary of Θ_0 or when the support of the distributions varies with θ—see the example sheet), but this is certainly not the biggest limitation.

The justification of MLEs relied on what turned out to be an extremely powerful idea: we regarded our estimator $\widehat{\theta}_n$ applied to data $X_1, \ldots, X_n$ as embedded within an infinite sequence of estimators, and aimed to understand properties of $\widehat{\theta}_n$ by understanding its limiting behaviour. Therein however lies a fundamental weakness of our entire analysis. We have not put forward any guarantees on the behaviour of MLEs at a *finite* sample size n. This gap in our argument opens the door to other inference strategies that may be superior, at least in some ways, in such finite samples. One possibility for improvement is to leverage any vague prior information we may have about the parameter of interest.

2 Bayesian inference

Bayesian inference is an approach to inference based on regarding the parameter of interest as random, and specifying a *prior distribution* for this. This prior distribution can represent (subjective) beliefs about the parameter. A more broad perspective however would regard the methods resulting from thinking in this way as precisely that: methods, which we can assess in the same sort of way as other inference methods.

Example 7. Consider a simple model where $\Theta = \{\theta_1, \theta_2\}$. We regard our target parameter θ as a random variable taking values in Θ with prior probabilities $\pi_j := \mathbb{P}(\theta = \theta_j)$. Let f_j be the pmf of our data $X \in \mathcal{X}$ (considered discrete here for transparency) given that

$\theta = \theta_j$. Then if x is our realised data,

$$\begin{aligned}\mathbb{P}(\theta = \theta_j \,|\, X = x) &= \frac{\mathbb{P}(X = x \text{ and } \theta = \theta_j)}{\mathbb{P}(X = x)} \\ &= \frac{\mathbb{P}(X = x \,|\, \theta = \theta_j)\mathbb{P}(\theta = \theta_j)}{\mathbb{P}(X = x \,|\, \theta = \theta_1)\mathbb{P}(\theta = \theta_1) + \mathbb{P}(X = x \,|\, \theta = \theta_2)\mathbb{P}(\theta = \theta_2)} \\ &= \frac{\pi_j f_j(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}.\end{aligned}$$

Thus we should prefer the hypothesis $H_0 : \theta = \theta_1$ over $H_1 : \theta = \theta_2$ if

$$\frac{\mathbb{P}(\theta = \theta_1 \,|\, X = x)}{\mathbb{P}(\theta = \theta_2 \,|\, X = x)} = \frac{f_1(x)}{f_2(x)} \frac{\pi_1}{\pi_2} > 1.$$

To justify this more formally, we can consider an arbitrary approach $\delta : \mathcal{X} \to \Theta$ for deciding between θ_1 and θ_2 based on data X which incurs a loss of 1 when we make the wrong decision. We will return to this idea of measuring the quality of an estimator (or more generally a decision making process) in the next chapter, but for now, note that the expected loss satisfies

$$\begin{aligned}\mathbb{E}\mathbb{1}_{\{\delta(X) \neq \theta\}} &= \mathbb{E}\left(\mathbb{1}_{\{\delta(X)=\theta_2\}}\mathbb{1}_{\{\theta=\theta_1\}} + \mathbb{1}_{\{\delta(X)=\theta_1\}}\mathbb{1}_{\{\theta=\theta_2\}}\right) \\ &= \mathbb{E}\left(\mathbb{E}(\mathbb{1}_{\{\delta(X)=\theta_2\}}\mathbb{1}_{\{\theta=\theta_1\}} + \mathbb{1}_{\{\delta(X)=\theta_1\}}\mathbb{1}_{\{\theta=\theta_2\}} \,|\, X)\right) \\ &= \mathbb{E}\left(\mathbb{1}_{\{\delta(X)=\theta_2\}}\mathbb{P}(\theta = \theta_1 \,|\, X) + \mathbb{1}_{\{\delta(X)=\theta_1\}}\mathbb{P}(\theta = \theta_2 \,|\, X)\right).\end{aligned}$$

To minimise this then $\delta(X)$ should always pick the hypothesis preferred by the rule above.

In summary, we see that if the prior on θ genuinely described our uncertainty about θ, then our inference on θ should be based on the posterior distribution given by $\{\mathbb{P}(\theta = \theta_1 \,|\, X = x), \mathbb{P}(\theta = \theta_2 \,|\, X = x)\}$.

Let us put the ideas above in a more general setting. In the Bayesian context, specifying a statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$ means that the distribution of data X given θ is

$$X \,|\, \theta \sim f(\cdot, \theta).$$

We also specify a *prior density* π which is the marginal density of θ. The *posterior density* is given by

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

We also regard the posterior as a random probability density function $\Pi(\theta \,|\, X)$[15]. As usual, when either X or θ are discrete, the densities above should be thought of as pmfs and the

[15] Note that the randomness is coming from the data X (even though we are thinking of θ as a random variable). There is some abuse of notation here: usually for e.g. a random variable X with density f, we would not also use X to denote a deterministic point where we might evaluate its density and would instead write $f(x)$. In the Bayesian context however, it is common to see with the parameter of interest θ.

integrals replaced with sums. Typically, we will deal with settings where we observe copies $X_1, \dots, X_n$ of X where $X \,|\, \theta \sim f(\cdot, \theta)$ that are *conditionally* i.i.d. *given θ*. We write this as $X_1, \dots, X_n \,|\, \theta \overset{\text{i.i.d.}}{\sim} f(\cdot, \theta)$. Note that the $X_1, \dots, X_n$ will typically not be *marginally* independent as they 'share' the same θ. Our posterior is

$$\theta \,|\, X_1, \dots, X_n \sim \Pi(\theta \,|\, x_1, \dots, x_n) = \frac{\pi(\theta) \prod_{i=1}^n f(x_i, \theta)}{\int_\Theta \pi(\theta') \prod_{i=1}^n f(x_i, \theta') \, d\theta'}.$$

The expression above can be viewed as a reweighted version of the likelihood function $L_n(\theta)$. Note that the denominator is simply a normalising factor and is constant in θ. It can often be ignored in calculations by spotting the form of the distribution.

Example 8. Suppose $X_1, \dots, X_n \,|\, \theta \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$ with prior $\theta \sim N(0, 1)$. The numerator of the posterior is proportional to (as a function of θ)

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

Thus we see that

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

In the example above, both the prior and the posterior were in the same distributional family (they were both normal). This motivates the following definition.

Definition 7. In a statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$, if the prior $\pi(\theta)$ and the posterior $\Pi(\theta|X)$ belong to the same family of distributions, the prior is called a *conjugate* prior.

Other examples (see example sheet) of conjugacy include:

- Beta prior and binomial sampling;

- Gamma prior and Poisson sampling.

The posterior can be used in several ways for performing inference about θ:

- **Estimation:** We can use the posterior mean

  $$\bar{\theta} = \bar{\theta}(X) = \mathbb{E}(\theta \,|\, X) = \int_{\theta \in \Theta} \theta \Pi(\theta \,|\, X) \, d\theta$$

  for example, or another summary of the posterior such as the mode or median.

- **Uncertainty quantification:** Any subset $\widehat{C} = \widehat{C}(X) \subseteq \Theta$ such that

$$\int_{\widehat{C}} \Pi(\theta \,|\, X)\, d\theta = \mathbb{P}(\theta \in \widehat{C} \,|\, X) = 1 - \alpha$$

  is a $(1-\alpha)$-level *credible set* for θ.

- **Hypothesis testing:** As in Example 7, we can decide between hypotheses $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ via the *Bayes factor*

$$\frac{\mathbb{P}(\theta \in \Theta_0 \,|\, X)}{\mathbb{P}(\theta \in \Theta_1 \,|\, X)} = \frac{\int_{\Theta_0} f(X,\theta)\pi(\theta)\, d\theta}{\int_{\Theta_1} f(X,\theta)\pi(\theta)\, d\theta}.$$

2.1 Uninformative priors

We motivated the Bayesian approach through a desire to leverage prior information for inference. In many situations, such prior information cannot be easily summarised in the form of a prior probability distribution over the parameter of interest. It is nevertheless interesting to look at Bayesian methods in this context as well. What sort of prior would be sensible to use in such a setting?

Consider the case where $X \,|\, \theta \sim \mathrm{Bern}(\theta)$. It would appear that the only sensible choice of 'ignorant' prior in this case is $\theta \sim U[0,1]$. However the principle to represent ignorance by uniform priors on the parameter space is logically flawed. Indeed, we could reparametrise our model via $\psi = \theta^{100}$. The implied prior on ψ would then be

$$\pi^{(\psi)}(\psi) = \pi(\theta(\psi)) \cdot \left| \frac{d\theta(\psi)}{d\psi} \right| = \frac{1}{100}\psi^{-99/100},$$

which seems *in*formative as it puts much more mass close to 0 than 1. Therefore the principle of using uniform priors is not invariant to reparametrisations.

To achieve this form of invariance, we can use the Jeffreys prior:

Definition 8. The prior $\pi(\theta)$ proportional to $\sqrt{\det I(\theta)}$ is called the *Jeffreys prior.*

Note that it may be the case that then $\int_\Theta \pi(\theta)\, d\theta = \infty$: any such a prior is called *improper*. Although the prior then would not represent a probability distribution, the posterior

$$\frac{f(x,\theta)\pi(\theta)}{\int_\Theta f(x,\theta')\pi(\theta')\, d\theta'}$$

may still be a well-defined probability density (though it cannot be interpreted as a conditional density). To see how the Jeffreys prior restores the desired invariance, consider for simplicity the case where $p = 1$ and observe that under regularity conditions, the Fisher information $I^{(\psi)}(\psi)$ in the ψ parametrisation satisfies

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

as $\mathbb{E}_\psi(S(\theta(\psi)) = 0$. Thus with the Jeffreys prior,

$$\sqrt{\det I^{(\psi)}(\psi)} \propto \pi(\theta(\psi))\left|\frac{d\theta(\psi)}{d\psi}\right|$$

2.2 Frequentist analysis of Bayesian methods

Particularly when using an uninformative prior, it is hard to defend a dogma that all inference should be based on the posterior without other justification. This motivates studying Bayesian methods from a 'frequentist' perspective, that is by studying their behaviour on average over hypothetical repetitions of the 'experiment' used to produce the data. This is just a fancy way of describing what we have been doing all along with e.g. checking whether a given confidence region contains the true parameter at least $1-\alpha$ of the time 'on average'.

Example 8 continued. Recall that when $X_1,\ldots,X_n \,|\, \theta \overset{\text{i.i.d.}}{\sim} N(\theta,1)$ with $\theta \sim N(0,1)$, we have that the posterior mean

$$\bar{\theta}_n := \mathbb{E}(\theta \,|\, X_1,\ldots,X_n) = \frac{n}{n+1}\bar{X}_n.$$

This is not exactly the MLE $\widehat{\theta}_n := \bar{X}_n$, but is close. Consider now the setting where $X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim} N(\theta_0,1)$ for a deterministic $\theta_0 \in \mathbb{R}$. Under this sampling scheme, we have $\sqrt{n}(\widehat{\theta}_n - \bar{\theta}_n) \xrightarrow{p} 0$, so by Slutsky, their limiting distributions are identical. Moreover, a credible set of the form

$$\widehat{C}_n := \left\{\theta' : |\theta' - \bar{\theta}| \le \frac{R_n}{\sqrt{n}}\right\},$$

where R_n is taken such that $\int_{\widehat{C}_n} \Pi(\theta \,|\, X_1,\ldots,X_n)\,d\theta = 1-\alpha$ will share the frequentist coverage guarantee $\mathbb{P}_{\theta_0}(\theta_0 \in \widehat{C}_n) \to 1-\alpha$ of the standard (Wald) confidence interval centred at the MLE (see Example Sheet). (Note that R_n is a random variable that depends on the data $X_1,\ldots,X_n$.)

In the above example, we observed a close relationship between likelihood-based and Bayesian inference. Remarkably, this asymptotic equivalence persists across all sufficiently regular models and priors. To see why such a result could hold, suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f(\cdot, \theta_0)$ and that the prior π is continuous and positive at θ_0. Consider reparametrising the posterior in terms of $h = \sqrt{n}(\theta - \widehat{\theta}_n)$. We obtain

$$\begin{aligned}\Pi(\theta(h) \,|\, X_1, \dots, X_n) \frac{1}{\sqrt{n}} &=: \widetilde{\Pi}(h \,|\, X_1, \dots, X_n) \\ &= \frac{L_n(\widehat{\theta}_n + h/\sqrt{n}) \pi(\widehat{\theta}_n + h/\sqrt{n})}{\int L_n(\widehat{\theta}_n + h'/\sqrt{n}) \pi(\widehat{\theta}_n + h'/\sqrt{n}) \, dh'} \\ &\propto \frac{L_n(\widehat{\theta}_n + h/\sqrt{n})}{L_n(\widehat{\theta}_n)} \pi(\widehat{\theta}_n + h/\sqrt{n}).\end{aligned}$$

Now, for any fixed h, we have, arguing as in the proof of Wilks theorem,

$$\begin{aligned}\log\left(\frac{L_n(\widehat{\theta}_n + h/\sqrt{n})}{L_n(\widehat{\theta}_n)}\right) &= \ell_n(\widehat{\theta}_n + h/\sqrt{n}) - \ell_n(\widehat{\theta}_n) \\ &= -\frac{1}{2} h^\top J_n(\widetilde{\theta}_n) h\end{aligned}$$

where $\widetilde{\theta}_n = \widehat{\theta}_n + th/\sqrt{n}$, some $t \in [0, 1]$. By consistency of the MLE and Lemma 14, $J_n(\widetilde{\theta}_n) \xrightarrow{p} I_1(\theta_0)$, so by the CMT, we see that

$$\frac{L_n(\widehat{\theta}_n + h/\sqrt{n})}{L_n(\widehat{\theta}_n)} \pi(\widehat{\theta}_n + h/\sqrt{n}) \xrightarrow{p} \exp\left(-\frac{1}{2} h^\top I_1(\theta_0) h\right) \pi(\theta_0).$$

If we could additionally show that the integrals over h of the two sides above converged in probability, dividing by the normalising constants and appealing to Slutsky, we would have

$$\widetilde{\Pi}(h \,|\, X_1, \dots, X_n) \xrightarrow{p} \widetilde{\phi}(h)$$

for each h, where $\widetilde{\phi}$ is the $N_p(0, I_1(\theta_0)^{-1})$ density. It turns out this can be strengthened to

$$\int |\widetilde{\Pi}(h \,|\, X_1, \dots, X_n) - \widetilde{\phi}(h)| \, dh \xrightarrow{a.s.} 0,$$

or, in the original parametrisation $\theta = \widehat{\theta}_n + h/\sqrt{n}$,

$$\int |\Pi(\theta \,|\, X_1, \dots, X_n) - \widehat{\phi}_n(\theta)| \, d\theta \xrightarrow{a.s.} 0,$$

where $\widehat{\phi}_n$ is the (random) $N_p(\widehat{\theta}_n, I_n(\theta_0)^{-1})$ density. This result, which holds under relatively mild regularity conditions, is known as the *Bernstein-von Mises theorem*[16].

[16] See `http://www.statslab.cam.ac.uk/~nickl/Site/__files/stat2013.pdf` for a detailed proof.

Proposition 19. *Consider the above setup (assuming the Bernstein-von Mises theorem holds) with $p = 1$. Let*

$$\widehat{C}_n := \left\{\theta' : |\widehat{\theta}_n - \theta'| \leq \frac{R_n}{\sqrt{n}}\right\},$$

where for $\alpha \in (0, 1)$, R_n is chosen such that $\int_{\widehat{C}_n} \Pi(\theta \,|\, X_1, \ldots, X_n)\, d\theta = 1 - \alpha$. Then $\mathbb{P}_{\theta_0}(\theta_0 \in \widehat{C}_n) \to 1 - \alpha$.

Proof. First observe that R_n satisfies

$$\int_{-R_n}^{R_n} \widetilde{\Pi}(h \,|\, X_1, \ldots, X_n)\, dh = 1 - \alpha.$$

But

$$\begin{aligned}\left|\int_{-R_n}^{R_n} \widetilde{\Pi}(h \,|\, X_1, \ldots, X_n)\, dh - \int_{-R_n}^{R_n} \widetilde{\phi}(h)\, dh\right| &\leq \int |\widetilde{\Pi}(h \,|\, X_1, \ldots, X_n) - \widetilde{\phi}(h)|\, dh \\ &\overset{a.s.}{\to} 0\end{aligned}$$

by the Bernstein-von Mises theorem. Thus, writing

$$\widetilde{\Phi}(t) := \int_{-t}^{t} \widetilde{\phi}(h)\, dh,$$

we have $\widetilde{\Phi}(R_n) \overset{a.s.}{\to} 1 - \alpha$. Now $\widetilde{\Phi} : (0, \infty) \to (0, \infty)$ is continuous and strictly increasing, so it has a continuous inverse $\widetilde{\Phi}^{-1} : (0, \infty) \to (0, \infty)$. Hence by the CMT, $R_n \overset{a.s.}{\to} \widetilde{\Phi}^{-1}(1 - \alpha)$. By Slutsky,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \frac{\widetilde{\Phi}^{-1}(1 - \alpha)}{R_n} \overset{d}{\to} \widetilde{Z}$$

where $\widetilde{Z} \sim N(0, I_1(\theta_0)^{-1})$. Hence

$$\begin{aligned}\mathbb{P}_{\theta_0}(\theta_0 \in \widehat{C}_n) &= \mathbb{P}_{\theta_0}\left(-\widetilde{\Phi}^{-1}(1 - \alpha) \leq \sqrt{n}(\widehat{\theta}_n - \theta_0) \frac{\widetilde{\Phi}^{-1}(1 - \alpha)}{R_n} \leq \widetilde{\Phi}^{-1}(1 - \alpha)\right) \\ &\to \mathbb{P}(-\widetilde{\Phi}^{-1}(1 - \alpha) \leq \widetilde{Z} \leq \widetilde{\Phi}^{-1}(1 - \alpha)) = 1 - \alpha.\end{aligned}$$

□

Overall we see that Bayesian methods enjoy the same favourable asymptotic properties as likelihood based inference. Of course, as discussed at the end of the last chapter, optimal asymptotic properties need not translate to optimal finite sample performance. In the next chapter we therefore turn to the issue of finite sample performance of estimators and study this from first principles.

3 Decision theory

Given a statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$ and data $X \in \mathcal{X}$, we can cast many statistical tasks as *decision problems*, where our goal is to find an appropriate *decision rule* $\delta : \mathcal{X} \to \mathcal{A}$, where $\mathcal{A}$ is a set of *actions*.

(i) **Hypothesis testing:** we can take $\mathcal{A} = \{0, 1\}$ with δ a test function.

(ii) **Estimation:** $\mathcal{A} = \Theta$ and $\delta(X) = \widehat{\theta}(X)$ is an estimator.

To measure the quality of an action, we can use a *loss function*

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

For example:

(i) **Hypothesis testing:** for testing simple hypotheses we may take $L(a, \theta) = \mathbb{1}_{\{a \neq \theta\}}$.

(ii) **Estimation:** $L(a, \theta) = (a - \theta)^2$ or $L(a, \theta) = |a - \theta|$ in one dimension.

To assess the performance of a decision rule, we can consider its average loss or *risk*:

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

(i) **Hypothesis testing:** $R(\delta, \theta) = \mathbb{P}_\theta(\delta(X) \neq \theta)$ is either the probability of a type I error or a type II error, depending on the value of θ.

(ii) **Estimation:** The *quadratic risk* is also known as the *mean squared error* (MSE)

$$R(\widehat{\theta}, \theta) = \mathbb{E}_\theta(\widehat{\theta} - \theta)^2.$$

Example 9. Consider estimating θ in a $\text{Bin}(n, \theta)$ model where $\theta \in [0, 1]$ under quadratic risk. The MLE $\widehat{\theta}(X) = X/n$ satisfies

$$R(\widehat{\theta}, \theta) = \text{Var}_\theta(\widehat{\theta}) = \frac{\theta(1 - \theta)}{n}.$$

On the other hand, the (naive) estimator $\widetilde{\theta} = 1/2$ has

$$R(\widetilde{\theta}, \theta) = (\theta - 1/2)^2,$$

which is then apparently preferable when θ is sufficiently close to $1/2$.

3.1 Bayes risk

One issue with using risk to compare the quality of two decision rules is that one must fix a value of $\theta \in \Theta$ at which the comparison is to be made. We can instead average this risk over different θ values.

Definition 9. Given a prior density π on Θ and model $\{f(\cdot, \theta) : \theta \in \Theta\}$, the π-Bayes risk of a decision rule δ is

$$R_\pi(\delta) := \int_\Theta R(\delta, \theta)\pi(\theta)\, d\theta = \mathbb{E}(L(\delta(X), \theta)),$$

where in the final equality both θ and X are random, $X \,|\, \theta \sim f(\cdot, \theta)$, and the marginal density of θ is π. Any minimiser of the π-Bayes risk is called a *π-Bayes rule.*

Example 9 continued. Consider the uniform prior $\pi = \mathbb{1}_{[0,1]}$. We have

$$R_\pi(\widehat{\theta}) = \frac{1}{n}\int_0^1 \theta(1-\theta)\, d\theta = \frac{1}{6n}.$$

On the other hand,

$$R_\pi(1/2) = \int_0^1 (\theta - 1/2)^2\, d\theta = \frac{1}{3} \times 2 \times \frac{1}{2^3} = \frac{1}{12}.$$

How can one find π-Bayes rules? Observe that

$$R_\pi(\delta) = \mathbb{E}[\mathbb{E}\{L(\delta(X), \theta) \,|\, X\}].$$

Thus to minimise the π-Bayes risk over δ, it suffices to set $\delta(x)$ for each $x \in \mathcal{X}$ to be the minimiser of the *posterior risk*

$$\mathbb{E}[L(\delta(x), \theta) \,|\, X = x] = \int_\Theta L(\delta(x), \theta)\Pi(\theta \,|\, x)\, d\theta.$$

Writing $\delta_\Pi : \mathcal{X} \to \mathcal{A}$ for this minimiser (assumed to be unique), we also have conversely that any π-Bayes rule δ must satisfy

$$\mathbb{P}(\delta(X) = \delta_\Pi(X)) = 1.$$

(Note that in the above, X follows its marginal distribution $\int f(x, \theta)\pi(\theta)\, d\theta$.) Indeed, we must have

$$R_\pi(\delta) - R_\pi(\delta_\Pi) = \mathbb{E}[\underbrace{\mathbb{E}\{L(\delta(X), \theta) \,|\, X\} - \mathbb{E}\{L(\delta_\Pi(X), \theta) \,|\, X\}}_{\geq 0}] = 0.$$

But a fact from *Probability and Measure* tells us that if U is a non-negative random variable, then $\mathbb{E}U = 0$ if and only if $\mathbb{P}(U = 0) = 1$.

Example 10. The posterior quadratic risk is minimised by the posterior mean. Indeed, fixing $x \in \mathcal{X}$ and writing $\bar{\theta} = \bar{\theta}(x)$ for the posterior mean, we have

$$\begin{aligned}\mathbb{E}[(\delta - \theta)^2 \,|\, X = x] &= \mathbb{E}[(\delta - \bar{\theta} + \bar{\theta} - \theta)^2 \,|\, X = x] \\ &= (\delta - \bar{\theta})^2 + \mathbb{E}[(\bar{\theta} - \theta)^2 \,|\, X = x].\end{aligned}$$

The following result shows that the property of unbiasedness and that of being π-Bayes for the quadratic risk, are largely incompatible.

Proposition 20. *Suppose $\widehat{\theta} = \widehat{\theta}(X)$ is an unbiased estimator of θ, so $\mathbb{E}_\theta(\widehat{\theta}) = \mathbb{E}(\widehat{\theta} \,|\, \theta) = \theta$. If $\widehat{\theta}$ is also π-Bayes, for some prior π in the quadratic risk, then*

$$\mathbb{P}(\widehat{\theta} = \theta) = \int \mathbb{1}_{\{\widehat{\theta}(x) = \theta'\}} f(x, \theta') \pi(\theta') \, d\theta' dx = 1.$$

Proof. It suffices to show that

$$\mathbb{E}(\widehat{\theta} - \theta)^2 = \mathbb{E}(\widehat{\theta^2}) - 2\mathbb{E}(\theta\widehat{\theta}) + \mathbb{E}(\theta^2) = 0. \tag{3.1}$$

But

$$\mathbb{E}(\theta\widehat{\theta}) = \mathbb{E}\{\theta \mathbb{E}(\widehat{\theta} \,|\, \theta)\} = \mathbb{E}(\theta^2),$$

and also $\widehat{\theta} = \mathbb{E}(\theta \,|\, X)$ almost surely, so

$$\mathbb{E}(\theta\widehat{\theta}) = \mathbb{E}\{\theta \mathbb{E}(\theta \,|\, X)\} = \mathbb{E}[\mathbb{E}\{\theta \mathbb{E}(\theta \,|\, X) \,|\, X\}] = \mathbb{E}(\widehat{\theta^2}).$$

We therefore see that (3.1) holds. □

3.2 Minimax risk

While the Bayes risk removes the ambiguity of fixing on a particular value of θ, a criticism of this approach for comparing estimators could be that it involves having to fix a prior distribution, which may be just as problematic. An alternative is to consider the worst case risk $R(\delta, \theta)$ over all values of θ.

Example 9 continued. Recall that when $X \sim \mathrm{Bin}(n, \theta)$, the MLE $\widehat{\theta} = X/n$. We have

$$\sup_{\theta \in [0,1]} R(\widehat{\theta}, \theta) = \sup_{\theta \in [0,1]} \frac{\theta(1 - \theta)}{n} = \frac{1}{4n}.$$

On the other hand,

$$\sup_{\theta \in [0,1]} R(1/2, \theta) = \frac{1}{4}.$$

Definition 10. The *minimax risk* is the infimum (over all possible decision rules) of the maximal risk over the parameter space Θ:

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

A decision rule δ that has maximal risk $\sup_{\theta \in \Theta} R(\delta, \theta)$ attaining the minimax risk is said to be *minimax.*

Perhaps surprisingly, Bayes rules can be helpful for finding minimax rules.

Theorem 21. *Let π be a prior on Θ and suppose δ_π is a decision rule such that*

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

(In particular this will occur if δ_π has constant risk $R(\delta_\pi, \theta)$ in θ.) If δ_π is a (unique) π-Bayes rule, then it is a (unique) minimax rule.

Proof. We have that for any decision rule δ,

$$\sup_{\theta \in \Theta} R(\delta, \theta) \geq \int_\Theta R(\delta, \theta)\pi(\theta)\, d\theta = R_\pi(\delta) \geq R_\pi(\delta_\pi) = \sup_{\theta \in \Theta} R(\delta_\pi, \theta),$$

so $\inf_\delta \sup_{\theta \in \Theta} R(\delta, \theta) = \sup_{\theta \in \Theta} R(\delta_\pi, \theta)$. If δ_π is a unique π-Bayes rule, then the second inequality above would be strict for $\delta \neq \delta_\pi$, so no such δ can have maximal risk equal to the minimax risk. □

A prior satisfying the hypothesis of the theorem is necessarily a 'worst-case' prior in the following sense:

Definition 11. A prior π is *least favourable* if given a π-Bayes estimator δ_π, for any other prior λ and λ-Bayes estimator δ_λ, we have $R_\lambda(\delta_\lambda) \leq R_\pi(\delta_\pi)$.

Indeed, then

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

with the last equality following by assumption.

Example 9 continued. One can show that $\widehat{\theta}$ is not minimax. Instead we may take a Beta(a, b) prior $\pi_{a,b}$ on $\theta \in [0, 1]$ and writing $\bar{\theta}_{a,b}$ for the posterior mean, solve the set of equations

$$R(\bar{\theta}_{a,b}, \theta) = \text{const.} \qquad \text{for all } \theta \in [0, 1].$$

This will yield a unique Bayes rule with constant risk, which is thus the unique minimax rule (see Example Sheet).

3.3 Admissibility

We motivated the notions of a π-Bayes risk and the maximum risk as ways of addressing the fact that there is no clear way of ordering the risk functions $\theta \mapsto R(\delta, \theta)$ of different decision rules in order of preference. There is however a natural partial order among risk functions:

Definition 12. A decision rule δ is *inadmissible* if there exists another decision rule δ' that *dominates* δ in the sense that

$$R(\delta', \theta) \leq R(\delta, \theta) \quad \text{for all } \theta \in \Theta \qquad \text{and} \qquad R(\delta', \theta) < R(\delta, \theta) \quad \text{for some } \theta \in \Theta.$$

Otherwise δ is *admissible.*

An estimator being admissible does not necessarily mean it is a 'sensible' estimator. Indeed any constant estimator is admissible. On the other hand, if an estimator is inadmissible, any estimator that dominates it should always be preferred.

Note that a decision rule being minimax does not guarantee admissibility, a fact which underlines how in summarising the risk function by taking the maximum value, some information has been lost. However:

Proposition 22. *If for a prior π the π-Bayes rule is unique, then it is admissible.*

Proof. Let δ_π be π-Bayes and suppose decision rule δ satisfies $R(\delta, \theta) \leq R(\delta_\pi, \theta)$ for all $\theta \in \Theta$. Then

$$R_\pi(\delta) = \int_\Theta R(\delta, \theta)\pi(\theta)\, d\theta \leq \int_\Theta R(\delta_\pi, \theta)\pi(\theta)\, d\theta = R_\pi(\delta_\pi)$$

so δ is π-Bayes and hence $\delta = \delta_\pi$ by uniqueness. □

The following provides a helpful connection between admissibility and minimaxity:

Proposition 23. *If decision rule δ is admissible and has constant risk, it is minimax.*

Proof. If δ is not minimax, then there exists δ' with

$$\sup_\theta R(\delta', \theta) < \sup_\theta R(\delta, \theta) = \inf_\theta R(\delta, \theta),$$

contradicting admissibility of δ. □

Together, Propositions 22 and 23 provide a convenient way of finding an estimator that is both minimax and admissible: a unique π-Bayes rule with constant risk is guaranteed to have this property. However not all estimators with this property arise in this way. Below we show that the MLE in a $N(\theta, 1)$ model is minimax and admissible (recall from Proposition 20 the unbiased MLE cannot be a π-Bayes rule for any prior π[17]). To derive

[17] However in some sense it is a limit of the Bayes rules δ_{τ^2} for priors $N(0, \tau^2)$ when $\tau \to \infty$. A result due to Wald shows that all minimax rules are limits of Bayes rules.

this result, we will show admissibility directly from the Cramér–Rao lower bound: recall that for any estimator $\widetilde{\theta}$, we have (under regularity conditions) writing $b(\theta) := \mathbb{E}_\theta(\widetilde{\theta}) - \theta$,

$$\mathrm{Var}_\theta(\widetilde{\theta}) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta\widetilde{\theta}\right)^2}{I_n(\theta)} = \frac{(b'(\theta)+1)^2}{I_n(\theta)}.$$

Theorem 24. *Consider the model* $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$ *where* $\theta \in \Theta = \mathbb{R}$. *Then* $\widehat{\theta} = \widehat{\theta}(X_1, \dots, X_n) = \bar{X}$ *is admissible and minimax for estimating* θ *in quadratic risk.*

Proof. The MLE has constant risk $R(\widehat{\theta}, \theta) = \mathbb{E}_\theta(\bar{X} - \theta)^2 = 1/n$. Thus it suffices to show $\widehat{\theta}$ is admissible. Now for any estimator $\widetilde{\theta}$, we have the variance–bias decomposition

$$R(\widetilde{\theta}, \theta) = b(\theta)^2 + \mathrm{Var}_\theta(\widetilde{\theta}) \geq b(\theta)^2 + \frac{1}{n}(1 + b'(\theta))^2$$

using the Cramér–Rao lower bound for the final inequality (noting that the Gaussian model satisfies the regularity conditions). Suppose now that $R(\widetilde{\theta}, \theta) \leq R(\widehat{\theta}, \theta)$, so

$$b(\theta)^2 + \frac{1}{n}(1 + b'(\theta))^2 \leq \frac{1}{n}. \tag{3.2}$$

We see that b is bounded from above and below, so by the mean value theorem, there exist sequences $\theta_k^+ \to \infty$ and $\theta_k^- \to -\infty$ with $b'(\theta_k^+) \to 0$ and $b'(\theta_k^-) \to 0$. But from (3.2) then $b(\theta_k^+) \to 0$ and $b(\theta_k^-) \to 0$. Since from (3.2) we know that b is nondecreasing ($b'(\theta) \leq 0$), this implies that $b(\theta) = 0$ for all $\theta \in \mathbb{R}$, so $R(\widetilde{\theta}, \theta) = 1/n = R(\widehat{\theta}, \theta)$ and $\widehat{\theta}$ is admissible. □

3.4 The James–Stein estimator

Consider now the multivariate setting where we seek to estimate the mean vector $\theta \in \mathbb{R}^p$ in a $X \sim N_p(\theta, I)$ model (we consider only one observation for simplicity). A natural loss function here is $L(a, \theta) = \|a - \theta\|^2$ with associated risk for an estimator $\widehat{\theta}$ given by

$$R(\widehat{\theta}, \theta) := \mathbb{E}_\theta\{\|\widehat{\theta}(X) - \theta\|^2\} = \sum_{j=1}^p \mathbb{E}_\theta(\widehat{\theta}_j(X) - \theta_j)^2.$$

If, given only X_j, the ‘best’ (in the sense of being admissible and minimax) way to estimate θ_j is via X_j itself, surely the best way to estimate θ is simply using the MLE $\widehat{\theta}(X) = X$? This intuition is correct for $p = 2$, but, in a result that shocked the statistical world upon its discovery, was shown to be false for $p \geq 3$: the MLE $\widehat{\theta}$ is in fact inadmissible in this case! Surprisingly, one can improve on the MLE uniformly by using all components of X to estimate each individual component of θ.

Definition 13. The James–Stein estimator is given by

$$\widehat{\theta}_{\mathrm{JS}} = \left(1 - \frac{p-2}{\|X\|^2}\right) X.$$

Note that $R(\widehat{\theta}, \theta) = \mathbb{E}_\theta \|X - \theta\|^2 = p$. To compute the risk of the James–Stein estimator, we use the following lemma.

Lemma 25 (Stein's lemma)**.** *Let $Z \sim N(\mu, 1)$ and let $g : \mathbb{R} \to \mathbb{R}$ be a bounded, differentiable function such that $\mathbb{E}|g'(Z)| < \infty$. Then*

$$\mathbb{E}[(Z - \mu)g(Z)] = \mathbb{E}[g'(Z)].$$

Proof.

$$\begin{aligned}\mathbb{E}[(Z - \mu)g(Z)] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(z)(z - \mu) \exp\left(-\frac{1}{2}(z - \mu)^2\right) dz \\ &= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(z) \left(\frac{d}{dz} \exp\left(-\frac{1}{2}(z - \mu)^2\right)\right) dz \\ &= -\frac{1}{\sqrt{2\pi}} \left[g(z) \exp\left(-\frac{1}{2}(z - \mu)^2\right)\right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g'(z) \exp\left(-\frac{1}{2}(z - \mu)^2\right) dz \\ &= \mathbb{E}[g'(Z)],\end{aligned}$$

using the boundedness of g for the final equality. □

Theorem 26. *Let $X \sim N_p(\theta, I)$ for $p \geq 3$. The risk of the James–Stein estimator satisfies $R(\widehat{\theta}_{\text{JS}}, \theta) < p$ for all θ, so in particular the MLE is inadmissible.*

Proof. In the below, we drop the subscript θ in the expectations for simplicity. We have

$$\begin{aligned}R(\widehat{\theta}_{\text{JS}}, \theta) = \mathbb{E}\|\widehat{\theta}_{\text{JS}} - \theta\|^2 &= \mathbb{E}\left\|X - \theta - \frac{p-2}{\|X\|^2}X\right\|^2 \\ &= p + (p-2)^2\mathbb{E}\|X\|^{-2} - 2(p-2)\sum_{j=1}^{p} \mathbb{E}\left(\frac{X_j(X_j - \theta_j)}{\|X\|^2}\right). \qquad (3.3)\end{aligned}$$

Consider the final term. We have, by the tower property,

$$\mathbb{E}\left(\frac{X_j(X_j - \theta_j)}{\|X\|^2}\right) = \mathbb{E}\left[\mathbb{E}\left(\frac{X_j(X_j - \theta_j)}{\|X\|^2} \,\Big|\, X_{-j}\right)\right],$$

where $X_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$. We can write the RHS as $\mathbb{E}[(X_j - \theta_j)g_j(X_j) \mid X_{-j}]$ where

$$g_j(x) = \frac{x}{x^2 + \|X_{-j}\|^2},$$

which is bounded provided $\|X_{-j}\|^2 \neq 0$, which occurs almost surely. Also

$$g_j'(x) = \frac{\|X_{-j}\|^2 - x^2}{(x^2 + \|X_{-j}\|^2)^2}$$

is similarly bounded, so applying Stein's lemma (conditionally) we get

$$\begin{aligned}\mathbb{E}[(X_j-\theta_j)g_j(X_j)\,|\,X_{-j}] &= \mathbb{E}[g_j'(X_j)\,|\,X_{-j}] \\ &= \mathbb{E}\left(\frac{\|X\|^2-2X_j^2}{\|X\|^4}\,\middle|\,X_j\right).\end{aligned}$$

Thus

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

Returning to (3.3) we thus have

$$R(\widehat{\theta}_{\mathrm{JS}},\theta) = p-(p-2)^2\mathbb{E}\|X\|^{-2} < p. \qquad \square$$

Remark 10. One can show (see example sheet) that

$$R(\widehat{\theta}_{\mathrm{JS}},\theta) \le p - \frac{(p-2)^2}{p-2+\|\theta\|^2},$$

so the improvement over the MLE is most substantial when $\|\theta\|^2$ is small and p is large.

3.5 Shrinkage

Given the surprising nature of Theorem 26, it is natural to be sceptical: is this just a quirk of Gaussianity? Not really (see below): what is crucial however is that the loss involve all components of θ. The intuition is that estimating all of θ when p is large, is a hard problem and in such so-called 'high-dimensional problems' an estimator can do well by sacrificing some bias if it results in an appreciable reduction in variance. To see this, consider estimating $\theta := \mathbb{E}X$ given data $X \in \mathbb{R}^p$ where $\mathrm{Cov}(X) = I$ (this slightly generalises the setting from earlier). Let $\widehat{\theta}_c := (1-c)X$: the interpretation is that when $c \in [0,1]$, $1-c$ is a factor by which we are 'shrinking' the natural estimator X towards the origin. (The James–Stein estimator uses the data-driven choice $c = (p-2)\|X\|^{-2}$.) We have

$$\begin{aligned}\mathbb{E}\|\widehat{\theta}_c-\theta\|^2 &= \mathbb{E}\|(1-c)X-(1-c)\theta-c\theta\|^2 \\ &= \underbrace{(1-c)^2p}_{\text{variance}} + \underbrace{c^2\|\theta\|^2}_{\text{squared bias}} \\ &= p + c^2(p+\|\theta\|^2) - 2cp \\ &= (p+\|\theta\|^2)\left(c-\frac{p}{p+\|\theta\|^2}\right)^2 - \frac{p^2}{p+\|\theta\|^2} + p \\ &= (p+\|\theta\|^2)\left(c-\frac{p}{p+\|\theta\|^2}\right)^2 + \frac{p\|\theta\|^2}{p+\|\theta\|^2}.\end{aligned}$$

Thus using the optimal choice $c^* := p/\{p + \|\theta\|^2\}$ gives a risk of

$$\frac{p}{1 + p/\|\theta\|^2},$$

which is a large improvement on the risk p realised by $\widehat{\theta}_0 = X$ when $\|\theta\|^2$ is small. This is of course an unfair comparison as $\widehat{\theta}_{c^*}$ requires knowledge of $\|\theta\|^2$. However, we can try to mimic this oracular choice by estimating $\|\theta\|^2$: this seems an easier task than estimating all of $\theta \in \mathbb{R}^p$. Note that $\mathbb{E}\|X\|^2 = p + \|\theta\|^2$: using $\|X\|^2$ to estimate $\|\theta\|^2$ gives the final estimator

$$\left(1 - \frac{p}{\|X\|^2}\right) X;$$

the only difference with the James–Stein estimator is that $p - 2$ has been replaced by p here. This heuristic argument suggests that the favourable properties of the James–Stein estimator can extend beyond Gaussianity.

Is the James–Stein estimator admissible? It turns out $\widehat{\theta}_{\mathrm{JS}}$ is in turn dominated (in the $N_p(\theta, I)$ model) by the *positive-part James Stein estimator*[18]

$$\widehat{\theta}_{\mathrm{JS+}} := \left(1 - \frac{p-2}{\|X\|^2}\right)_+ X,$$

where $(u)_+ = u\mathbb{1}_{[0,\infty)}(u)$. This remedies an undesirable feature of the regular James–Stein estimator whereby if $\|X\|^2$ were small, $\widehat{\theta}_{\mathrm{JS}}$ could have opposite signs to X.

Where does all this leave maximum likelihood estimation? While the MLE can be beaten in finite samples, as we have seen, there is a fairly strong sense in which it is asymptotically unbeatable (some indication of this is given on the Example Sheet, though a general result is beyond the scope of this course). The main message is that when the parameter to be estimated has high dimension, some modifications to the basic maximum likelihood scheme to reduce variance are helpful; in fact a large part of modern statistics has been and continues to be devoted to developing such strategies.

4 Multivariate analysis

4.1 Classification

One decision problem of great practical importance is the so-called (two-class) *classification problem*. It is a form of regression problem where we have an input (or predictors) $X \in \mathcal{X}$ and a binary outcome (or class label) $Y \in \{0, 1\}$. We can characterise the joint distribution of (X, Y) in two ways:

[18] In fact $\widehat{\theta}_{\mathrm{JS+}}$ is itself inadmissible: this comes as a consequence of a general result that all admissible estimators in this model must be smooth functions of the observations.

1. We first generate X according to its marginal distribution and then draw Y according to the *regression function*

$$\mathbb{P}(Y = 1 \ | X = x) = \mathbb{E}(Y \,|\, X = x) =: \eta(x).$$

2. First draw Y according to prior probabilities $\pi_0 := \mathbb{P}(Y = 0)$, $\pi_1 := \mathbb{P}(Y = 1)$ and next generate X via

$$X \,|\, Y = 0 \sim f_0(\cdot) \quad \text{or} \quad X \,|\, Y = 1 \sim f_1(\cdot).$$

Suppose for now that the joint distribution of (X, Y) is known to us. Our goal is to predict Y given data X. To view this as a decision problem, observe that Y here plays the role of θ previously. A natural loss function to use is

$$L(\delta(X), Y) = \mathbb{1}_{\{\delta(X) \neq Y\}},$$

where decision rule δ is known in this context as a *classifier*. The corresponding π-Bayes risk is

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

To find a π-Bayes decision rule δ_π, in this context known as a *Bayes classifier*, for each $x \in \mathcal{X}$, we can choose $\delta_\pi(x)$ to minimise the posterior risk $\mathbb{P}(\delta(x) \neq Y \,|\, X = x)$.

Proposition 27. *A Bayes classifier is given by*

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

If

$$\mathbb{P}\left(\frac{f_1(X)\pi_1}{f_0(X)\pi_0} = 1\right) = 0,$$

then any Bayes classifier δ satisfies $\mathbb{P}(\delta(X) = \delta_\pi(X)) = 1$.

Proof. See Example Sheet. □

When $X \,|\, \{Y = j\} \sim N_p(\mu_j, \Sigma)$, the Bayes classifier takes a particularly simple form. We have

$$\begin{aligned} \log\left(\frac{f_1(X)\pi_1}{f_0(X)\pi_0}\right) &= \log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}(X - \mu_1)^\top \Sigma^{-1}(X - \mu_1) + \frac{1}{2}(X - \mu_0)^\top \Sigma^{-1}(X - \mu_0) \\ &= \log\left(\frac{\pi_1}{\pi_0}\right) + \frac{1}{2}\left(\mu_0^\top \Sigma^{-1}\mu_0 - \mu_1^\top \Sigma^{-1}\mu_1\right) + X^\top \Sigma^{-1}(\mu_1 - \mu_0) \end{aligned}$$

We thus see the Bayes classifier only depends on X through the linear function $X^\top \Sigma^{-1}(\mu_1 - \mu_0)$. It thus defines a linear decision boundary where $\delta_\pi(x) = 1$ for x on one side of the boundary, and $\delta_\pi(x) = 0$ otherwise. This method for classification is known as *linear discriminant analysis*. Typically in practice, the prior probabilities π_0, π_1, means μ_0, μ_1 and covariance Σ would be unknown. Instead, we would have available *training data* $(X_1, Y_1), \ldots, (X_n, Y_n)$ formed of i.i.d. copies of (X, Y) with which to estimate these unknown quantities as follows:

$$\begin{aligned}
n_j &:= \sum_{i=1}^n \mathbb{1}_{\{Y_i = j\}}, \\
\pi_j &:= n_j / n, \\
\widehat{\mu}_j &:= \frac{1}{n_j} \sum_{i : Y_i = j} X_i \\
\widehat{\Sigma} &:= \frac{1}{n-2} \sum_{j=0,1} \sum_{i : Y_i = j}^n (X_i - \widehat{\mu}_j)(X_i - \widehat{\mu}_j)^\top.
\end{aligned}$$

Note that the $1/(n-2)$ factor rather than $1/n$ in the covariance matrix estimate makes it unbiased (see Example Sheet).

4.2 Correlation and partial correlation

Regression and classification problems involve learning aspects of the relationship between an outcome variable and predictors. In other settings we may have multivariate data but there may be no distinguished outcome variable. Instead we might seek to understand the relationship between pairs of variables.

Consider a random vector $X = (X^{(1)}, \ldots, X^{(p)}) \in \mathbb{R}^p$ with $\mathrm{Cov}(X) = \Sigma$ and $\min_j \Sigma_{jj} > 0$. As a first attempt to formalise the idea of variables being 'related to one another' we might look at which pairs of variables are dependent. Recall that if $X \sim N_p(\mu, \Sigma)$, then

$$X^{(j)} \perp\!\!\!\perp X^{(k)} \iff \mathrm{Cov}(X^{(j)}, X^{(k)}) = \Sigma_{jk} = 0.$$

A convenient measure of the strength of the dependence is then given by the *correlation*

$$\rho_{jk} := \mathrm{Corr}(X^{(j)}, X^{(k)}) := \frac{\mathrm{Cov}(X^{(j)}, X^{(k)})}{\sqrt{\mathrm{Var}(X^{(j)})\mathrm{Var}(X^{(k)})}} = \frac{\Sigma_{jk}}{\sqrt{\Sigma_{jj}\Sigma_{kk}}}.$$

Note that

$$\begin{aligned}
|\mathrm{Cov}(X^{(j)}, X^{(k)})| = |\mathbb{E}(X^{(j)} - \mu_j)(X^{(k)} - \mu_k)| &\leq \sqrt{\mathbb{E}[(X^{(j)} - \mu_j)^2]\mathbb{E}[(X^{(k)} - \mu_j)^2]} \\
&= \sqrt{\mathrm{Var}(X^{(j)})\mathrm{Var}(X^{(k)})},
\end{aligned}$$

by the Cauchy–Schwarz inequality, so $\rho_{jk} \in [-1, 1]$. We get

$$\begin{aligned}
\rho_{jk} \in \{-1, 1\} &\iff X^{(j)} - \mu_j = c(X^{(k)} - \mu_k) \text{ a.s. for some } c \in \mathbb{R} \\
&\iff X^{(j)} = m + cX^{(k)} \text{ a.s. for some } c, m \in \mathbb{R}.
\end{aligned}$$

To estimate ρ_{jk} given i.i.d. copies $X_1, \ldots, X_n$ of X, we can use the sample correlation given by

$$\widehat{\rho}_{jk} := \frac{\widehat{\Sigma}_{jk}}{\sqrt{\widehat{\Sigma}_{jj}\widehat{\Sigma}_{kk}}} \qquad \text{where} \qquad \widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top.$$

One can show that when $X \sim N_p(\mu, \Sigma)$ with $\mu \in \mathbb{R}^p$, and Σ lies in the set of positive definite matrices, $\widehat{\Sigma}$ is the MLE of Σ. Our result on plug-in MLEs (Proposition 17) shows that the above is then the MLE of the correlation ρ_{jk}. Moreover, Question 9 of Example Sheet 1 and an application of Slutsky shows that $\widehat{\rho}_{jk} \xrightarrow{p} \rho_{jk}$ even when the data is non-Gaussian (and additionally $\sqrt{n}(\widehat{\rho}_{jk} - \rho_{jk})$ will have a Gaussian limiting distribution).

One issue with basing our idea of when variables are related to one another on dependence or correlation is that many pairs of variables may exhibit dependence without a very meaningful connection between them. For example, in human populations, height and literacy levels are positively correlated. While this may at first appear interesting or alarming, a little thought reveals that this fact is an expected consequence of babies not knowing how to read! If we were to look only at the literacy levels of those individuals of a given age a, then we would not expect to see such a relationship. The statistical property of *conditional independence* captures this idea:

Definition 14. Given random vectors X, Y and Z, we say X is *conditionally independent* of Y given Z, and write $X \perp\!\!\!\perp Y \,|\, Z$, if

$$f_{XY|Z}(x, y|z) = f_{X|Z}(x|z)f_{Y|Z}(y|z) \qquad \text{whenever } f_Z(z) > 0.$$

(Here, for example $f_{XY|Z}(x, y|z) = f_{XY}(x, y)/f_Z(z)$, $f_Z(z) > 0$ is the conditional density of (X, Y) given Z.) If not we say X and Y are *conditionally dependent* given Z and write $X \not\perp\!\!\!\perp Y \,|\, Z$. Equivalently,

$$X \perp\!\!\!\perp Y \,|\, Z \iff f_{X|YZ}(x|y, z) = m(x, z)$$

for some function m whenever $f_{YZ}(y, z) > 0$, and moreover this m will then be the conditional density $f_{X|Z}$.

The interpretation of $X \perp\!\!\!\perp Y \,|\, Z$ is that 'knowing Z makes X unimportant for learning Y (and vice versa)'.

A key fact about jointly Gaussian random variables is that the conditional distributions are also Gaussian:

Proposition 28. *Suppose*

$$(Y, W) \sim N_{d+k}\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \underbrace{\begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}}_{=:\Sigma}\right)$$

where Σ is positive definite. Then

$$Y \,|\, W = w \sim N_d(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(w - \mu_2),\, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

Proof. Our idea is to write $Y = MW + (Y - MW)$ for a matrix $M \in \mathbb{R}^{d \times k}$ such that $Y - MW \perp\!\!\!\perp W$. Since jointly Gaussian random vectors are independent if and only if they are uncorrelated, this is equivalent to asking for

$$0 = \operatorname{Cov}(W, Y - MW) = \Sigma_{21} - \Sigma_{22} M^\top,$$

which occurs when we take $M^\top = \Sigma_{22}^{-1}\Sigma_{21}$. Because $Y - MW \perp\!\!\!\perp W$, the distribution of $Y - MW$ conditional on $W = w$ is equal to its unconditional distribution. As a linear transformation of the Gaussian random vector (Y, W), $Y - MW$ is Gaussian and hence its distribution is characterised by its mean and variance, which we now compute:

$$\begin{aligned}
\mathbb{E}(Y - MW) &= \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \\
\operatorname{Cov}(Y - MW) &= \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21} - 2\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
&= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.
\end{aligned}$$

On the other hand, MW is simply equal to Mw conditional on $W = w$. Putting things together gives the result. □

In the setting of the result above, suppose that $Y \in \mathbb{R}$ and write $W = (X, Z) \in \mathbb{R} \times \mathbb{R}^p$. Then we may write

$$Y = \underbrace{\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2}_{\text{intercept}} + \underbrace{\Sigma_{12}\Sigma_{22}^{-1}}_{\text{coefficient vector}} \begin{pmatrix} X \\ Z \end{pmatrix} + \varepsilon$$

where $\varepsilon \sim N(0, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$, independently of (X, Z). We thus have a normal linear regression model with response Y on predictors (X, Z). Importantly, if the component of the coefficient vector corresponding to X were zero, we would have that the conditional distribution of $Y \,|\, X, Z$ would not depend on X, i.e. $Y \perp\!\!\!\perp X \,|\, Z$. Suppose we have data $(X_i, Y_i, Z_i)_{i=1}^n$ formed of i.i.d. copies of (X, Y, Z). Let

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^n, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n, \quad \mathbf{Z} = \begin{pmatrix} Z_1^\top \\ \vdots \\ Z_n^\top \end{pmatrix} \in \mathbb{R}^{n \times p}.$$

The above suggests measuring the strength of the conditional (in)dependency by examining the coefficient corresponding to $\mathbf{X}$ after linearly regressing $\mathbf{Y}$ onto $(\mathbf{1}, \mathbf{X}, \mathbf{Z})$ where $\mathbf{1} \in \mathbb{R}^n$ is a vector of ones (the intercept). This measure however does not reflect the symmetry in X and Y of the conditional independence relationship.

The fact that the conditional distribution of $(X, Y) \,|\, Z$ is Gaussian offers an alternative to the regression approach above: we can instead examine the correlation of X and Y in the conditional distribution $(X, Y) \,|\, Z$. We have

$$X \perp\!\!\!\perp Y \,|\, Z \iff 0 = \frac{\operatorname{Cov}(X, Y \mid Z)}{\sqrt{\operatorname{Var}(X \mid Z)\operatorname{Var}(Y \mid Z)}} =: \rho_{XY|Z}.$$

The quantity on the right is the *partial correlation* of X and Y given Z. From Proposition 28, we know that $\mathrm{Cov}(X, Y \mid Z = z)$ is constant in z, and similarly for $\mathrm{Var}(X \mid Z)$ and $\mathrm{Var}(Y \mid Z)$. Thus

$$\rho_{XY|Z} = \frac{\mathbb{E}[\{X - \mathbb{E}(X \mid Z)\}\{Y - \mathbb{E}(Y \mid Z)\}]}{\sqrt{\mathbb{E}[\{X - \mathbb{E}(X \mid Z)\}^2]\mathbb{E}[\{Y - \mathbb{E}(Y \mid Z)\}^2]}}.$$

Let us write $P \in \mathbb{R}^{n \times n}$ for the orthogonal projection onto the column space of $(\mathbf{1}, \mathbf{Z}) \in \mathbb{R}^{n \times (p+1)}$. The sample version of the partial correlation is then given by

$$\widehat{\rho}_{XY|Z} := \frac{\{(I - P)\mathbf{X}\}^\top \{(I - P)\mathbf{Y}\}}{\|(I - P)\mathbf{X}\| \|(I - P)\mathbf{Y}\|} = \frac{\mathbf{X}^\top (I - P)\mathbf{Y}}{\|(I - P)\mathbf{X}\| \|(I - P)\mathbf{Y}\|}.$$

Similarly to the sample correlation, one can show this is the maximum likelihood estimate of $\rho_{XY|Z}$. In fact there is a close connection between the regression approach and partial correlation:

Proposition 29. *In the setting above, let $Q \in \mathbb{R}^{n \times n}$ be the orthogonal projection matrix onto the column space of $\mathbf{W} := (\mathbf{1}, \mathbf{X}, \mathbf{Z}) \in \mathbb{R}^{n \times (p+2)}$, assumed to have full column rank $p + 2$. Then the F-statistic for testing the significance of $\mathbf{X}$ in a normal linear model of $\mathbf{Y}$ on $(\mathbf{1}, \mathbf{X}, \mathbf{Z})$ is*

$$\frac{\|(Q - P)\mathbf{Y}\|^2}{\frac{1}{n-p-2}\|(I - Q)\mathbf{Y}\|^2} = (n - p - 2)\frac{\widehat{\rho}^2_{X,Y|Z}}{1 - \widehat{\rho}^2_{X,Y|Z}}.$$

In particular, under the null $H_0 : X \perp\!\!\!\perp Y \mid Z$,

$$(n - p - 2)\frac{\widehat{\rho}^2_{X,Y|Z}}{1 - \widehat{\rho}^2_{X,Y|Z}} \sim F_{1,n-p-2}.$$

Proof. Recall (from IB Statistics)[19] that $Q - P$ is an orthogonal projection with rank 1, so $Q - P = vv^\top / \|v\|^2$ for some vector $v \in \mathbb{R}^n$ with $v \neq 0$. Also since

$$I - P = (I - Q) + (Q - P)$$

and both Q and P have columns in the column space of $\mathbf{W}$, $(I - Q)(Q - P) = 0$, so

$$\|(I - P)\mathbf{Y}\|^2 = \|(I - Q)\mathbf{Y}\|^2 + \|(Q - P)\mathbf{Y}\|^2.$$

Now v is the only eigenvector of $Q - P$ (up to a multiplicative constant) and $(Q - P)\mathbf{X} = (I - P)\mathbf{X}$, so $(Q - P)(I - P)\mathbf{X} = (I - P)\mathbf{X}$. Thus

$$Q - P = \frac{(I - P)\mathbf{X}\mathbf{X}^\top (I - P)}{\|(I - P)\mathbf{X}\|^2},$$

[19] See page 12 of `https://www.statslab.cam.ac.uk/~rds37/teaching/statistical_modelling/notes.pdf`.

so

$$\|(Q-P)\mathbf{Y}\|^2 = \frac{(\mathbf{X}^\top(I-P)\mathbf{Y})^2}{\|(I-P)\mathbf{X}\|^2} = \|(I-P)\mathbf{Y}\|^2\widehat{\rho}^2_{XY|Z}.$$

Thus

$$\frac{\|(Q-P)\mathbf{Y}\|^2}{\|(I-Q)\mathbf{Y}\|^2} = \frac{\widehat{\rho}^2_{XY|Z}}{1-\widehat{\rho}^2_{XY|Z}},$$

which gives the result. □

Remark 11. The result also holds when $p = 0$ i.e. there is no $\mathbf{Z}$: If $X \perp\!\!\!\perp Y$ then writing $\widehat{\rho}$ for the sample correlation of $\mathbf{X}$ and $\mathbf{Y}$,

$$(n-2)\frac{\widehat{\rho}^2}{1-\widehat{\rho}^2} \sim F_{1,n-2}.$$

Remark 12. While the correlation can make sense as an indicator of dependence even when (X, Y) are not Gaussian in that we always have

$$X \perp\!\!\!\perp Y \implies \mathrm{Corr}(X, Y) = 0,$$

if (X, Y, Z) are not jointly Gaussian, it is possible to have $X \perp\!\!\!\perp Y \mid Z$ and *not* have $\widehat{\rho}_{XY|Z} \xrightarrow{p} 0$.

4.3 Principal component analysis

Let $X_1, \ldots, X_n$ be i.i.d. copies of a random vector $X \in \mathbb{R}^p$. When the dimension p of the data is large, it is often of interest to reduce the dimension in some way while trying to retain as much 'information' as possible. The method of *principal component analysis* (PCA) aims to maximise the variability of the compressed data. Given a target dimension $k \leq p$, it works as follows:

1. Form the sample covariance matrix

$$\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\top.$$

2. Find k unit norm eigenvectors $\widehat{v}_1, \ldots, \widehat{v}_k$ corresponding to the k largest eigenvalues of $\widehat{\Sigma}$ (for simplicity, we assume the eigenvalues are unique, so the eigenvectors are unique up to an arbitrary sign).

3. Writing $\widehat{V} \in \mathbb{R}^{p\times k}$ for the matrix with jth column $\widehat{v}_j$, set $U_i := \widehat{V}^\top X_i$; $U_1, \ldots, U_n$ then forms the compressed data.

To understand the motivation for this, observe first that $\widehat{\Sigma}$ estimates $\mathrm{Cov}(X) =: \Sigma$ (recall that it is for example the MLE in a $X \sim N_p(\mu, \Sigma)$ model, though we do not assume this here). Thus $\widehat{v}_j$ estimate the corresponding population eigenvector v_j given by the jth column of V defined through the eigendecomposition

$$\Sigma = V\Lambda V^\top.$$

Here $V \in \mathbb{R}^{p\times p}$ is an orthogonal matrix and $\Lambda \in \mathbb{R}^{p\times p}$ is a diagonal matrix with diagonal entries given by (assumed unique) eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_p \geq 0$.

Then v_1 can be interpreted as the unit vector w such that $\mathrm{Var}(w^\top X)$ is maximal. Indeed, writing $\alpha := V^\top w$, note that $\|\alpha\| = 1$. Then

$$\mathrm{Var}(w^\top X) = w^\top \Sigma w = \alpha^\top V^\top V \Lambda V^\top V \alpha = \sum_{j=1}^{p} \alpha_j^2 \lambda_j \leq \lambda_1,$$

with equality if and only if $\alpha = (1, 0, \ldots, 0)^\top$, i.e. $W = v_1$. Similarly, one can show that v_j is the unit vector, orthogonal to $\{v_1, \ldots, v_{j-1}\}$, upon which the projection of X has maximal variance.

5 Nonparametric inference and Monte Carlo techniques

5.1 The Jackknife

Consider the following setting: we have i.i.d. data $X_1, \ldots, X_n$ and have constructed an estimator $\widehat{\theta}_n = T_n(X_1, \ldots, X_n)$ of a parameter of interest $\theta \in \mathbb{R}$. We would now like to understand the statistical properties of $\widehat{\theta}_n$. The approach we have seen so far in the course for doing this involves applying various stochastic convergence results. However if $\widehat{\theta}_n$ is very complicated, this may be difficult.

If instead of just a single dataset $X_1, \ldots, X_n$, we had available multiple versions of this dataset, we could apply T_n to each to these and hence estimate e.g. the mean and variance of $\widehat{\theta}_n$ or indeed its entire distribution. Resampling techniques aim to mimic this setup (roughly speaking): they involve forming multiple versions of the data from the original dataset, and applying the estimator to each such copy of the data.

The *jackknife* leaves each observation X_i out of the dataset in turn, to give n perturbed versions of our original data. This approach can be used to estimate the bias of $\mathbb{E}_{\theta_0}\widehat{\theta}_n$, where θ_0 is the true parameter, as follows.

Definition 15. Let $\widehat{\theta}_n^{(-i)} := T_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. The *jackknife bias estimate* is defined as

$$\widetilde{B}_n := (n-1)\left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\theta}_n^{(-i)} - \widehat{\theta}_n\right).$$

The *jackknife bias-corrected estimate* of θ is then

$$\widetilde{\theta}_n := \widehat{\theta}_n - \widetilde{B}_n.$$

Theorem 30. *Suppose the bias* $B_n := \mathbb{E}_{\theta_0}(\widehat{\theta}_n) - \theta_0$ *satisfies*

$$B_n = \frac{a}{n} + \frac{b}{n^2} + O(n^{-3}) \tag{5.1}$$

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

$$\mathbb{E}_{\theta_0}(\widetilde{\theta}_n) = \theta_0 + O(n^{-2}).$$

Proof. Observe that $\widetilde{\theta}_n = n\widehat{\theta}_n - \frac{n-1}{n}\sum_{i=1}^n \widehat{\theta}_n^{(-i)}$. Thus

$$\begin{aligned}\mathbb{E}_{\theta_0}(\widetilde{\theta}_n) &= n(\theta_0 + B_n) - (n-1)(\theta_0 + B_{n-1}) \\ &= \theta_0 + \left(a + \frac{b}{n}\right) - \left(a + \frac{b}{n-1}\right) + O(n^{-2}) \\ &= \theta_0 - \frac{b}{n(n-1)} + O(n^{-2}) = \theta_0 + O(n^{-2}). \qquad \square\end{aligned}$$

Example 11. For a concrete example of where the bias condition (5.1) holds, suppose $\mathbb{E}(X_i) =: \mu \in \mathbb{R}$ and we wish to estimate $\theta = \mu^2$ using $\widehat{\theta}_n = \bar{X}_n^2$. Then

$$\mathbb{E}_\mu(\bar{X}_n^2) - \mu^2 = \mathbb{E}_\mu(\bar{X}_n - \mu + \mu)^2 - \mu^2 = \mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X_i)}{n},$$

so the bias condition is indeed satisfied.

More generally, suppose now $\mu \in \mathbb{R}^p$, $\theta = g(\mu)$ and $\widehat{\theta}_n = g(\bar{X}_n)$ for some smooth $g : \mathbb{R}^p \to \mathbb{R}$. Then from a Taylor expansion, we have

$$\widehat{\theta}_n - \theta \approx \nabla g(\mu)^\top(\bar{X}_n - \mu) + \frac{1}{2}(\bar{X}_n - \mu)^\top \nabla^2 g(\mu)(\bar{X}_n - \mu).$$

The first term has mean 0 and for the second term, we have

$$\begin{aligned}\mathbb{E}[(\bar{X}_n - \mu)^\top \nabla^2 g(\mu)(\bar{X}_n - \mu)] &= \mathbb{E}[\mathrm{tr}\{(\bar{X}_n - \mu)^\top \nabla^2 g(\mu)(\bar{X}_n - \mu)\}] \\ &= \mathbb{E}[\mathrm{tr}\{(\bar{X}_n - \mu)(\bar{X}_n - \mu)^\top \nabla^2 g(\mu)\}] \\ &= \frac{\mathrm{tr}\{\mathrm{Cov}(X_i)\nabla^2 g(\mu)\}}{n},\end{aligned}$$

so the bias condition can be expected to hold (and one can show it is satisfied in even greater generality than this).

5.2 The bootstrap

The *bootstrap* takes the idea of the jackknife of reusing the data to understand the distribution of estimators even further. To introduce it, we make the following definition:

Definition 16. Given i.i.d. data $X_1, \dots, X_n$, the *empirical distribution* $\mathbb{P}_n$ is the (random) discrete distribution that places a mass of $1/n$ at each observation X_i, so for a set A,

$$\mathbb{P}_n(A) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(X_i).$$

A sample $X_1^*, \dots, X_n^* \,|\, X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_n$ is known as a *bootstrap sample.* (Note here the X_j^* are *conditionally* i.i.d. given the data $X_1, \dots, X_n$.) Thus $X_1^*, \dots, X_n^*$ is a random resample of the $X_1, \dots, X_n$ with replacement.

Importantly $\mathbb{P}_n$ is something we observe since it depends entirely on the data, whereas the underlying distribution P of the data would typically be unknown to us. Given a parameter $\theta = \theta(P)$ (for example, this could be the mean or median of the distribution), and corresponding estimator $\widehat{\theta}_n := T_n(X_1, \dots, X_n)$, suppose we wish to form a confidence interval for θ based on $\widehat{\theta}_n$. Consider for simplicity the setting where $\theta \in \mathbb{R}$. The central idea of the bootstrap procedure is to approximate the (unknown) distribution function F_n of

$$\sqrt{n}\{\underbrace{T_n(X_1, \dots, X_n)}_{=:\widehat{\theta}_n} - \theta(P)\} \tag{5.2}$$

by the (random but in principle known) distribution function $\widehat{F}_n$ of

$$\sqrt{n}\{\underbrace{T_n(X_1^*, \dots, X_n^*)}_{=:\widehat{\theta}_n^*} - \widehat{\theta}_n\} \,|\, X_1, \dots, X_n.$$

To gain some intuition for why such an approximation may work, one should think of $\widehat{\theta}_n$ as playing the role of $\theta(\mathbb{P}_n)$, the parameter in the empirical distribution. The quality of this approximation is often best when the quantity in (5.2) is a so-called *pivot*, meaning that its distribution is the same for all values of θ under consideration.

In practice $\widehat{F}_n$ will typically be infeasible to compute since there are n^n possible values $(X_1^*, \dots, X_n^*)$ could take. We can however approximate $\widehat{F}_n$ by first drawing independent bootstrap samples $(X_1^{(b)}, \dots, X_n^{(b)})$ for $b = 1, \dots, B$ with B large, and then forming

$$\widehat{F}_n^{(B)}(t) := \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\left\{\sqrt{n}(T_n(X_1^{(b)}, \dots, X_n^{(b)}) - \widehat{\theta}_n) \le t\right\}}.$$

To see how the principle above may be used to construct a confidence interval for θ, fix $\alpha \in (0, 1)$ and let $l_n := F_n^{-1}(\alpha/2)$ and $u_n := F_n^{-1}(1 - \alpha/2)$, where for simplicity we have implicitly assumed F_n is continuous and strictly increasing. Observe that a $(1 - \alpha)$-level confidence interval for θ is given by

$$C_n := \{\theta : l_n \le \sqrt{n}(\widehat{\theta}_n - \theta) \le u_n\}.$$

To describe how we may approximate this, we need one more definition.

Definition 17. Given a distribution function $F : \mathbb{R} \to [0,1]$, the *quantile function* $F^{-1} : (0,1) \to \mathbb{R}$ is

$$F^{-1}(p) := \inf\{t : F(t) \geq p\}.$$

(Note that if F is strictly increasing, then the quantile function is simply the inverse of F.)

Let $\widehat{l}_n := \widehat{F}_n^{-1}(\alpha/2)$ and $\widehat{u}_n := \widehat{F}_n^{-1}(1-\alpha/2)$. A bootstrap confidence interval is then given by

$$\widehat{C}_n := \{\theta : \widehat{l}_n \leq \sqrt{n}(\widehat{\theta}_n - \theta) \leq \widehat{u}_n\}.$$

The validity of this approach rests on $\widehat{l}_n$ and $\widehat{u}_n$ approaching l_n and u_n. This can be expected when $F_n \xrightarrow{d} F$ and $\widehat{F}_n$ approaches this limiting distribution F (see Example Sheet for details). To show this latter fact in full generality is beyond the scope of this course: we shall study the special case where

$$\begin{aligned} &\theta(P) \text{ is the mean of the distribution } P \text{ and} \\ &\widehat{\theta}_n = \bar{X}_n \text{ is the sample mean.} \end{aligned} \tag{5.3}$$

We make use of the following 'nonasymptotic central limit theorem' whose statement is *non-examinable*.

Theorem 31 (Berry–Esseen theorem)**.** *Suppose $Z_1, \ldots, Z_n$ are i.i.d. with mean $\mu \in \mathbb{R}$ and variance σ^2. Then for any $\delta \in (0,1]$,*

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\left(\sqrt{n}(\bar{Z}_n - \mu) \leq t\right) - \Phi(t/\sigma)\right| \leq \frac{8\mathbb{E}(|Z_1-\mu|^{2+\delta})}{\sigma^{2+\delta}n^{\delta/2}}.$$

Theorem 32. *Suppose $X_1, X_2, \ldots$ are i.i.d. with mean θ. In the setting of* (5.3) *suppose that for some $\delta > 0$, $\mathbb{E}|X_1 - \theta|^{2+\delta} < \infty$ and let $\sigma^2 := \mathrm{Var}(X_1) > 0$. Then*

$$\sup_{t\in\mathbb{R}} |\widehat{F}_n(t) - \Phi(t/\sigma)| \xrightarrow{a.s.} 0.$$

Proof. Write $\widehat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ for the sample variance. Recall from Example 5 that $\widehat{\sigma}_n^2 \xrightarrow{a.s.} \sigma^2$ (the claim was stated with convergence in probability, but the argument therein also yields the stronger almost sure convergence). We apply the Berry–Esseen theorem[20] to $X_1^*, \ldots, X_n^*$ (conditional on $X_1, \ldots, X_n$). We have $\widehat{\theta}_n^* = \frac{1}{n}\sum_{i=1}^n X_i^* =: \bar{X}_n^*$ and

$$\mathbb{E}(X_i^* \,|\, X_1, \ldots, X_n) = \bar{X}_n, \qquad \mathrm{Var}(X_i^* \,|\, X_1, \ldots, X_n) = \widehat{\sigma}_n^2.$$

Thus

$$A_n := \sup_{t\in\mathbb{R}} \Big| \underbrace{\mathbb{P}\left(\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq t \,|\, X_1, \ldots, X_n\right)}_{=\widehat{F}_n(t)} - \Phi(t/\widehat{\sigma}_n)\Big| \leq 8\frac{\frac{1}{n}\sum_{i=1}^n |X_i - \bar{X}_n|^{2+\delta}}{\widehat{\sigma}_n^{2+\delta}n^{\delta/2}}.$$

[20] Why wouldn't a regular CLT work? The issue is that here we do not have i.i.d. data from a single fixed distribution, but rather the distribution $\mathbb{P}_n$ is changing with n. There is a version of the CLT for triangular arrays that can be used here, though the proof is slightly more involved. The upshot is that it avoids the assumption $\mathbb{E}|X_1 - \theta|^{2+\delta}$ that we have had to make here.

Now $|X_i - \theta - (\bar{X}_n - \theta)| \leq 2\max(|X_i - \theta|, |\bar{X}_n - \theta|)$ so

$$\frac{1}{2^{2+\delta} n} \sum_{i=1}^{n} |X_i - \bar{X}_n|^{2+\delta} \leq \frac{1}{n} \sum_{i=1}^{n} |X_i - \theta|^{2+\delta} + |\bar{X}_n - \theta|^{2+\delta} \overset{a.s.}{\to} \mathbb{E}|X_1 - \theta|^{2+\delta}$$

by SLLN and CMT. Thus $A_n \overset{a.s.}{\to} 0$. By the triangle inequality,

$$\sup_{t\in\mathbb{R}} |\widehat{F}_n(t) - \Phi(t/\sigma)| \leq A_n + \underbrace{\sup_{t\in\mathbb{R}} |\Phi(t/\sigma) - \Phi(t/\widehat{\sigma}_n)|}_{=:B_n}$$

so it suffices to show $B_n \overset{a.s.}{\to} 0$. Note first that for a sequence $a_n \to a > 0$, $\sup_t |\Phi(a_n t) - \Phi(at)| \to 0$. Indeed, we have that for all n sufficiently large, $a_n > a/2$, so by the MVT, for such n,

$$|\Phi(a_n t) - \Phi(at)| \leq |a_n - a||t|\phi(at/2).$$

but $\sup_t |t|\phi(at/2) < \infty$, so the above tends to 0. Now by the CMT, $\widehat{\sigma}_n^{-1} \overset{a.s.}{\to} \sigma$ so from the above, $B_n \overset{a.s.}{\to} 0$. □

Remark 13. The version of the bootstrap discussed above is sometimes known as the *nonparametric bootstrap* to distinguish it from the *parametric bootstrap*. The latter works in a setting where we have a parametric model $\{P_\theta : \theta \in \Theta\}$. Instead of estimating the distribution P_{θ_0} by $\mathbb{P}_n$, we can form an estimate $\widehat{\theta}_n$ of θ_0, and use $P_{\widehat{\theta}_n}$ in place of $\mathbb{P}_n$, so we draw $X_1^*, \ldots, X_n^* \,|\, X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_{\widehat{\theta}_n}$.

5.3 Monte Carlo methods

We have seen how while the bootstrap resampling distribution function $\widehat{F}_n$ is in principle known to us, we nevertheless will typically need to approximate it through simulation. Another large class of methods that require us to compute quantities relating to potentially complex distributions are Bayesian methods: for example there may be no closed-form formulas for the posterior mean or quantiles of the posterior distribution. In such situations, numerical simulation techniques can be very useful, and we now study the general problem of simulating from a known distribution.

As a starting point, we shall assume we can generate[21] $U_1, U_2, \ldots \overset{\text{i.i.d.}}{\sim} U[0,1]$. If F is a distribution function on $\mathbb{R}$, we can always then generate i.i.d. draws from F via

$$F^{-1}(U_1), F^{-1}(U_2), \ldots \overset{\text{i.i.d.}}{\sim} F$$

where F^{-1} is the quantile function of F. Indeed, when F is strictly increasing and continuous, $\mathbb{P}(F^{-1}(U_1) \leq t) = \mathbb{P}(U_1 \leq F(t)) = F(t)$; for the general case see the example

[21] In practice we would only be able to generate a *pseudo-random* uniform sample, but algorithms are sufficiently advanced that to a large extent we can work as if we in fact have a uniform sample.

sheet. With this, we can approximate, for example $\mathbb{E}g(X)$ where $X \sim F$ by appealing to the SLLN:

$$\frac{1}{N}\sum_{i=1}^{N} g(X_i) \overset{a.s.}{\to} \mathbb{E}g(X),$$

where we have written $X_i := F^{-1}(U_i)$. Sometimes however it is not possible to compute F^{-1} explicitly, in which case we need to resort to other methods.

5.3.1 Importance sampling

Suppose f is a (potentially multivariate) density from which it is hard to simulate. Suppose however there is a density h whose support includes that of f, from which we can simulate easily. Observe that

$$\mathbb{E}_{X\sim h}\left(g(X)\frac{f(X)}{h(X)}\right) = \int_{\mathcal{X}} g(x)\frac{f(x)}{h(x)}h(x)\,dx = \mathbb{E}_{X\sim f}(g(X)).$$

As a consequence, for $X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} h$, we have

$$\frac{1}{N}\sum_{i=1}^{N} g(X_i)\frac{f(X_i)}{h(X_i)} \overset{a.s.}{\to} \mathbb{E}_{X\sim f}(g(X)).$$

5.3.2 Accept–reject algorithm

An alternative when $f(x) \le Mh(x)$ for all $x \in \mathcal{X}$ is the following:

1. Generate $X \sim h$ and independently $U \sim U[0,1]$.

2. If $U \le \frac{f(X)}{Mh(X)}$, output $Y = X$; otherwise return to step 1.

Then $Y \sim f$ (see example sheet). Note that here the computation required to generate a single draw is random (and will tend to be lower if M is lower).

5.3.3 Markov chains and invariant measures

One very important class of procedures for generating samples from a density f involves constructing a Markov chain which has f as its so-called invariant distribution. Recall that a (discrete-time) Markov chain $X_0, X_1, X_2, \ldots$ is a sequence of random variables where for any $m \ge 1$ and any (measurable) $B \subseteq \mathcal{X}$,

$$\mathbb{P}(X_m \in B \mid X_{m-1} = t, X_{m-2} = t_{m-2}, \ldots, X_0 = t_0) = \mathbb{P}(X_m \in B \mid X_{m-1} = t) =: K(t, B)$$

where K is the *Markov transition kernel* for the chain. The corresponding transition pdf k (if it exists) satisfies

$$K(t, B) = \int_B k(t, s)\,ds$$

for all (measurable) $B \subseteq \mathcal{X}$. (We will later encounter a case where K is a mixture of a discrete and a continuous distribution.) Note that if $X_{m-1} \sim f$, then the distribution of X_m is given by

$$\int_{\mathcal{X}} K(t, \cdot) f(t)\, dt.$$

Definition 18. A pdf f on $\mathcal{X}$ is *invariant* or *stationary* for K if

$$\int_{\mathcal{X}} K(t, B) f(t)\, dt = \int_B f(t)\, dt$$

for all (measurable) $B \subseteq \mathcal{X}$.

Results in ergodic theory (see Probability and Measure) imply that, under certain conditions on the Markov chain, the distribution of X_N converges to its unique invariant distribution. (We will not concern ourselves with the detailed conditions in this course.) Moreover, we also have

$$\frac{1}{N} \sum_{i=1}^{N} g(X_i) \overset{a.s.}{\to} \mathbb{E}_{X \sim f}(g(X)).$$

We now look at some important examples of this key idea.

5.3.4 Gibbs sampler

The Gibbs sampler is a useful method for generating samples from a multivariate distribution. We illustrate the idea in the bivariate case. Suppose $(X, Y) \sim f$ and we can simulate from each of the conditionals $f_{Y|X}(\cdot|x)$ and $f_{X|Y}(\cdot|y)$. As an illustration of when such a situation could arise, consider the following example.

Example 12. Recall (Example Sheet 2, Question 11) that when $X_1, \ldots, X_n \,|\, \mu, \sigma^2 \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ with improper prior density $\pi(\mu, \sigma) \propto \sigma^{-2}$, the posterior

$$\Pi(\sigma, \mu \,|\, X_1, \ldots, X_n) \propto \sigma^{-(n+2)} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \right\}.$$

Thus writing $\omega := \sigma^{-2}$ for the precision,

$$\begin{aligned} \mu \,|\, \omega, X_1, \ldots, X_n &\sim N(\bar{X}, 1/(\omega n)) \\ \omega \,|\, \mu, X_1, \ldots, X_n &\sim \text{Gamma}\left(\frac{n+1}{2}, \frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^2 \right). \end{aligned}$$

The Gibbs sampler takes the following steps. We initialise $X_0 = x$ and then for $m = 1, 2, \ldots$ iteratively perform:

1. $Y_m \sim f_{Y|X}(\cdot \,|\, X = X_{m-1})$,

2. $X_m \sim f_{X|Y}(\cdot \,|\, Y = Y_m)$.

This generates a Markov chain $(Y_1, X_1), (Y_2, X_2), \ldots$ where the joint density f is invariant. To see this, we may argue as follows. The transition density here is

$$k((y_1, x_1), (y_2, x_2)) = f_{Y|X}(y_2|x_1) f_{X|Y}(x_2|y_2),$$

so

$$\begin{aligned}\int\!\!\int f_{Y|X}(y_2|x_1) f_{X|Y}(x_2|y_2) f_{XY}(x_1, y_1)\, dy_1\, dx_1 &= f_{X|Y}(x_2|y_2) \int f_{Y|X}(y_2|x_1) f_X(x_1)\, dx_1 \\ &= f_{X|Y}(x_2|y_2) f_Y(y_2) \\ &= f_{XY}(x_2, y_2).\end{aligned}$$

The method generalises to larger numbers of variables by cycling through each variable in turn.

5.3.5 Metropolis–Hastings

The Gibbs sampler, while simple, has the issue that the full conditionals may often be tricky to sample from. The Metropolis–Hastings algorithm is a powerful method that only requires an auxiliary *proposal* conditional density $q(\cdot|t)$ from which we can simulate. Given an initial $X_0 = x$ it proceeds as follows for $m = 1, 2, \ldots$:

1. Draw $S_m \,|\, X_m \sim q(\cdot|X_m)$.
2. Let $A_m \,|\, X_m, S_m \sim \mathrm{Bern}(a(X_m, S_m))$ where

$$a(t, s) := \min\left(\frac{f(s)}{f(t)} \frac{q(t|s)}{q(s|t)}, 1\right).$$

Set $X_{m+1} := A_m S_m + (1 - A_m) X_m$.

Importantly, the Metropolis–Hastings algorithm only requires evaluation of the ratio $f(s)/f(t)$, rather than $f(t)$ itself. This is particularly useful when f is a posterior density since then the normalisation factor does not need to be computed.

Theorem 33. *In the setting above, suppose $q(s|t) > 0$ for all $s, t \in \mathcal{X}$, where $\mathcal{X}$ is the support of f. Then f is invariant for the transition kernel K of the Markov chain $X_1, X_2, \ldots$ generated by the Metropolis–Hastings algorithm.*

Proof. We have

$$\begin{aligned}K(t, B) &= \mathbb{P}(X_{m+1} \in B \,|\, X_m = t) \\ &= \mathbb{P}(X_{m+1} \in B, A_m = 1 \,|\, X_m = t) + \mathbb{P}(X_{m+1} \in B, A_m = 0 \,|\, X_m = t) \\ &= \int_B \underbrace{\mathbb{P}(A_m = 1 \,|\, S_m = s, X_m = t)}_{a(t,s)} q(s|t)\, ds + \mathbb{1}_B(t) \mathbb{P}(A_m = 0 \,|\, X_m = t).\end{aligned}$$

Now

$$\begin{aligned}\mathbb{P}(A_m = 0 \,|\, X_m = t) &= \int_{\mathcal{X}} \mathbb{P}(A_m = 0 \,|\, S_m = s, X_m = t) q(s|t)\, ds \\ &= 1 - \int_{\mathcal{X}} a(t,s) q(s|t)\, ds.\end{aligned}$$

Also (interchanging the order of integration, which is always justified when the integrand is non-negative),

$$\begin{aligned}\int_{\mathcal{X}} \int_B a(t,s) q(s|t)\, ds f(t)\, dt &= \int_B \int_{\mathcal{X}} \min(f(s)q(t|s), f(t)q(s|t))\, dt\, ds \\ &= \int_B f(s) \int_{\mathcal{X}} a(s,t) q(t|s)\, dt\, ds \\ &= \int_B f(s) \mathbb{P}(A_m = 1 \,|\, X_m = s)\, ds.\end{aligned}$$

Thus

$$\begin{aligned}\int_{\mathcal{X}} K(t,B) f(t)\, dt &= \int_B f(s)\mathbb{P}(A_m = 1 \,|\, X_m = s)\, ds + \int_B f(t) \mathbb{P}(A_m = 0 \,|\, X_m = t)\, dt \\ &= \int_B f(t)\, dt\end{aligned}$$

as required. □

Example 13 ((Special case of) preconditioned Crank–Nicolson (pCN))**.** Consider a parametric model $\{f(\cdot, \theta) : \theta \in \mathbb{R}^p\}$ where we take a Bayesian approach with prior $\theta \sim N_p(0, I)$. We wish to sample from the posterior

$$\Pi(\theta \,|\, X) \propto f(X, \theta) \exp(-\|\theta\|^2/2).$$

If we take $q(\cdot|t) \sim N_p(t\sqrt{1-2\delta}, 2\delta I)$ for a tuning parameter $\delta \in (0, 1/2)$, then we see $a(t,s)$ has the particularly simple form (see example sheet)

$$a(t,s) = \min\left(\frac{f(X,s)}{f(X,t)}, 1\right).$$

5.4 Introduction to nonparametric statistics

We have spent much of the course studying *parametric* statistical models $\{P_\theta : \theta \in \Theta\}$, where when such a model is well-specified, the goal of understanding the distribution of the data P_{θ_0} can be translated to estimating $\theta_0 \in \Theta$. However, such a model can be hard to defend when the sample size is large, and in such settings, it is often of interest to estimate the distribution of the data without recourse to a potentially restrictive statistical model. Tasks of this nature fall within the realm of *nonparametric statistics.* Here we only provide the briefest introduction to this rich and exciting area by focusing on the problem of estimating the distribution function F based on i.i.d. data $X_1, \ldots, X_n \in \mathbb{R}$.

Definition 19. The *empirical distribution function* $\widehat{F}_n$ based on a sample $X_1, \ldots, X_n$ is given by

$$\widehat{F}_n(t) := \mathbb{P}_n((-\infty, t]) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty,t]}(X_i).$$

The SLLN guarantees that for each fixed $t \in \mathbb{R}$, $\widehat{F}_n(t) \overset{a.s.}{\to} F(t)$. However, similarly to the way our ULLN strengthened the SLLN, we have the following uniform convergence result:

Theorem 34 (Glivenko–Cantelli)**.**

$$\sup_{t\in\mathbb{R}} |\widehat{F}_n(t) - F(t)| \overset{a.s.}{\to} 0.$$

Proof. We only consider the case where F is continuous for simplicity. Fix $m \in \mathbb{N}$ and pick $-\infty = t_0 < t_1 < \cdots < t_{m-1} < t_m = \infty$ such that $F(t_j) - F(t_{j-1}) = 1/m$ for $j = 1, \ldots, m$. Now for all $t \in \mathbb{R}$, there exists $j \in \{1, \ldots, m\}$ such that $t \in [t_{j-1}, t_j]$, so

$$\begin{aligned}\widehat{F}_n(t) - F(t) &\leq \widehat{F}_n(t_j) - F(t_{j-1}) = \widehat{F}_n(t_j) - F(t_j) + \frac{1}{m},\\ \widehat{F}_n(t) - F(t) &\geq \widehat{F}_n(t_{j-1}) - F(t_j) = \widehat{F}_n(t_{j-1}) - F(t_{j-1}) - \frac{1}{m}.\end{aligned}$$

Thus

$$\sup_t |\widehat{F}_n(t) - F(t)| \leq \underbrace{\max_{j=0,\ldots,m} |\widehat{F}_n(t_j) - F(t_j)|}_{\overset{a.s.}{\to} 0 \text{ by SLLN}} + \frac{1}{m}.$$

We therefore have that writing $\Omega_m := \{\limsup_{n\to\infty} \sup_t |\widehat{F}_n(t) - F(t)| \leq 1/m\}$, we have $\mathbb{P}(\Omega_m) = 1$. But $\Omega_\infty := \cap_{m=1}^\infty \Omega_m = \{\lim_{n\to\infty} \sup_t |\widehat{F}_n(t) - F(t)| = 0\}$ and

$$\mathbb{P}(\Omega_\infty^c) = \mathbb{P}(\cup_{m=1}^\infty \Omega_m^c) \leq \sum_{m=1}^\infty \mathbb{P}(\Omega_m^c) = 0. \qquad \square$$

While the above result is encouraging, it does not directly allow us to conduct inference about the unknown F. Observe however that when F is continuous and strictly increasing,

$$\sup_{t\in\mathbb{R}} |\widehat{F}_n(t) - F(t)| = \sup_{t\in\mathbb{R}} |\widehat{F}_n(F^{-1}(t)) - \underbrace{F(F^{-1}(t))}_{=t}|.$$

Now $\{X_i \in (-\infty, F^{-1}(t)]\} = \{F(X_i) \leq t\} = \{F(X_i) \in (-\infty, t]\}$. Moreover $F(X_i) \sim U[0,1]$ since $\mathbb{P}(F(X_i) \leq t) = \mathbb{P}(X_i \leq F^{-1}(t)) = F(F^{-1}(t)) = t$. Thus

$$\widehat{F}_n(F^{-1}(t)) \overset{d}{=} \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty,t]}(U_i)$$

where $U_1, \ldots, U_n \overset{\text{i.i.d.}}{\sim} U[0,1]$. In particular, we see that the distribution of

$$\sup_{t \in \mathbb{R}} |\widehat{F}_n(t) - F(t)|$$

is precisely the same for all continuous, strictly increasing F, and can be determined through simulation! In fact:

Theorem 35 (Kolmogorov–Smirnov)**.** *If F is a continuous distribution function, then*

$$\sqrt{n} \sup_t |\widehat{F}_n(t) - F(t)| \overset{d}{\to} K,$$

where K has a so-called Kolmogorov *distribution for all F. Moreover the distribution function of K is continuous.*

This result can be used to construct asymptotically valid confidence bands for F (see example sheet), or test the null hypothesis $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F_0$ for some known continuous distribution function F_0.