SB2.1

Foundations of Statistical Inference

Michaelmas Term 2022

George Deligiannidis

Contents

This set of notes is largely based on notes prepared by Damian Falck in MT2020 based on slides and lectures given by Julien Berestycki. The material goes back to material prepared by Judith Rousseau.

This is still work in progress and may contain typos or even errors. I would very much appreciate your help in improving them so if you spot anything please let me know at `deligian@stats.ox.ac.uk`.

Chapter 0

Notations

The situations of interest to us in this course start in general with having observed some data x, where x is a point in $\mathcal{X}$.

Example. Consider a large field of soybean plants. During 7 weeks, each Monday 5 plants are randomly chosen and the average height recorded.

The data are $x = \{5, 13, 16, 13, 23, 33, 40\}$. Here $\mathcal{X} = \mathbb{R}_+^7$.

We will consider x as the ***realisation*** of a random variable X, taking values in some measurable space $(\mathcal{X}, \mathcal{F})$, where the distribution of X is (at least partly) unknown. **Statistical inference is about using x to gain information on the distribution of X.**

We write $\mathcal{P}(\mathcal{X})$ for the collection of all probability measures on $(\mathcal{X}, \mathcal{F})$. We will usually be interested in a smaller ***class*** or ***family*** of possible distributions $\mathcal{P} \subset \mathcal{P}(\mathcal{X})$ parametrised by some parameter θ:

Definition 0.1. A set $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where the P_θ are probability distributions on $\mathcal{X}$, is called a ***statistical model***. Here Θ is the ***parameter space***.

If P_θ is *absolutely continuous* w.r.t. to some reference measure μ (for our purposes the Lebesgue measure), we write $f(x, \theta)$, or $p_\theta(x)$ for its probability *density* function, whereas if P_θ is *discrete* we write $f(x, \theta)$ for its probability *mass* function. We will often identify a distribution $\mathbb{P}_\theta$ with its density. We write $\mathbb{E}_\theta[\cdot]$ and $\mathbb{P}_\theta[\cdot]$ to mean expectations/probabilities under P_θ; so in $\mathbb{E}_\theta[\phi(X)]$, for example, we take X to have distribution P_θ.

Other possible notations for the same mass/density include $p_\theta(x)$, $p(x, \theta)$, $p(x \mid \theta)$, $f(x \mid \theta)$, $\mathbb{P}_\theta(X = x)$ (in the discrete case), and $L(\theta; x)$.

Remark (**Remark on use of notation**). Throughout this course we will freely drift between different notations for the same objects. This is somewhat intentional.

Chapter 1

Exponential Families

1.1 Definition and examples

There is one particular class of statistical models that will come up time and time again in our journey, and to which many of the common distributions belong:

Definition 1.1. A family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probabilities (pmf or pdf) on some set $\mathcal{X}$, indexed by θ is called an ***exponential family*** if there exists $k \in \mathbb{N}$, functions $\eta_1, \ldots, \eta_k, B : \Theta \mapsto \mathbb{R}$, statistics $T_1, \ldots, T_k : \mathcal{X} \mapsto \mathbb{R}$ and a non-negative real-valued function h on $\mathcal{X}$ such that the pdf/pmfs $p(x;\theta)$ of P_θ have the form

$$p(x;\theta) = \exp\left[\sum_{i=1}^{k} \eta_i(\theta) T_i(x) - B(\theta)\right] h(x). \tag{1.1}$$

Remark. By changing the set $\mathcal{X}$ if necessary, we can always focus on the case where $0 < h(x) < \infty$, for all $x \in \mathcal{X}$ without changing the exponential family.

Remark. The above should be implicitly understood as saying that

$$\mathrm{d}P_\theta(\cdot) = p(x;\cdot)\mathrm{d}\mu,$$

for some reference measure μ. When $x \in \mathbb{R}^d$, μ will typically be Lebesgue measure, and for discrete sets it will be the counting measure.

The η_i are called the ***natural*** or ***canonical*** parameters, and the $T_i(x)$ are called the ***natural*** or ***canonical*** observations.

Since for all $\theta \in \Theta$

$$1 = \int_x p(x;\theta)\,\mathrm{d}x = \exp(-B(\theta)) \left[\int_x h(x) \exp\left(\sum_{i=1}^{k} \eta_i(\theta) T_i(x)\right) \mathrm{d}x\right],$$

we can think of $\exp(-B(\theta))$ as a ***normalisation***. Observe that B only depends on θ through $\eta(\theta)$. (In the above expression the integral should be a sum if $p(x;\theta)$ is a pmf, or we can think of $p(x;\theta)$ as *distribution* to encompass both cases).

Often, it is useful to use the η_i as the parameters and to write the model in its ***canonical form***,

$$p(x;\eta) = \exp\left[\sum_{i=1}^{n} \eta_i T_i(x) - B(\eta)\right] h(x).$$

(Note this is possible even if $\theta \mapsto \eta$ is not one-to-one. Note also that there is a slight abuse of notation as $(x, \theta) \mapsto p(x; \theta)$ and $(x, \eta) \mapsto p(x; \eta)$ are not the same function).

In general θ and x can be multidimensional.

Examples (Common 1-parameter exponential families).

- **Poisson distribution.** For the Poi(θ) distribution, the mass function $f(x;\theta) = \frac{e^{-\theta}\theta^x}{x!}$ ($x = 0, 1, 2, \ldots$) can be written as

$$f(x;\theta) = \frac{1}{x!}\mathrm{e}^{-\theta + x\log\theta} \tag{1.2}$$

$$= h(x)\exp(\eta(\theta)x - B(\theta)) \tag{1.3}$$

  with $h(x) = 1/x!$, $\eta(\theta) = \log\theta$, $B(\theta) = \theta$ and $T(x) = x$. The natural parameter is $\log\theta$.

- **Binomial distribution with known number of trials.** For the Bin(n, p) distribution, considering n to be known and p to be the parameter, the mass function may be written as

$$f(x;p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{1.4}$$

$$= \binom{n}{x} \exp\left[x(\log p - \log(1-p)) + n\log(1-p)\right] \tag{1.5}$$

  (for $x = 0, 1, \ldots, n$). So $h(x) = \binom{n}{x}$, $T(x) = x$, $\eta(p) = \log\frac{p}{1-p}$, and $B(p) = -n\log(1-p)$.

- **Gaussian distribution with known variance.** For the $\mathcal{N}(\mu, 1)$ distribution (for example), the density may be written as

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2}\right] = \frac{\exp\left(-\frac{x^2}{2}\right)}{\sqrt{2\pi}}\exp\left[\mu x - \frac{\mu^2}{2}\right],$$

  so $h(x) = \frac{\exp\left(-\frac{x^2}{2}\right)}{\sqrt{2\pi}}$, $\eta(\mu) = \mu$, $T(x) = x$ and $B(\mu) = \frac{\mu^2}{2}$.

Examples (Common 2-parameter exponential families).

- **Gamma distribution.** For the Gamma(α, β) distribution, with $\theta = (\alpha, \beta)$, we have mass function

$$f(x;\theta) = \frac{\beta^\alpha x^{\alpha-1}\mathrm{e}^{-\beta x}}{\Gamma(\alpha)}\mathbb{1}_{x\geqslant 0} \tag{1.6}$$

$$= \exp\Bigg[\underbrace{(\alpha-1)}_{\eta_1(\theta)}\underbrace{\log x}_{T_1(x)} - \underbrace{\beta}_{\eta_2(\theta)}\underbrace{x}_{T_2(x)} - \underbrace{(\log(\Gamma(\alpha)) - \alpha\log\beta)}_{B(\theta)}\Bigg]\underbrace{\mathbb{1}_{x\geqslant 0}}_{h(x)}. \tag{1.7}$$

- **Gaussian distribution.** For the $\mathcal{N}(\mu, \sigma^2)$ distribution, with $\theta = (\mu, \sigma^2)$, we have mass function

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1.8}$$

$$= \exp\Bigg[\underbrace{-\frac{1}{2\sigma^2}}_{\eta_1(\theta)}\underbrace{x^2}_{T_1(x)} + \underbrace{\frac{\mu}{\sigma^2}}_{\eta_2(\theta)}\underbrace{x}_{T_2(x)} - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)}_{B(\theta)}\Bigg]. \tag{1.9}$$

1.1.1 Support and counterexamples

One important property of exponential families, is that all distributions in such a family are equivalent as stated by the following proposition.

Proposition 1.2. *Two probability measures $\mathbb{P}$ and $\mathbb{Q}$ are said to be equivalent if we have $\mathbb{P}(N) = 0$ iff $\mathbb{Q}(N) = 0$. If $\mathcal{P} = \{p(x;\theta) : \theta \in \Theta\}$ is an exponential family, then all $p(\cdot;\theta)$ are equivalent.*

Proof. Take $\theta_1 \neq \theta_2 \in \Theta$ and suppose $\mathbb{P}_{\theta_1}(N) = 0$. Write $\mathbf{1}_N$ for the indicator function of N.

$$\mathbb{P}_{\theta_1}(N) = e^{-B(\theta_1)} \int \exp\left(\sum_j \eta_j(\theta_1) T_j(x)\right) h(x)\mathbf{1}_N(x)dx = 0 \tag{1.10}$$

This implies that $h(x)\mathbf{1}_N(x) = 0$ for Lebesgue almost all x and therefore that

$$\mathbb{P}_{\theta}(N) = e^{-B(\theta)} \int \exp\left(\sum_j \eta_j(\theta) T_j(x)\right) h(x)\mathbf{1}_N(x)dx = 0 \tag{1.11}$$

for arbitrary $\theta \in \Theta$. □

Definition 1.3. For a distribution P with density f, the ***support*** of $\mathcal{P}$, equivalently of f, is defined as the set

$$\mathsf{supp}(\mathcal{P}) = \{x : f(x) > 0\}.$$

Corollary 1.4. *In an exponential family $\mathcal{P} = \{f(x;\theta), \theta \in \Theta\}$ the support of $f(x;\theta)$ does* **not** *depend on θ. We will write $\mathcal{A}$ for the common support of the $f(x;\theta)$.*

In fact, in the general case where h is allowed to vanish on a subset of $\mathcal{X}$, $\mathcal{A} = \{x : h(x) > 0\}$. It can be easily seen that $P_\theta(\mathcal{A}) = 1$ for all θ.

Example. $f(x;\theta) = \mathrm{e}^{\theta - x}\mathbb{1}_{x>\theta}$ is *not* an exponential family.

Example. Another example of a family which is ***not*** exponential is the Cauchy family with location parameter μ:

$$f(x;\mu) = \frac{1}{\pi(1 + (x-\mu)^2)}.$$

1.2 Parsimonious parametrisation

Exponential families $\mathcal{P}$ generally have multiple representations. Although $\eta = (\eta_1, \eta_2, \ldots, \eta_k)$, $T = (T_1, \ldots, T_k)$ and k are not uniquely determined we call (1.1) a ***k-dimensional family***. We will see that optimal statistical procedures will only depend on the k-dimensional statistic T and it therefore makes sense to choose k to be as small as possible. An exponential family with minimal number of summands is called a ***strictly k-parameter exponential family***.

Definition 1.5. A class of probability measures $\mathcal{P} = \{p(x;\theta) : \theta \in \Theta\}$ which is an exponential family is said to be ***strictly k-parameter*** when k is minimal, that is if $\mathcal{P} = \{g(x;\theta) : \theta \in \Theta\}$ with

$$g(x;\theta) = \tilde{h}(x) \exp\left\{\sum_{i=1}^{l} \tilde{\eta}_i(\theta)\tilde{T}_i(x) - \tilde{B}(\theta)\right\},$$

then $l \geq k$.

Definition 1.6. A representation of an exponential family $\mathcal{P}$ of the form (1.1) is called ***minimal*** if k is minimal; that is for any other representation of $\mathcal{P}$ in terms of $\{\tilde{\eta}_j : j \leq l\}, \{\tilde{T}_j : j \leq l\}$ we have $l \geq k$.

This means that one cannot find $s < k$, such that $p(x;\theta)$ can be written in the form (1.1) with s replacing k and some new statistics $T'_1, \ldots, T'_s$, new functions $\eta'_1, \ldots, \eta'_s$ and B' on Θ and a new function h'.

It can be easily seen that if the $\{T_i\}_{i=1}^k$ satisfy some ***affine*** relationship of the form $\sum_i c_i T_i(x) = c_0$ for all $x \in \mathcal{A}$, then we can rewrite one of the T_i in terms of the rest and a constant, reducing k by 1. Similarly if the $\{\eta_i\}_{i=1}^k$ satisfy an affine relationship. In fact we can keep reducing k until $\{\eta_i\}_{i=1}^k$ and $\{T_i\}_{i=1}^k$ become ***affinely independent***.

Definition 1.7. The functions $T_1, \ldots, T_n$ are called ***affinely independent*** (***$\mathcal{P}$-affine independent*** in [LZ16]) if for any[a] $c_0, \ldots, c_n \in \mathbb{R}$,

$$\left(\sum_{j=1}^n c_j T_j(x) = c_0 \ \mu\text{-almost everywhere}\right) \implies \Big(c_j = 0 \text{ for } j = 0, \ldots, n\Big).$$

The functions $\eta_1, \ldots, \eta_n$ are ***affinely independent*** if

$$\left(\sum_{j=1}^n c_j \eta_j(\theta) = c_0 \ \forall \theta \in \Theta\right) \implies \Big(c_j = 0 \text{ for } j = 0, \ldots, k\Big).$$

[a] Recall from Corollary 1.4 that $\mathcal{A}$ denotes the common support of the exponential family defined in Corollary 1.4.

Remark. In the case of the functions $T_1, \ldots, T_n$ it may help to intuitively understand affine independence as saying that $c_0, \ldots, c_n \in \mathbb{R}$,

$$\left(\sum_{j=1}^n c_j T_j(x) = c_0 \ \forall x \in \mathcal{A}\right) \implies \Big(c_j = 0 \text{ for } j = 0, \ldots, n\Big),$$

where $\mathcal{A}$ was defined in Corollary 1.4 as the common support of the exponential family—notice that $P_\theta(\mathcal{A}) = 1$ for all θ.

Remark. If the functions $\{\eta_j(\cdot)\}$ are not affinely-independent then they are contained in a $k-1$-dimensional hyperplane. Similarly for $\{T_j(\cdot)\}$, ignoring a set of measure zero.

Remark. It is easy to see that *affine independence* is stronger than *linear independence*. For example the functions $\{x, x+1\}$ are linearly independent viewed as functions $\mathbb{R} \mapsto \mathbb{R}$, but not affine independent. Affine independence of $\{f_1, \ldots, f_k\}$ means that $\{f_1, \ldots, f_k, \mathbb{1}\}$ are linearly independent where $\mathbb{1}$ denotes the constant function.

We argued earlier that we iteratively exploit any affine relationships, reducing the dimension of the representation, until we arrive at an affinely independent representation. We will now establish that *every* affinely independent representation has the same dimension.

Proposition 1.8. *For every exponential family $\mathbb{P} = \{p(x;\theta) : \theta \in \Theta\}$, with $p(x;\theta)$ of the form (1.1), there exists a $k' \leq k$ such that (1.1) has a k'-parameter, affinely independent representation. Any affinely independent representation has dimension k'.*

Proof. First of all we can always arrive at an affinely independent representation by iteratively exploiting affine relationships until none exist. Therefore we may assume that $\mathbb{P}$ is given by (1.1),

with (1.1) affinely independent. Define

$$\mathcal{H} := \{(\eta_i(\theta))_{i=1}^k : \theta \in \Theta\} \subset \mathbb{R}^k.$$

By definition, since $\{\eta_i\}_i$ are affinely independent, $\mathcal{H}$ is not contained in any $(k-1)$-dimensional hyperplane. Consider the collection $\mathcal{R}$ of log-likelihood ratios

$$\mathcal{R} := \{x \mapsto \log p(x;\theta) - \log p(x;\theta') : \theta, \theta' \in \Theta\},$$

where

$$\log \frac{p(x;\theta)}{p(x;\theta')} = \sum_{i=1}^k (\eta_i(\theta) - \eta_i(\theta'))T_i(x) + B(\theta') - B(\theta).$$

Then let V be the vector space spanned by $\mathcal{R}$ and the constant function $\mathbf{1}_{\mathcal{X}}$,

$$V := \left\{ \beta_0 + \sum_{i=1}^m \beta_i f_i : m \in \mathbb{N}, \beta_i \in \mathbb{R}, f_i \in \mathcal{R} \right\}.$$

Clearly $V \subset W$ where $W := \mathsf{span}(\mathbf{1}_{\mathcal{X}}, T_1, \ldots, T_k)$, and by affine independence of $\{T_i\}$, W is $(k+1)$-dimensional. In addition, since $\{\eta_i\}$ are affinely independent, $\mathcal{H}$ is not contained in any $(k-1)$-dimensional hyperplane. By affine independence again, the collection

$$\left\{ x \mapsto \sum_{i=1}^k \beta_i T_i(x) : \beta = (\beta_i) \in \mathcal{H} \right\},$$

does not lie in any $(k-1)$-dimensional hyperplane of W and therefore V does not lie in any k-dimensional hyperplane of W. Therefore $V = W$. Finally notice that the likelihood ratios are independent of the particular representation chosen in (1.1), in particular independent of the reference measure and of h. Therefore so is the vector space V. Since V is also independent of the choice of T, k is determined uniquely. □

Proposition 1.9. *Suppose that $\mathcal{P}$ is given by (1.1). Then $\mathcal{P}$ is strictly k-parameter if and only if in (1.1) the functions $\{\eta_i : i \le k\}$ and the statistics $\{T_i : i \le k\}$ are affinely independent.*

Proof. One direction is obvious, since if the functions $\{\eta_i(\cdot)\}$ and the statistics $\{T_i(\cdot)\}_i$ in a k-parameter family are not affinely independent then we can find a $(k-1)$-dimensional representation and therefore the family is not strictly k-parameter.

For the other direction, we assume that the exponential family $\mathcal{P}$ is given by $\mathcal{P} = \{f(x;\theta) : \theta \in \Theta\}$, where

$$f(x;\theta) = h(x) \exp\left\{ \sum_{i=1}^k \eta_i(\theta) T_i(x) - B(\theta) \right\}, \quad \theta \in \Theta \tag{1.12}$$

where the functions $\{\eta_i(\cdot)\}$ and the statistics $\{T_i(\cdot)\}_i$ are affinely independent. We need to prove that $\mathcal{P}$ is strictly k-parameter.

We can argue in two ways.

Way 1: Notice that any minimal representation must be affinely independent, since otherwise we can always reduce the dimension violating minimality. We also know from Proposition 1.8 that any affinely independent representation has the same dimension k. Recall from the proof of Proposition 1.8 that the vector space V spanned by the collectino of the likelihood ratios $\{x \mapsto \log p(x;\theta)/p(x;\theta') : \theta, \theta' \in \Theta\}$ is $(k+1)$-dimensional and is independent of the representation. Therefore any representation must have dimension at least k.

Way 2: Suppose that $\mathcal{P}$ also has another representation

$$\mathcal{P} = \{f(x;\theta) : \theta \in \Theta\} = \{g(x;\theta) : \theta \in \Theta\}$$

where

$$g(x;\theta) = \tilde{h}(x) \exp\left\{\sum_{i=1}^{l} \tilde{\eta}_i(\theta)\tilde{T}_i(x) - \tilde{B}(\theta)\right\}, \quad \theta \in \Theta. \tag{1.13}$$

Next notice that likelihood ratios are independent of the parameterisation. Then we have

$$\frac{f(\theta,x)f(\theta_0,x_0)}{f(\theta_0,x)f(\theta,x_0)} = \frac{h(x)\exp\left\{\eta(\theta)\cdot t(x) - B(\theta)\right\} \times h(x_0)\exp\left\{\eta(\theta_0)\cdot t(x_0) - B(\theta_0)\right\}}{h(x)\exp\left\{\eta(\theta_0)\cdot t(x) - B(\theta_0)\right\} \times h(x_0)\exp\left\{\eta(\theta)\cdot t(x_0) - B(\theta)\right\}} \tag{1.14}$$

$$= \exp\left\{[\eta(\theta) - \eta(\theta_0)]\cdot[t(x) - t(x_0)]\right\}, \tag{1.15}$$

is independent of the parameterisation.

Fix a $P_0 \in \mathcal{P}$ corresponding to θ_0, θ_0. For $\mathcal{P}$, given equivalently by (1.12) and (1.13) with θ and θ respectively, we have that

$$\left[\eta(\theta) - \eta(\theta_0)\right]\cdot\left[T(x) - T(x_0)\right] = \left[\tilde{\eta}(\theta) - \tilde{\eta}(\theta_0)\right]\cdot\left[\tilde{T}(x) - \tilde{T}(x_0)\right]. \tag{1.16}$$

Since by assumption the functions $\{\eta_i(\cdot) : i = 1,\dots,k\}$ are affinely independent, they are not contained in any $(k-1)$-dimensional hyperplane. Therefore we can find $\theta_0,\dots,\theta_k$ such that the vectors $\eta(\theta_0),\dots,\eta(\theta_k)$ are not contained in any $(k-1)$-dimensional hyperplane and therefore the vectors $\eta(\theta_1) - \eta(\theta_0),\dots,\eta(\theta_k) - \eta(\theta_0)$ span $\mathbb{R}^k$. Applying with $\theta = \theta_i$, $i = 1,\dots,k$, we get k equations which in matrix notation can be written as

$$\begin{pmatrix} \eta(\theta_1) - \eta(\theta_0) \\ \vdots \\ \eta(\theta_k) - \eta(\theta_0) \end{pmatrix} \times \left[T(x) - T(x_0)\right] = \begin{pmatrix} \tilde{\eta}(\theta_1) - \tilde{\eta}(\theta_0) \\ \vdots \\ \tilde{\eta}(\theta_k) - \tilde{\eta}(\theta_0) \end{pmatrix} \times \left[\tilde{T}(x) - \tilde{T}(x_0)\right] \tag{1.17}$$

where $\eta, \tilde{\eta}$ are written as row vectors and $T, \tilde{T}$ are column vectors.

Since $\eta(\theta_i) - \eta(\theta_0), i = 1,\dots,k$ span $\mathbb{R}^k$ we can invert the left-most matrix and obtain an equation of the form $T(x) = A\tilde{T}(x) + b$, where A is a constant (in x) $k \times l$ matrix and b a constant vector. Notice that since $\{T(x) : x \in \mathcal{X}\}$ is not contained in any $(k-1)$-dimensional hyperplane, the image of A must have dimension k; therefore $\mathsf{rank}(A) = k$ and thus $l \geq k$.

Using the same reasoning, and the fact that $\{T_i(\cdot) : i = 1,\dots k\}$ are affinely independent, we can find $x_0,\dots,x_k$ so that $T(x_i) - T(x_0), i = 1,\dots,k$ span $\mathbb{R}^k$ and working in a similar way as before we can obtain an equation of the form $\eta(\theta) = C\tilde{\eta}(\theta) + d$ with C a constant $k \times l$ matrix and d a constant vector.

This proves that if in (1.12) $\{\eta_i(\cdot) : i = 1,\dots,k\}$ and $\{T_i(\cdot) : i = 1,\dots,k\}$ are affinely independent, then any other representation must have dimension $l \geq k$ and therefore $\mathcal{P}$ is strictly k-parameter. □

Remark. We have now seen that a representation is minimal if and only if it is affinely independent. In fact, often in the literature, a representation is called minimal if it is affinely independent.

If $X \sim f(x;\theta)$, then $T = (T_1(X),\dots,T_N(X))$ is a random vector. Let $\mathrm{Cov}_\theta(T)$ be its covariance matrix under $f(x;\theta)$. The following gives a condition for the statistics $\{T_i\}$ to be affine-independent.

Proposition 1.10. *The functions T_i are $\mathcal{P}$-affinely independent if and only if for some θ, and thus for all θ, $\mathrm{Cov}_\theta(T)$ is positive definite.*

Proof. As before let $X \sim f(x;\theta)$. First or all notice that the following are all equivalent:

(i) $\mathrm{Cov}_\theta(T)$ is positive definite;

(ii) for all $\eta \neq 0$
$$\sum_{ij} \eta_i \,\mathrm{Cov}_\theta(T)_{ij}\eta_j = \mathrm{Var}_\theta\left(\sum \eta_i T_i(X)\right) > 0;$$

(iii) for all $\eta \neq 0$ the mapping $x \mapsto \sum \eta_j T_j(x)$ is not P_θ-a.s. constant;

(iv) for all $\eta \neq 0$ we have $P_\theta^{\otimes 2}(X_\eta) > 0$ where

$$X_\eta := \left\{(x,x') \in \mathcal{X}^2 : \sum \eta_j \left(T_j(x) - T_j(x')\right) \neq 0\right\}, \tag{1.18}$$

and $P_\theta^{\otimes 2} = P_\theta \otimes P_\theta$ is the product measure.

Since, for all θ', $P_{\theta'}$ is equivalent to P_θ, it easily follows that $P_{\theta'}^{\otimes 2}$ is equivalent to $P_\theta^{\otimes 2}$; it suffices to write down the Radon-Nikodym derivative. Therefore $P_{\theta'}^{\otimes 2}(X_\eta) > 0$ for all θ', and therefore by the above equivalences we have that $\mathrm{Cov}_{\theta'}(T)$ is positive definite for all θ'.

Suppose first that the T_j are affine independent. Then for any $\eta \neq 0$, we have that $x \mapsto \sum_j \eta_j T_j(x)$ cannot be constant μ- and thus also P_θ-almost everywhere for any θ; equivalently for all $\eta \neq 0$ and θ we have $P_\theta^{\otimes 2}(X_\eta) > 0$ with X_η as in (1.18). Therefore for any $\eta \neq 0$ and any θ, letting X, X' be i.i.d. from P_θ we have

$$0 < \mathbb{E}\left[\left(\sum \eta_j T_j(X) - \sum \eta_j T_j(X')\right)^2\right] = 2\,\mathrm{Var}_\theta\left(\sum \eta_j T_j(X)\right) = 2\sum_{ij} \eta_i \,\mathrm{Cov}_\theta(T)_{ij}\eta_j. \tag{1.19}$$

This proves that $\mathrm{Cov}_\theta(T)$ is positive definite for all θ.

For the other direction suppose that the T_j are not affine independent: there exists an $\eta \neq 0$ such that $x \mapsto \sum \eta_j T_j(x)$ is μ-a.e. constant, and therefore $P_\theta^{\otimes 2}(X_\eta) = 1$ for all θ. Therefore

$$0 = \mathbb{E}\left[\left(\sum \eta_j T_j(X) - \sum \eta_j T_j(X')\right)^2\right] = 2\,\mathrm{Var}_\theta\left(\sum \eta_j T_j(X)\right) = 2\sum_{ij} \eta_i \,\mathrm{Cov}_\theta(T)_{ij}\eta_j, \tag{1.20}$$

and thus we see that Cov_θ is not positive definite. □

Remark. Above we used the fact that with Z, Z' i.i.d. $\mathbb{E}[(Z-Z')^2] = 2\,\mathrm{Var}(Z)$ and that if $Z = Z'$ a.s. with Z, Z' i.i.d. then Z, Z' are almost surely constant by the equality case of Jensen's inequality since $\mathbb{E}[Z^2] = \mathbb{E}[Z \times Z'] = \mathbb{E}[Z]^2$.

Example. Suppose X takes values in $\{1,2,3\}$ with $\mathbb{P}(X = i) = p_i$ for $i = 1,2,3$, so that $\theta = (p_1, p_2, p_3)$. Then

$$p(x;\theta) = p_1^{I_1(x)} p_2^{I_2(x)} p_3^{I_3(x)} \qquad \text{where } I_i(x) := \mathbb{1}_{x=i} \tag{1.21}$$
$$= \exp(I_1(x)\log(p_1) + I_2(x)\log(p_2) + I_3(x)\log(p_3)), \tag{1.22}$$

so X belongs to a 3-parameter exponential family, ***but*** $I_1(x) + I_2(x) + I_3(x) = 1$ so it is *not* strictly 3-dimensional. Indeed,

$$p(x;\theta) = \exp\left(I_1(x)\log\left(\frac{p_1}{1-(p_1+p_2)}\right) + I_2(x)\log\left(\frac{p_2}{1-(p_1+p_2)}\right) + \log(p_3)\right)$$

so it is a strictly 2-dimensional exponential family.

1.3 The parameter space

Definition 1.11. The ***parameter space*** is defined to be

$$\Theta := \left\{ \theta : \int h(x) \exp\left[\textstyle\sum_{i=1}^n \eta_i(\theta) T_i(x)\right] \mathrm{d}x < \infty \right\},$$

i.e. the set of θ for which the integrand can be normalized to become a probability density function.

Definition 1.12. The ***natural parameter space*** is defined to be

$$\Xi := \left\{ \eta = (\eta_1, \ldots, \eta_n) : \int h(x) \exp\left[\textstyle\sum_{i=1}^n \eta_i T_i(x)\right] \mathrm{d}x < \infty \right\},$$

i.e. the set of η for which we can define $B(\eta) := \log \int h(x) \exp\left[\sum_{i=1}^n \eta_i T_i(x)\right] \mathrm{d}x$ so that

$$\tilde{f}(x;\eta) = e^{-B(\eta)} h(x) \exp\left[\textstyle\sum_{i=1}^n \eta_i T_i(x)\right]$$

is a pdf/pmf on $\mathcal{X}$.

Observe that you we always have $\eta(\Theta) \subset \Xi$, although it may be the case that $\eta(\Theta) \neq \Xi$).

Theorem 1.13. *The natural parameter space Ξ of a strictly k-parameter exponential family is convex and contains a non-empty k-dimensional ball.*

Proof. Take $\eta, \eta' \in \Xi$ and let $\alpha \in (0,1)$. Define $B(\eta) = \log \int \exp\big(\sum_i \eta_i T_i(x)\big) h(x) \, \mathrm{d}x$. Then

$$B(\alpha\eta + (1-\alpha)\eta') = \log \int \exp(\alpha \textstyle\sum_i \eta_i T_i(x) + (1-\alpha) \sum_i \eta_i' T_i(x)) h(x) \, \mathrm{d}x \tag{1.23}$$

$$= \log \int \left[\exp(\textstyle\sum_i \eta_i T_i(x)) h(x)\right]^{\alpha} \left[\exp(\textstyle\sum_i \eta_i' T_i(x)) h(x)\right]^{1-\alpha} \mathrm{d}x \tag{1.24}$$

$$\text{(using } h = h^{\alpha} h^{1-\alpha}\text{)} \tag{1.25}$$

$$\leqslant \log \left(\int \exp(\textstyle\sum_i \eta_i T_i(x)) h(x) \, \mathrm{d}x \right)^{\alpha} \left(\int \exp(\textstyle\sum_i \eta_i' T_i(x)) h(x) \, \mathrm{d}x \right)^{1-\alpha} \tag{1.26}$$

$$\text{by Hölder's inequality} \tag{1.27}$$

$$= \alpha B(\eta) + (1-\alpha) B(\eta') < \infty. \tag{1.28}$$

Notice that if Ξ is contained in a $(k-1)$-dimensional subspace, then so would $\eta(\Theta) \subset \Xi$, which is impossible since $\{\theta \mapsto \eta_i(\theta) : i = 1, \ldots, k\}$ are affinely independent. Therefore Ξ is not contained in any $(k-1)$-dimensional subspace and must therefore contain $k+1$ points, say $\xi_1, \ldots, \xi_{k+1}$, not all lying in the same $(k-1)$-subspace; since Ξ is convex it must also contain the convex hull of these points; since they don't all lie in the same $(k-1)$-subspace Ξ contain a non-empty, open k-dimensional ball. □

Definition 1.14. If the image of the parameter space $\eta(\Theta) \subseteq \Xi$ for a strictly k-parameters exponential family contains a k-dimensional open set, then it is called ***full rank***.

Full rank is also sometimes referred to as ***regular***. See the discussion of curved exponential families below for examples of families which are not full rank.

Remark. In the one dimensional case, being full-rank is easily checked: it suffices that $\eta(\Theta)$ contains an interval.

Theorem 1.15. *Let $\mathcal{P}$ be a strictly k-parameter exponential family with natural parameter space Ξ. Then for all $\eta \in \mathrm{Int}(\Xi)$:*

(a) all moments of T (with respect to $f(x;\eta)$) exist;

(b) $\mathbb{E}_\eta[T_i(X)] = \dfrac{\partial}{\partial \eta_i} B(\eta)\ \forall i$; *and*

(c) $\mathrm{Cov}_\eta(T_i, T_j) = \dfrac{\partial^2}{\partial \eta_i \partial \eta_j} B(\eta)\ \forall i, j.$

Proof. Recall that

$$\exp(B(\eta)) = \int \exp\left(\sum_{i=1}^{k} \eta_i T_i(x)\right) h(x)\mathrm{d}x.$$

Then for $s = (s_1, \ldots, s_k)$ consider the moment generating function of $T(X)$, when $X \sim P_\eta$,

$$M_{T(X)}(s) = \mathbb{E}_\eta\left[\exp\left(\sum_{i=1}^{k} s_i T_i(X)\right)\right] \tag{1.29}$$

$$= \int \exp\Big(\sum_{i=1}^{k} (\eta_i + s_i) T_i(X) - B(\eta)\Big) h(x)\mathrm{d}x \tag{1.30}$$

$$= \exp\left\{B(\eta + s) - B(\eta)\right\}. \tag{1.31}$$

Since by assumption $\eta \in \mathsf{Int}(\Xi)$, there exists a $\delta > 0$, such that for all $y \in B(\eta, \delta)$ we have $y \in \Xi$ and therefore for all $|s| < \delta$ we have

$$M_{T(X)}(s) = \exp\left[-B(\eta) + B(\eta + s)\right] < \infty.$$

This implies all moments are finite.

For the other two parts, start by differentiating $\exp(B(\eta))$ to get

$$\frac{\partial \exp(B(\eta))}{\partial \eta_1} = \lim_{s \to 0} \frac{1}{s} \int \exp\Big(\sum_{i=1}^{k} \eta_i T_i(X)\Big) \left[\exp(sT_1(x)) - 1\right] h(x)\mathrm{d}x \tag{1.32}$$

$$\exp(B(\eta)) \frac{\partial B(\eta)}{\partial \eta_1} = \int \exp\Big(\sum_{i=1}^{k} \eta_i T_i(X)\Big) T_1(x) h(x)\mathrm{d}x \tag{1.33}$$

$$\frac{\partial B(\eta)}{\partial \eta_1} = \mathbb{E}_\eta[T_1(X)], \tag{1.34}$$

where the differentiation under the integral sign may be justified by dominated convergence. The last statement follows by differentiating once more. □

1.4 Curved exponential families

Definition 1.16. A family $\mathcal{P} = \{p(x;\theta) : \theta \in \Theta\}$ of probabilities (pmf or pdf) indexed by θ is called a ***curved exponential family*** if there exists $q < k \in \mathbb{N}$, real-valued functions $\eta_1, \ldots, \eta_k$ and B on $\Theta \subseteq \mathbb{R}^q$, real-valued statistics $T_1, \ldots, T_k$ and a non-negative real-valued function h on $\mathcal{X}$ such that

1. $\exists \theta : \mathrm{Cov}_\theta T$ is positive definite

2. the pdf/pmfs $p(x;\theta)$ have the form

$$p(x;\theta) = \exp\left[\sum_{i=1}^{k} \eta_i(\theta)T_i(x) - B(\theta)\right] h(x). \tag{1.35}$$

Positive-definiteness of the covariance matrix guarantees that k is not an arbitrary large number. In fact it guarantees that the exponential family is strictly k-parameter, although the parameter space forms a lower-dimensional, non-linear submanifold of the ***natural*** parameter space.

Example. (Normal Distribution). Let the statistical model be the class of all normal distributions with $N(\mu, \mu^2)$ where μ is unknown and $\mu \neq 0$ with parameter $\theta = \mu \in \mathbb{R}^*$.

(1.36)
$$p(x;\theta) = \frac{1}{\sqrt{2\pi}\mu} \exp\left\{-\frac{(x-\mu)^2}{2\mu^2}\right\} = \frac{1}{\sqrt{2\pi}\mu} \exp\left\{\frac{(-x^2+2x\mu-\mu^2}{2\mu^2}\right\} = \frac{1}{\sqrt{2\pi}\mu} \exp\left\{--\frac{x^2}{2\mu^2} + \frac{x}{\mu} - \frac{1}{2}\right\}$$

We thus have

$$T_1(x) = x, T_2(x) = x^2, \eta_1(\theta) = \mu^{-1}, \eta_2(\theta) = -\mu^{-2}/2.$$

This examples satisfies the definition of Curved Exponential families because the parameter θ is one dimensional but results in a 2-parameter exponential family. The covariance matrix can be calculated with the known moments

$$\mathrm{Cov}_\theta T = \begin{pmatrix} \mu^2 & 2\mu^3 \\ 2\mu^3 & 6\mu^4 \end{pmatrix}$$

The covariance matrix is shown to be positive definite for all $\theta \in \Theta$ (the determinant is $2\mu^6 > 0$). We see that T_1 and T_2 are $\mathcal{P}$-affine independent and the familly is strictly 2-parameters (the η_i are constrained, but not linearly).

Example. Suppose $X_1 \sim \mathcal{N}(\theta, 1)$ and $X_2 \sim \mathcal{N}(\frac{1}{\theta}, 1)$ are independent. Their joint distribution has log-density

$$\log f(x;\theta) = -\frac{(x_1-\theta)^2}{2} - \frac{(x_2-\frac{1}{\theta})^2}{2} + \text{constant} \tag{1.37}$$

$$= x_1\theta + x_2\frac{1}{\theta} - \frac{\theta^2}{2} - \frac{\theta^{-2}}{2} + \text{terms in } (x_1, x_2) \text{ alone}, \tag{1.38}$$

so that $\eta_1 = \theta$, $\eta_2 = \frac{1}{\theta}$, $T_1 = x_1$ and $T_2 = x_2$. This is a $(2,1)$-curved family.

Observe that $\eta(\Theta) = \{(\theta, \frac{1}{\theta}) \in \mathbb{R}^2 : \theta \in \mathbb{R} \setminus \{0\}\}$ is a one-dimensional manifold.

Further examples can be found in this document. See also here for some examples about curved/full-rank families.

Finally, let us formulate the following important statement about the distribution of a sample of independent r.v.s distributed according to a distribution from an exponential family:

Theorem 1.17. *(a) If $X_1, \ldots, X_n$ is a sample of independent r.v.s with distributions belonging to an exponential family, then the joint distribution of the vector $\mathbf{X} = (X_1, ..., X_n)$ is an element of an exponential family.*

(b) If $\mathbf{X} = (X_1, \ldots, X_n)$ are i.i.d. samples from a k-parameter exponential distribution of the form (1.1) with functions $\eta = (\eta_1, \ldots, \eta_k)$ and $T = (T_1, \ldots, T_k)$, then the distribution of $\mathbf{X}$ belongs

to a a k-parameter exponential family with natural observation $T_{(n)}(\mathbf{x}) := \sum_{i=1}^{n} T(x_i)$.

Remark. The fact that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ also belongs to a k parameter exponential family, independently of n, is quite important.

Proof. Left as an exercise. □

Chapter 2

Sufficiency and Minimality

2.1 Sufficiency

We may often be interested in summarising a set of data without losing any information about the parameter we're trying to estimate. A statistic that does this is said to be ***sufficient***:

Definition 2.1. Suppose $X \sim f(x;\theta)$ for some parameter θ.

A ***statistic*** $T(X)$ is a function of the data which does not depend on θ.

A statistic $T(X)$ is said to be ***sufficient*** for θ if the conditional distribution of X given T does not depend on θ. That is,

$$f(x \mid t, \theta) = f(x \mid t).$$

Remark. In particular, this means that for any function g the map $\theta \mapsto \mathbb{E}_\theta[g(X) \mid T = t]$ is constant.

We can think of a sufficient statistic as 'wrapping up' all the information there is about θ somehow.

Example. Let $X_1, \dots, X_n$ be independent $\text{Ber}(p)$ random variables, so that $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$, and let $T = \sum_{i=1}^n X_i$, so that $T \sim \text{Bin}(n, p)$. Then, writing $X = (X_1, \dots, X_n)$, for any $x \in \{0,1\}^n$ and $t \in \{0, \dots, n\}$ we have

$$f(x \mid t, p) = \mathbb{P}(X = x \mid T = t, p) = \frac{\mathbb{P}(X = x, T = t \mid p)}{\mathbb{P}(T = t \mid p)} \tag{2.1}$$

$$= \frac{\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} \mathbb{1}_{\sum x_i = t} \tag{2.2}$$

$$= \frac{p^t(1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} \mathbb{1}_{\sum x_i = t} = \binom{n}{t}^{-1} \mathbb{1}_{\sum x_i = t}, \tag{2.3}$$

which has no dependence on p. So T is sufficient for p.

The intuitive meaning of this is that only the *number* of successes matters for estimating p; the order in which successes arrive shouldn't change your guess for p.

Theorem 2.2 (The Factorisation Criterion). *Suppose $X \sim f(x;\theta)$. Then a statistic $T(X)$ is sufficient for θ if and only if f can be written as*

$$f(x;\theta) = g(T(x), \theta)h(x)$$

for some non-negative functions g, h.

Proof for the discrete case. Suppose T is sufficient and write $t = T(x)$. So

$$f(x;\theta) = \mathbb{P}_\theta(X = x) = \mathbb{P}_\theta(X = x, T = t) = \mathbb{P}_\theta(X = x \mid T = t)\,\mathbb{P}_\theta(T = t).$$

Then just note that as T is sufficient, $\mathbb{P}_\theta(X = x \mid T = t) =: h(x)$ is independent of θ, and $\mathbb{P}_\theta(T = t) =: g(t, \theta)$ only depends on t and θ.

Conversely, suppose $f(x;\theta) = g(t,\theta)h(x)$ for some non-negative functions g, h. So

$$\mathbb{P}_\theta(T = t) = \sum_{x:T(x)=t} \mathbb{P}_\theta(X = x) = \sum_{x:T(x)=t} f(x;\theta) = g(t,\theta) \sum_{x:T(x)=t} h(x).$$

Thus $\mathbb{P}_\theta(X = x \mid T = t) = \frac{\mathbb{P}_\theta(X=x,T=t)}{\mathbb{P}_\theta(T=t)} = \frac{\mathbb{P}_\theta(X=x)}{\mathbb{P}_\theta(T=t)} = \frac{h(x)}{\sum_{y:T(y)=t} h(y)}$, which has no dependence on θ! So T is sufficient for θ.

The general case is much more complicated to prove; see [CP97, Example 6] for a proof using disintegrations. □

The following corollary is obvious.

Corollary 2.3. *Let $\mathcal{P} = \{f(\cdot;\theta) : \theta \in \Theta\}$ be a k-parameter exponential family, where*

$$f(x;\theta) = \exp\left\{\sum_{i=1}^{k} \eta_i(\theta)T_i(x) - B(x)\right\} h(x).$$

Then $\big(T_1(x), \ldots, T_k(x)\big)$ is sufficient for θ.

Remark. An important case of the above corollary is the distribution of $\boldsymbol{X} = (X_1, \ldots, X_n)$ are i.i.d. samples from an exponential family $\mathcal{P} = \{f(\cdot;\theta) : \theta \in \Theta\}$. In that case the above corollary implies that $\boldsymbol{T}_{(n)}(\boldsymbol{x}) = \big(\sum_{i=1}^n T_1(x_i), \ldots, \sum_{i=1}^n T_k(x_i)\big)$ is sufficient.

The next natural question to ask is to what extent we can summarise a set of data — by how much we can reduce it — without losing information about θ. This brings us to the concept of ***minimal sufficiency***.

Example. Let X_1, X_2, X_3 be independent $\mathrm{Ber}(p)$ random variables modelling three coin tosses (so 0 means heads and 1 means tails). Consider the following four statistics:

1. $T_1(X) = (X_1, X_2, X_3)$,
2. $T_2(X) = (X_1, \sum_{i=1}^3 X_i)$,
3. $T_3(X) = \sum_{i=1}^3 X_i$,
4. $T_4(X) = \mathbb{1}_{T_3(X)=0}$.

Which of these are sufficient for p?

2.2 Minimality

Suppose X takes values in $\mathcal{X}$. A statistic $T : \mathcal{X} \mapsto \mathcal{T}$ induces a partition of the sample space $\mathcal{X}$

$$\mathcal{X} = \bigcup_{t\in\mathcal{T}} \{x \in \mathcal{X} : T(x) = t\};$$

the partition is equivalently generated by the equivalence relation $x \sim y \iff T(x) = T(y)$.

Example (continued). The following diagrams show the partitions induced by the statistics $T_1, \dots, T_4$:

HHH	THT	HTT	HTH
TTH	THH	HHT	TTT

1. $T_1(X) = (X_1, X_2, X_3)$

HHH	THT	HTT	HTH
TTH	THH	HHT	TTT

2. $T_2(X) = \left(X_1, \sum_{i=1}^3 X_i\right)$

HHH	THT	HTT	HTH
TTH	THH	HHT	TTT

3. $T_3(X) = \sum_{i=1}^3 X_i$

HHH	THT	HTT	HTH
TTH	THH	HHT	TTT

4. $T_4(X) = \mathbb{1}_{T_3(X)=0}$

In each case T is constant within each class.

Summarising our data through a statistic can be thought of as keeping track only of the equivalence class which contains our sample. Therefore a statistic T (equivalently a partition of the sample space) is sufficient, if the conditional distribution of X given the equivalence class it belongs to is independent of the parameter θ, i.e. the same for all distributions in our model.

A finer partition corresponds to keeping "more information" about our sample. We want to be as economical as possible, that is keep as little information about the sample without losing any information about the parameter θ. In other words, among all sufficient statistics, we want to choose the one generating the coarsest partition.

Consider the case where $\mathcal{X}$ is a finite set, and let Π, Π' be two partitions, such that Π' is finer than Π, that is each element of Π can be written as a union of elements of Π'. Equivalently, we can express this as a function sending multiple elements of Π' to one element of Π. This leads us to the following definition.

Definition 2.4. A statistic is ***minimal sufficient*** if it can be expressed as a function of any other sufficient statistic.

The following result gives a way of checking minimal sufficiency.

Theorem 2.5. *A statistic T is minimal sufficient if and only if*

$$T(x) = T(y) \iff \frac{f(y;\theta)}{f(x;\theta)} \text{ is independent of } \theta.$$

Remark. It is useful to think about the above result in terms of partitions: it says that in order to define a partition corresponding to a minimal sufficient statistic, the likelihood ratio for any two elements in the same class must be independent of θ.

Example (continued). In the coin-tossing example, first consider T_2. With $x = \text{TTH}$ and $y = \text{HTT}$, we have $f(x;p) = f(y;p) = p^2(1-p)$, so that $\frac{f(x,p)}{f(y,p)} = 1$, but clearly $T_2(X) \neq T_2(Y)$, so T_2 is not minimal sufficient.

Considering T_4 instead, take $x = \text{HTH}$ and $y = \text{TTT}$. So clearly $T_4(x) = T_4(y)$, but $\frac{f(y;p)}{f(x;p)} = \frac{p^3}{p(1-p)^2} = \frac{p^2}{(1-p)^2}$, which does depend on p. So T_4 is also not minimal sufficient.

Proof of theorem. ($\Longleftarrow$) Suppose T is a statistic such that $T(x) = T(y)$ if and only if $\frac{f(y;\theta)}{f(x;\theta)}$ is equal to some $k(x,y)$ independent of θ.

Sufficiency. In the discrete case,

$$f(x \mid t, \theta) = \mathbb{P}_\theta(X = x \mid T = t) = \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}_\theta(T = t)} = \frac{f(x;\theta)}{\sum_{y:T(y)=t} f(y;\theta)} \tag{2.4}$$

$$= \frac{f(x;\theta)}{\sum_{y:T(y)=t} f(x;\theta)k(x,y)} \tag{2.5}$$

$$= \left(\sum_{y:T(y)=t} k(x,y) \right)^{-1} \tag{2.6}$$

which is independent of θ, so T is sufficient. For the continuous case, replace the sum with an integral.

Minimality. Now suppose $U : \mathcal{X} \mapsto \mathcal{U}$ is another sufficient statistic and that $U(x) = U(y)$ for some x, y. Since U is sufficient, by the factorisation criterion we have

$$\frac{f(y;\theta)}{f(x;\theta)} = \frac{g(U(y),\theta)h(y)}{g(U(x),\theta)h(x)} = \frac{h(y)}{h(x)}$$

which is independent of θ. So by hypothesis, $T(x) = T(y)$ and hence Π_U is finer than Π_T, where Π_U, Π_T be the partitions induced by U, T respectively. We now show that T must be a function of U. Without loss of generality we assume that

$$\mathcal{U} = \bigcup_{A \in \Pi_U} \{U(x) : x \in A\}, \qquad \mathcal{T} = \bigcup_{B \in \Pi_T} \{T(x) : x \in B\}.$$

We define a function $\phi : \mathcal{U} \mapsto \mathcal{T}$ as follows: for each $u \in \mathcal{U}$, there exists an $A \in \Pi_U$ such that $u = U(x)$ for all $x \in A$ and a $t \in \mathcal{T}$ such that for all $x \in A$, $T(x) = t$; let $\phi(u) = t$. In this way $T(x) = \phi(U(x))$ Hence T is minimal sufficient.

($\implies$) Conversely, suppose T is minimal sufficient. Take x, y such that $T(x) = T(y)$. Then by the factorisation criterion,

$$\frac{f(y;\theta)}{f(x;\theta)} = \frac{g(T(y),\theta)h(y)}{g(T(x),\theta)h(x)} = \frac{h(y)}{h(x)}$$

which does not depend on θ. (Note this only used the sufficiency of T.)

For the other direction, we need to show that if $f(x;\theta)/f(y;\theta)$ is independent of θ then $T(x) = T(y)$. Start by writing $x \sim y$ whenever $f(x;\theta) = k(x,y)f(y;\theta)$ for all θ (for some function $k(x,y)$). It is easy to check that this is an equivalence relation. For each equivalence class $[x]$ choose a representative $\overline{x}$ and define G to be the representative function (i.e. $G(y) = \overline{x}$ for all $y \in [x]$). So G is a statistic constant on the equivalence classes. It is also sufficient, by the factorisation criterion, since $f(x;\theta) = k(x,\overline{x})f(\overline{x};\theta) = k(x,G(x))f(G(x);\theta)$ for all x. So T is a function of G (by minimality) and hence is also constant on the equivalence classes, meaning $x \sim y \implies T(x) = T(y)$. □

2.3 Minimal sufficiency in exponential families

Let us turn to the case of exponential families.

Theorem 2.6. *Suppose the functions* $f(x;\theta) = \exp\left[\sum_{j=1}^k \eta_j(\theta)T_j(x) - B(\theta)\right] h(x)$ *form a strictly* k*-parameter exponential family. Let* $X = (X_1, \ldots, X_n)$*, where* $X_1, \ldots, X_n \sim f(x,\theta)$ *are i.i.d. Then:* $T_{(n)} = \left(\sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_k(X_i)\right)$ *is minimal sufficient.*

Remark. Since the vector $X = (X_1, \ldots, X_n)$ is strictly k-parameter exponential, we could just say $T(X) = (T_1(X), \ldots, T_k(X))$ is minimal sufficient.

Proof of theorem. Just note that

$$\frac{f((x_1,\dots,x_n);\theta)}{f((y_1,\dots,y_n);\theta)} = \frac{\prod_{i=1}^n h(x_i)}{\prod_{i=1}^n h(y_i)} \exp\left[\sum_{j=1}^k \eta_j(\theta)\left(\sum_{i=1}^n T_j(x_i) - \sum_{i=1}^n T_j(y_i)\right)\right]$$

which is independent of θ if and only if $\sum_{i=1}^n T_j(x_i) = \sum_{i=1}^n T_j(y_i)$ for all $j = 1,\dots,k$.

□

Examples.

1. **Bernoulli.** Let $X_1,\dots,X_n$ be i.i.d. Bernoulli trials with parameter p, and let $T(X) = \sum_{i=1}^n X_i$ be the number of successes. Then

$$\frac{f((x_1,\dots,x_n);p)}{f((y_1,\dots,y_n);p)} = \frac{p^{T(x)}(1-p)^{n-T(x)}}{p^{T(y)}(1-p)^{n-T(y)}} = p^{T(x)-T(y)}(1-p)^{T(y)-T(x)}$$

   which is independent of p if and only if $T(x) = T(y)$. So T is minimal sufficient.

2. **Uniform.** Let $X_1,\dots,X_n$ be i.i.d. random variables with $X_i \sim \mathcal{U}[a,b]$, taking the unknown parameter to be $\theta = (a,b)$. Then

$$f((x_1,\dots,x_n);\theta) = \prod_{i=1}^n \frac{1}{b-a}\mathbb{1}_{[a,b]}(x_i) = (b-a)^{-n}\mathbb{1}_{\min x_i \geqslant a}\mathbb{1}_{\max x_i \leqslant b},$$

   so by the factorisation criterion $T(x) = (\min x_i, \max x_i)$ is sufficient.

   Exercise: is it minimal sufficient?

3. **Normal.** Let $X = (X_1,\dots,X_n)$ be a sample of i.i.d. $\mathcal{N}(\mu,\sigma^2)$-distributed random variables. For the parameter $\theta = (\mu,\sigma^2) \in \mathbb{R}\times\mathbb{R}_+$, we have

$$\frac{f(x;\theta)}{f(y;\theta)} = \frac{(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right)}{(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2\right)} \tag{2.7}$$

$$= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2 - 2\mu\left(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i\right)\right)\right). \tag{2.8}$$

   This ratio is independent of θ if and only if $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ and $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$.
   Thus $T(X) = \sum X_i, \sum X_i^2$ is minimal sufficient.

   Note that $\overline{x} = \frac{1}{n}\sum x_i = \frac{T_1(x)}{n}$ and $S^2 = \frac{1}{n-1}\sum (x_i-\overline{x})^2 = \frac{1}{n-1}\left(\sum x_i^2 - n\overline{x}^2\right) = \frac{1}{n-1}(T_2(x) - \frac{1}{n}T_1(x)^2)$ are in one-to-one correspondence with $T(x)$, and hence $(\overline{X}, S^2)$ is also minimal sufficient for θ.

In closing our discussion on sufficiency, it is worth mentioning that although it is a nice property, that allows us to reduce the data to a bounded dimension statistic, it is not all that ***common***. In fact, only exponential families and some families with bounded support admit finite dimensional sufficient statistics for all sample sizes; this is the Pitman-Koopman-Darmois theorem.

Chapter 3

Fisher Information

Much of the material in this chapter can also be found (often with a nicer exposition) in your part A course.

We turn to the question now of whether there is some nice way to measure 'how much' information a given dataset contains about a particular parameter.

Let $f(x, \theta)$ be a parametric family of densities.

Definition 3.1. For each $x \in \mathcal{X}$, the ***likelihood function*** $L(\cdot, x) : \Theta \to \mathbb{R}_+$ is defined by $L(\theta, x) = f(x, \theta)$.

The ***log-likelihood*** is often written $\ell(\theta, x) \coloneqq \log L(\theta, x)$.

To simplify our analysis, we will need some ***regularity assumptions*** about our model. These will, primarily, allow us to use partial derivatives and to interchange them with sums/integrals without worrying too much (as we'll see).

Reg 1. *The distributions* $\{f(\cdot, \theta) : \theta \in \Theta\}$ *have common support, so that* $\mathcal{A} = \{x : f(x, \theta) > 0\}$ *is independent of* θ.

Remark. Distributions belonging to an exponential family satisfy Reg 1.

To proceed, we'll start by just looking at the one-dimensional case.

3.1 The one-dimensional case

Reg 2. $\Theta \subseteq \mathbb{R}$ *is an open interval (finite or infinite).*

Reg 3. *For all* $x \in \mathcal{A}$ *and for all* $\theta \in \Theta$, *the derivative* $\partial_\theta f(x, \theta)$ *exists and is finite. Furthermore, for each* $\theta \in \Theta$, *there exists a neighborhood* $I \subset \Theta$ *and an integrable function* $g_I(x)$ *such that*

$$|\partial_\theta f(x, \theta)| \leqslant g_I(x), \quad \forall \theta \in I, \forall x \in \mathcal{A}.$$

See Part A integration.

The following will be a useful tool to work with:

Definition 3.2. When Regs 1–3 are satisfied, for $x \in \mathcal{A}$ we define the ***score function***

$$S(\theta, x) = \ell'(\theta, x) = \frac{\partial \log L(\theta, x)}{\partial \theta}.$$

Now note the following handy fact (which is what motivates the regularity assumptions):

Lemma 3.3. *Under Regs 1–3, for continuous distributions*

$$\frac{\partial}{\partial \theta} \int_{\mathcal{A}} f(x, \theta)\, \mathrm{d}x = \int_{\mathcal{A}} \frac{\partial}{\partial \theta} f(x, \theta)\, \mathrm{d}x$$

and for discrete distributions

$$\frac{\partial}{\partial \theta} \sum_{x \in \mathcal{A}} f(x, \theta) = \sum_{x \in \mathcal{A}} \frac{\partial}{\partial \theta} f(x, \theta).$$

Proof. By the Leibniz integral rule. □

This allows us to see the following:

Theorem 3.4. *Under Regs 1–3,*

$$\mathbb{E}_\theta\, S(\theta, X) = 0 \ \forall \theta \in \Theta.$$

Proof. In the continuous case,

$$\mathbb{E}_\theta[S(\theta, X)] = \int_{\mathcal{A}} \ell'(\theta, x) f(x, \theta)\, \mathrm{d}x = \int_{\mathcal{A}} \frac{\frac{\partial}{\partial \theta} f(x, \theta)}{f(x, \theta)} f(x, \theta)\, \mathrm{d}x = \frac{\partial}{\partial \theta} \int_{\mathcal{A}} f(x, \theta)\, \mathrm{d}x = \frac{\partial}{\partial \theta} 1 = 0.$$

The discrete case is similar. □

Definition 3.5. When Regs 1–3 are satisfied, we define the ***Fisher information*** to be

$$I_X(\theta) = \mathrm{Var}_\theta[S(\theta, X)] = \mathbb{E}_\theta[(\ell'(\theta, X)^2].$$

Let us introduce one more regularity assumption now:

Reg 4. *The log-likelihood ℓ is twice-differentiable for all $x \in \mathcal{A}, \theta \in \Theta$, and*

$$\frac{\partial^2}{\partial \theta^2} \int_{\mathcal{A}} f(x, \theta)\, \mathrm{d}x = \int_{\mathcal{A}} \frac{\partial^2}{\partial \theta^2} f(x, \theta)\, \mathrm{d}x \qquad \textit{(for continuous distributions)}$$

or

$$\frac{\partial^2}{\partial \theta^2} \sum_{x \in \mathcal{A}} f(x, \theta)\, \mathrm{d}x = \sum_{x \in \mathcal{A}} \frac{\partial^2}{\partial \theta^2} f(x, \theta)\, \mathrm{d}x \qquad \textit{(for discrete distributions)}$$

for all $\theta \in \Theta$.

Definition 3.6 (Observed information). When Regs 1–4 are satisfied, for an observation $X = x$, we define the ***observed information*** to be

$$J(\theta; x) = -\ell''(\theta, x).$$

This allows us to derive an alternative form for the Fisher information which will be much more commonly used:

Theorem 3.7. *Under Regs 1–4,*

$$I_X(\theta) = -\,\mathbb{E}_\theta[\ell''(\theta, X)] = \mathbb{E}_\theta[J(\theta; X)].$$

Proof. In the continuous case,

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

By Reg 4,

$$\mathbb{E}_\theta\left[\left(\tfrac{\partial^2}{\partial\theta^2}f\right)/f\right] = \int_{\mathcal{A}}\left(\tfrac{\partial^2}{\partial\theta^2}f\right)/f\cdot f\,\mathrm{d}x = \int_{\mathcal{A}}\frac{\partial^2}{\partial\theta^2}f\,\mathrm{d}x = \frac{\partial^2}{\partial\theta^2}\int_{\mathcal{A}} f\,\mathrm{d}x = 0,$$

and thus

$$-\,\mathbb{E}_\theta[\ell''(\theta, X)] = \mathbb{E}_\theta\left[\left(\tfrac{\partial}{\partial\theta}f(X,\theta)/f\right)^2\right] = \mathbb{E}_\theta[(\ell'(\theta, X))^2].$$

The discrete case is similar. □

Proposition 3.8 (Properties of the Fisher information).

1. ***(Information grows with sample size.)*** *If X and Y are independent random variables, then*
$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta).$$
In particular, if $Z = (X_1, \ldots, X_n)$ where the X_i are i.i.d. copies of X, then $I_Z(\theta) = nI_X(\theta)$.
2. ***(Reparametrisation.)*** *If $\theta = h(\xi)$ where h is differentiable, then the Fisher information of X about ξ is*
$$I^*_X(\xi) = I_X(h(\xi))[h'(\xi)]^2.$$

Proof. (for the second statement) The log-likelihood w.r.t. $\mathbb{P}_{h(\xi)}$ is $\ell^*(\xi) = \ln p(x; h(\xi))$ thus the score function is

$$S^*(\xi; x) = \frac{\partial}{\partial\xi}\ln p(x; h(\xi)) = \frac{\partial}{\partial\theta}\ln p(x;\theta)\big|_{\theta=h(\xi)}h'(\xi)$$

and so

$$\mathrm{Var}_\xi S^*(\xi, X) = \mathrm{Var}_\xi\big(S(h(\xi), X)h'(\xi)\big) = I_X(h(\xi))[h'(\xi)]^2.$$ □

Example. Consider $X \sim N(0, \sigma^2)$. The parameter of interest is σ. Let us first compute $I_X(\sigma^2)$. Let us write $\theta = \sigma^2$. Then

$$\ell(\theta; x) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\theta) - \frac{1}{2\theta}x^2.$$

Thus the score function is $S(\theta; X) = \ell'(\theta; X) = -\frac{1}{2\theta} + \frac{1}{2\theta^2}X^2$. Since $\mathrm{Var}_\theta(X^2) = \mathbb{E}_\theta(X^4) - \theta^2 = 2\theta^2$ we get

$$I_X(\theta) = \frac{1}{4\theta^4}\mathrm{Var}_\theta(X^2) = \frac{1}{2\theta^2}.$$

Defining $h(\sigma) = \sigma^2$ we have $h'(\sigma) = 2\sigma$ and by the above Theorem

$$I^*_X(\sigma) = I_X(\sigma^2)[h'[\sigma)]^2 = \frac{1}{2\sigma^4}\cdot 4\sigma^2 = \frac{2}{\sigma^2}.$$

The direct way to compute $I_X^*(\sigma)$ is

$$\ell^*(\sigma; x) = -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2}x^2$$

so that

$$S^*(\sigma; x) = -\frac{1}{\sigma} + \frac{1}{\sigma^3}x^2$$

and

$$I_X^*(\sigma) = \mathrm{Var}_\sigma\left(\frac{1}{\sigma^3}X^2\right) = \frac{1}{\sigma^6}2\sigma^4 = \frac{2}{\sigma^2}.$$

3.2 The multivariate case

Let's extend this all to the multivariate case now — i.e. the case where $\theta \in \mathbb{R}^k$. Reg 1 remains unaltered but we have to adapt the other regularity assumptions:

Reg 2′. *$\Theta \subseteq \mathbb{R}^k$ is an open set.*

Reg 3′. *For all $x \in \mathcal{A}$ and for all $\theta \in \Theta$, the partial derivatives of $L(\theta, x)$ exist and are finite.*

Reg 4′. *All second order partial derivatives of log-likelihood ℓ exist and they can all be commuted with summation/integration over $\mathcal{A}$.*

We can now generalise our definitions:

Definition 3.9. When Regs 1, 2′, 3′ are satisfied, we define the ***score function*** to be

$$S(\theta, x) = \nabla_\theta \ell(\theta, x) = \left(\tfrac{\partial}{\partial\theta_1}\ell(\theta, x), \ldots, \tfrac{\partial}{\partial\theta_k}\ell(\theta, x)\right)^\mathsf{T}.$$

Definition 3.10. When Regs 1, 2′, 3′ are satisfied, we define the ***Fisher information*** matrix to be

$$I_X(\theta) = \mathrm{Cov}_\theta(S(\theta, X)),$$

so that

$$I_X(\theta)_{jr} = \mathbb{E}_\theta\left[\tfrac{\partial}{\partial\theta_j}\ell(\theta, X)\tfrac{\partial}{\partial\theta_r}\ell(\theta, X)\right].$$

Note the last line above used that the multi-dimensional score function also has zero expectation, which can be shown much like in the one-dimensional case.

Theorem 3.11. *Supposing Regs 1, 2′, 3′, 4′ hold, define the* **observed Fisher information** *matrix J by $J(\theta, x)_{jr} = -\dfrac{\partial^2\ell(\theta, x)}{\partial\theta_j\partial\theta_r}$ for $j, r = 1, \ldots, k$. Then*

$$I_X(\theta) = \mathbb{E}_\theta[J(\theta, X)].$$

Proof. Exercise (a generalisation of the one-dimensional case). □

Chapter 4

Point estimation

See Garthwaite, Joliffe and Jones p40, Liero and Zwanzig p71.

Definition 4.1. For any function $g : \Theta \to \Gamma$ (for some set Γ), an ***estimator*** of $\gamma = g(\theta)$ is a function $T : \mathcal{X} \to \Gamma$.

The value $T(X)$ is called the ***estimate*** of $g(\theta)$.

Definition 4.2. The ***bias*** of an estimator T for $\gamma = g(\theta)$ is

$$\text{bias}(T, \theta) = \mathbb{E}_\theta[T] - g(\theta).$$

T is called ***unbiased*** for $g(\theta)$ if $\mathbb{E}_\theta[T] = g(\theta)\ \forall \theta \in \Theta$.

Example. Suppose $X = (X_1, \ldots, X_n)$ is a sample of i.i.d. $\mathcal{N}(\mu, \sigma^2)$ random variables. Then $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ is an unbiased estimator for μ, and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \widehat{\mu})^2$ is an unbiased estimator for σ^2

(Exercise: prove this.)

4.1 The method of moments

A very simple approach for estimating functions of moments of a random variable is to replace all of the moments by their empirical values.

Formally, suppose $(X_1, \ldots, X_n)$ is a sample of i.i.d. P_θ-distributed random variables, where $\theta \in \Theta$ is the parameter. In general if $X \sim P_\theta$, then the moments $m_r = \mathbb{E}_\theta[X^r]$ for $r = 1, 2, \ldots$ depend on θ.

Assume there exists a function h such that $\gamma = h(m_1, \ldots, m_r)$.

Definition 4.3. For each $k = 1, \ldots, r$ let $\widehat{m}_k = \frac{1}{n}\sum_{i=1}^n X_i^k$. Then the ***moment estimator*** for γ is defined as

$$\widehat{\gamma}_{\mathsf{MME}} = h(\widehat{m}_1, \ldots, \widehat{m}_r).$$

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Poisson with parameter $\lambda > 0$. Since $m_1 = \mathbb{E}[X_1] = \lambda$, we

can use the sample mean $\widehat{m}_1 = \frac{1}{n}\sum_{i=1}^n X_i$, so that

$$\widehat{\lambda}_{\mathsf{MME}} = \widehat{m}_1 = \frac{1}{n}\sum_{i=1}^n X_i.$$

On the other hand, $\mathrm{Var}(X_i) = \lambda$ as well, so writing $\mathrm{Var}(X_i) = m_2 - m_1^2$ we can also use the estimator

$$\widehat{\lambda}_{\mathsf{MME}} = \widehat{m}_2 - \widehat{m}_1^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2.$$

Which estimator is "better"?

4.2 Maximum likelihood estimators

Definition 4.4. An estimator T is called a ***maximum likelihood estimator (MLE)*** for θ if

$$L(T(x), x) = \max_{\theta\in\Theta} L(\theta, x) \qquad \forall x \in \mathcal{X},$$

and is denoted by $\widehat{\theta}_{\mathsf{MLE}}$.

Theorem 4.5 (The Invariance Property). *If $\gamma = g(\theta)$ and g is bijective, then $\widehat{\theta}$ is a MLE for θ if and only if $\widehat{\gamma} = g(\widehat{\theta})$ is a MLE for γ.*

Proof. Part A statistics. □

In the case above, if g is *not* bijective, we define $\widehat{\gamma}_{\mathsf{MLE}} = g(\widehat{\theta}_{\mathsf{MLE}})$.

4.2.1 Finding the MLE

1. In many standard cases the function L is differentiable. In these cases, one way to maximise L is to differentiate it and set the (partial) derivatives with zero. If L achieves its maximum at an interior point of Θ then θ_{MLE} must be a solution of

$$\frac{\partial L(\theta; x)}{\partial \theta_j} = 0, j = 1, \ldots, k.$$

   Note that it is sometime easier to work with the log-likelihood ℓ instead of L and that maximise L is equivalent to maximising ℓ.

2. One then has to show that the solution to this system of equation is a local maximum (by considering the Hessian matrix and if the solution is not unique to show that it is a global maximum.

3. Notice that in some cases the maximum may occur at a boundary point, in which case the partial derivatives may not vanish.

Lemma 4.6 (MLE for exponential families). *Consider a k-parameter exponential family $\mathcal{P} = \{f(\cdot;\eta) : \eta\}$ in natural parameterisation*

$$L(\eta; x) = \exp\left\{\sum_{j=1}^k \eta_j T_j(x) - B(\eta)\right\} h(x). \tag{4.1}$$

Then any MLE of ϕ satisfies,

$$T_j(x) = \frac{\partial B}{\partial \eta_i}(\widehat{\eta}_{MLE}), \quad j = 1, \dots, k. \tag{4.2}$$

If $\mathcal{P}$ is strictly k-parameter and (4.2) admits a solution, then it is unique.

Proof. The first statement follows trivially from Theorem 1.15. For the second statement, also follows by Theorem 1.15, note that

$$\mathrm{Cov}_\eta(T) = \left(\frac{\partial^2}{\partial \eta_i \eta_j} B(\eta) \right)_{ij}.$$

By Propositions 1.9, 1.10 $\mathrm{Cov}_\eta(T)$ is strictly positive definite. Therefore the negative log-likelihood is strictly concave and therefore any local maximum is also a global maximum. □

4.3 Variance and mean squared error

Definition 4.7. The ***mean squared error (MSE)*** of an estimator T for $g(\theta)$ is defined as

$$\mathrm{MSE}_\theta(T) = \mathbb{E}_\theta[(T - g(\theta))^2].$$

(This is also often called the ***quadratic loss function***.)

Proposition 4.8. *In general, for an estimator T for $g(\theta)$,*

$$\mathrm{MSE}_\theta(T) = \mathrm{Var}_\theta(T) + \underbrace{(\mathbb{E}_\theta[T] - g(\theta))^2}_{\text{bias}^2}.$$

In particular, if T is unbiased, $\mathrm{MSE}_\theta(T) = \mathrm{Var}_\theta(T)$.

Proof. Exercise. □

Example. Let $X = (X_1, \dots, X_n)$ be a sample of i.i.d. $\mathcal{U}(0, \theta)$ random variables. Then $\widehat{\theta}_{\mathsf{MLE}} = X_{\max} = \max\{X_i : i = 1, \dots, n\}$.

It's easy to check that $\mathbb{E}_\theta(X_{\max}) = \frac{n}{n+1}\theta$ and $\mathrm{Var}_\theta(X_{\max}) = \frac{n}{(n+1)^2(n+2)}\theta^2$, so that

$$\mathrm{MSE}_\theta(X_{\max}) = \frac{2\theta^2}{(n+1)(n+2)}.$$

However, the estimator $\widehat{\theta} = \frac{n+1}{n} X_{\max}$ is unbiased, and indeed

$$\mathrm{MSE}_\theta(\widehat{\theta}) = \frac{\theta^2}{n(n+2)} < \mathrm{MSE}_\theta(\widehat{\theta}_{\mathsf{MLE}}).$$

Typically although we write $\widehat{\theta} = T(X)$, with $X \sim f(\cdot; \theta)$ for an estimator, the notation is overloaded when considering $X_1, \dots, X_n \overset{\text{iid}}{\sim} f(\cdot; \theta)$ with n varying. In this case we understand the estimator and the statistic as a family of estimators $\{\widehat{\theta}_n\}_n$ with $\widehat{\theta}_n := T_n(X_1, \dots, X_n)$ and $T_n : \mathcal{X} \mapsto \mathbb{R}^k$ for some k.

For most statistics, it is quite easy to see the dependence on n; e.g. for the sample mean $T_n(\boldsymbol{x}) = (x_1 + \cdots + x_n)/n$, or $T_n(\boldsymbol{x}) = \min_{i \le n} x_i$ etc.

As the sample size $n \to \infty$ we expect a good family of estimators $\widehat{\theta}_n = T_n(X_{1:n})$ to eventually recover

the value of the parameter we are estimating exactly, that is when $X_1, \dots, X_n \overset{\text{iid}}{\sim} f(\cdot;\theta)$ we expect $T_n(X_{1:n}) \to \theta$ in some sense, e.g. in probability, almost surely or in MSE. This is the concept of ***consistency***.

Definition 4.9 (Consistency). Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} f(\cdot;\theta)$ and consider the family of estimators $\widehat{\theta}_n = T_n(X_{1:n})$. We say that $T_n(X_{1:n})$ is

- ***consistent in probability*** if for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}_\theta \left[\|T_n(X_{1:n}) - \theta\| > \varepsilon \right] \to 0;$$

- ***consistent in MSE*** if

$$\lim_{n\to\infty} \mathsf{MSE}_\theta(T_n;\theta) \to 0;$$

- ***strongly consistent*** if

$$\lim_{n\to\infty} T_n(X_{1:n}) \to \theta), \quad \mathbb{P}_\theta\text{-almost surely.}$$

Any mode of convergence can be used to define a corresponding notion of consistency.

Chapter 5

MVUEs and the Cramer-Rao Lower Bound

Now that we have developed a few techniques for estimating a parameter, we need to evaluate how well various estimators actually work.

Suppose $X = (X_1, \ldots, X_n)$ is a random sample from the distribution P_θ. What is a 'good' estimator of θ?

A fairly natural pathway would be to try and minimise the MSE:

Definition 5.1. We say T_1 is a ***uniformly better*** estimator than T_2 (or ***better in quadratic mean***) if for all $\theta \in \Theta$,

$$\mathrm{MSE}_\theta(T_1) \leqslant \mathrm{MSE}_\theta(T_2).$$

Remark. If $\widehat{\theta} = \theta_0$, then $\mathrm{MSE}_{\theta_0}(\widehat{\theta}) = 0$. Hence no other estimator can be uniformly better!

5.1 The CRLB in the one-dimensional case

Definition 5.2. $T = T(X_1, \ldots, X_n)$ is the ***minimum variance unbiased estimator (MVUE)*** for θ (resp. for $g(\theta)$) if

- T is unbiased, and
- for all unbiased estimators $\tilde{T}$, $\mathrm{Var}_\theta(\tilde{T}) \geqslant \mathrm{Var}_\theta(T)\ \forall \theta \in \Theta$.

The estimator T is furthermore said to be ***regular*** if

$$\int_{\mathcal{A}} T(x) \frac{\partial}{\partial \theta} L(\theta; x) dx = \frac{\partial}{\partial \theta} \int_{\mathcal{A}} T(x) L(\theta; x) dx = \frac{\partial}{\partial \theta} \mathbb{E}_\theta[T(X)].$$

Theorem 5.3 (Cramer-Rao Lower Bound (CRLB) in 1 dimension). *Suppose Regs 2–4 hold and that $0 < I_X(\theta) < \infty$. Let $\gamma = g(\theta)$ where g is a continuously differentiable real-valued function with $g' \neq 0$.*

Let T be a **regular unbiased** *estimator of γ. Then*

$$\mathrm{Var}_\theta(T) \geqslant \frac{|g'(\theta)|^2}{I_X(\theta)},\ \forall \theta \in \Theta$$

with equality if and only if

$$T(x) - g(\theta) = \frac{g'(\theta)S(\theta, x)}{I_X(\theta)} \quad \forall x \in \mathcal{A}\ \forall \theta \in \Theta.$$

Remark. If T attains the CRLB,

$$\mathrm{Var}_\theta(T) = \frac{|g'(\theta)|^2}{I_X(\theta)},$$

then it is clearly a MVUE. There is no guarantee that there exists an estimator which attains the bound.

Remark. In the case $g(\theta) = \theta$ the CRLB is

$$\mathrm{Var}_\theta(T) \geqslant \frac{1}{I_X(\theta)}$$

and T attains the CRLB if and only if $S(\theta, x) = I_X(\theta)(T(x) - \theta)\ \forall x \in \mathcal{A}\ \forall \theta \in \Theta$, or equivalently $T(x) = \theta + \frac{S(\theta,x)}{I_X(\theta)}$.

Proof of theorem. Note that

$$\mathrm{Cov}_\theta(T, S(\theta, X)) = \mathbb{E}_\theta[TS(\theta, X)] \qquad \text{since } \mathbb{E}_\theta(S(\theta, X)) = 0 \tag{5.1}$$

$$= \int_{\mathcal{X}} T(x) \frac{\partial \log p(x, \theta)}{\partial \theta} p(x, \theta)\, \mathrm{d}x \tag{5.2}$$

$$= \int_{\mathcal{X}} T(x) \frac{\partial p(x, \theta)}{\partial \theta}\, \mathrm{d}x \tag{5.3}$$

$$= \frac{\partial}{\partial \theta} \int_{\mathcal{X}} T(x) p(x, \theta)\, \mathrm{d}x \qquad \text{(note this is where we need } T \text{ regular)} \tag{5.4}$$

$$= \frac{\partial}{\partial \theta} \mathbb{E}_\theta[T] = g'(\theta). \tag{5.5}$$

Now set $c(\theta) \coloneqq g'(\theta)/I_X(\theta)$. Then

$$0 \leqslant \mathrm{Var}_\theta(T - c(\theta)S(\theta, X)) = \mathrm{Var}_\theta T + c^2(\theta)\, \mathrm{Var}_\theta(S(\theta, X)) - 2c(\theta)\, \mathrm{Cov}_\theta(T, S(\theta, X)) \tag{5.6}$$

$$= \mathrm{Var}_\theta T + c^2(\theta) I_X(\theta) - 2c(\theta) g'(\theta) \tag{5.7}$$

$$= \mathrm{Var}_\theta(T) - \frac{|g'(\theta)|^2}{I_X(\theta)} \tag{5.8}$$

which is the CRLB. We have equality if and only if $T - c(\theta)S(\theta, X)$ is almost surely constant, and in that case it must be equal to its expectation $g(\theta)$:

$$T(x) - c(\theta)S(\theta, x) = g(\theta) \iff T(x) - g(\theta) = \frac{S(\theta, x)g'(\theta)}{I_X(\theta)}.$$

□

Is it common that there exists an estimator attaining the CRLB?

Corollary 5.4. *Suppose that $X \sim f(\cdot; \theta)$ for $\theta \in \Theta$. Under regularity conditions, if some estimator $\widehat{\gamma}$ of $\gamma = g(\theta)$ attains the CRLB, it follows that $\{f(\cdot; \theta) : \theta\}$ is in some exponential family.*

Proof. If $\widehat{\gamma} = T(x)$ attains the CRLB then we know that

$$T(x) = g(\theta) + \frac{g'(\theta)}{I_X(\theta)} \frac{\partial}{\partial \theta} \log f(x; \theta).$$

Rearranging we have

$$\frac{\partial}{\partial\theta}\log f(x;\theta) = (T(x)-\theta)\frac{g'(\theta)}{I_X(\theta)}.$$

By the fundamental theorem of calculus we have that

$$\log f(\theta;x) = C(x) + \eta(\theta)T(x) - B(\theta),$$

for some functions C, η, B and the conclusion follows. □

Corollary 5.5. *Suppose that $\mathbb{E}_\theta[T(X)] = \theta + b(\theta)$ (so that $b(\theta)$ is the bias of T) and that T is regular. Then*

$$Var_\theta(T(X)) \geqslant \frac{|1+b'(\theta)|^2}{I_X(\theta)}$$

Example. Suppose $X \sim \text{Bin}(n,\theta)$, where n is known. Our parameter of interest will be $\gamma = \theta(1-\theta)$ (so $g'(\theta) = 1-2\theta$). Hence

$$\ell(\theta,x) = \log\binom{n}{x} + (n-x)\log(1-\theta) + x\log\theta,$$

and therefore

$$S(\theta,x) = -\frac{n-x}{1-\theta} + \frac{x}{\theta},$$

so

$$\frac{\partial}{\partial\theta}S(\theta,x) = -\frac{n-x}{(1-\theta)^2} - \frac{x}{\theta^2}.$$

Thus the Fisher information is

$$I_X(\theta) = -\mathbb{E}_\theta\left[\tfrac{\partial}{\partial\theta}S(\theta,X)\right] \tag{5.9}$$

$$= \frac{n-\mathbb{E}_\theta[X]}{(1-\theta)^2} + \frac{\mathbb{E}_\theta[X]}{\theta^2} = \frac{n}{(1-\theta)\theta}. \tag{5.10}$$

Observe that $T(x) = \frac{1}{n-1}x\left(1-\frac{x}{n}\right)$ is unbiased for γ (check this as an exercise) and $\text{Var}_\theta(T) = \frac{\theta}{n} - \frac{\theta^2(5n-7)-4\theta^3(2n-3)+\theta^4(4n-6)}{n(n-1)}$ which is larger than the CRLB of $\frac{(1-2\theta)^2\theta(1-\theta)}{n}$.

5.2 Efficiency

Definition 5.6. The efficiency of an unbiased estimator T of $g(\theta)$ is the ratio of its variance and of the CRLB, that is

$$e(T,\theta) = \frac{[g'(\theta)]^2}{I_X(\theta)\text{Var}_\theta T}.$$

An unbiased estimator which attains the CRLB is called ***efficient***.

The following theorem is valid for exponential families.

Theorem 5.7. *Suppose that the distribution of $X = (X_1,\ldots,X_n)$ belongs to a one-parameter exponential family in ζ and T. Then the sufficient statistic T is an efficient estimator for the parameter $\gamma = g(\theta) = \mathbb{E}_\theta[T]$.*

Proof. The probability function of the sample is given by

$$p(x;\theta) = \exp(T(x)\zeta(\theta) - B(\theta))h(x)$$

and we have for the score function

$$S(\theta; x) = \frac{\partial}{\partial \theta} \ln p(x; \theta) = -B'(\theta) + \zeta'(\theta) T(x)$$

This means that $S(\theta; x)$ is a linear function of $T(x)$, hence

$$\mathrm{Cor}_\theta(S(\theta; X), T(X))^2 = 1. \tag{5.11}$$

From the proof of the CRLB we know that for an unbiased estimator T of $g(\theta)$:

$$\mathrm{Cov}_\theta(S(\theta; X), T(X)) = g'(\theta).$$

Thus by (5.11)

$$\frac{[g'(\theta)]^2}{\mathrm{Var}_\theta(S(\theta; X)\mathrm{Var}_\theta(T(X))} = 1$$

but since $\mathrm{Var}_\theta(S(\theta; X) = I_X(\theta)$ we have that

$$\mathrm{Var}_\theta(T(X)) = \frac{[g'(\theta)]^2}{I_X(\theta)}$$

which is the Cramer-Rao bound. □

5.3 The multivariate case

We turn now to the multivariate case. Suppose that $\gamma = g(\theta) \in \mathbb{R}^m$.

We will compare matrices using the Loewner order:

Definition 5.8. Let T, T^* be two unbiased estimators for γ. We say that T^* has a ***smaller*** covariance matrix than T at $\theta \in \Theta$ if

$$u^t(\mathrm{Cov}_\theta T^* - \mathrm{Cov}_\theta T)u \leqslant 0 \quad \forall u \in \mathbb{R}^m,$$

and we write $\mathrm{Cov}_\theta T^* \preceq \mathrm{Cov}_\theta T$.

Theorem 5.9 (Cramer-Rao Lower Bound in m dimensions). *Suppose Regs 1, 2′, 3′, 4′ hold and that $I_X(\theta)$ is not singular. Then the CRLB is*

$$\mathrm{Cov}_\theta T \succeq (D_\theta g)(\theta) I_X(\theta)^{-1} (D_\theta g)(\theta)^t \qquad \forall \theta \in \Theta,$$

where $D_\theta g$ is the Jacobian matrix, so $(D_\theta g)(\theta)_{ij} = \dfrac{\partial g_i(\theta)}{\partial \theta_j}$.

Proof. The idea is similar to the one-dimensional case; start by computing

$$\mathrm{Cov}_\theta \left[T(X) - D_\theta g(\theta) I_X(\theta)^{-1} S(X, \theta) \right] \succeq 0.$$

The rest is an excercise in matrix multiplications. □

Example. Let $X = (X_1, \ldots, X_n)$ be a random sample of $\mathcal{N}(\mu, \sigma^2)$ random variables, where our parameter of interest is $\theta = (\mu, \sigma^2)$. Recall from Part A Statistics that

$$I_X(\theta) = \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/2\sigma^4 \end{pmatrix}.$$

The estimators $\overline{X}$ and S^2 are independent, with $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$ and $\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}$. We can see that the CRLB is not attained.

5.4 MLEs and MVUEs

Note too the following, which shows that MLEs line up with MVUEs when the CRLB is attained:

Theorem 5.10. *Under Regs 1, 2′, 3′, 4′, if $\widehat{\theta}_{MLE}$ is the MLE for θ and if there exists $\tilde{\theta}$ which is unbiased and attains the CRLB, then $\tilde{\theta} = \widehat{\theta}_{MLE}$ almost surely.*

Proof. For simplicity we prove this statement in the case of a real-valued parameter. Suppose that $\tilde{\theta}$ is an efficient estimator of θ. Then we know that

$$\tilde{\theta}(x) - \theta = S(\theta; x)/I_X(\theta)$$

for all $\theta \in \Theta$ and $x \in \mathcal{A}$ So it holds in particular for $\theta = \widehat{\theta}_{\mathrm{MLE}}(x)$

$$\tilde{\theta}(x) - \widehat{\theta}_{\mathrm{MLE}}(x) = S(\widehat{\theta}_{\mathrm{MLE}}(x); x)/I_X(\widehat{\theta}_{\mathrm{MLE}}(x)).$$

Since $\widehat{\theta}_{\mathrm{MLE}}(x)$ maximizes $\ell(\cdot; x)$ we have $S(\widehat{\theta}_{\mathrm{MLE}}(x); x) = 0$ and thus

$$\tilde{\theta}(x) - \widehat{\theta}_{\mathrm{MLE}}(x) = 0.$$

□

Remark. Note that $\widehat{\theta}$ may be unbiased and MVUE without attaining the CRLB, in which case the conclusion above may be false.

Chapter 6

The Rao-Blackwell and Lehmann-Scheffé theorems

Of course, even when the CRLB is not achievable, we still want to be able to find a MVUE.

Theorem 6.1 (Rao-Blackwell Theorem). *Let $X \sim P_\theta$ and let T be a sufficient statistic. Let $\widehat{\gamma}$ be an unbiased estimator for $\gamma = g(\theta) \in \mathbb{R}^k$.*

Define $\widehat{\gamma}_T = \mathbb{E}_\theta[\widehat{\gamma} \mid T]$. Then:

1. *$\widehat{\gamma}_T$ is a function of T alone and does not depend on θ,*
2. *$\mathbb{E}_\theta[\widehat{\gamma}_T] = \gamma \ \forall \theta \in \Theta$,*
3. *$\mathrm{Cov}_\theta(\widehat{\gamma}_T) \preceq \mathrm{Cov}_\theta(\widehat{\gamma})$ (which reduces to $\mathrm{Var}_\theta(\widehat{\gamma}_T) \leqslant \mathrm{Var}_\theta(\widehat{\gamma})$, in the case $k = 1$).*

If $\mathrm{tr}(\mathrm{Cov}_\theta(\widehat{\gamma})) < \infty$ then $\mathrm{Cov}_\theta(\widehat{\gamma}_T) = \mathrm{Cov}_\theta(\widehat{\gamma})$ if and only if $\widehat{\gamma}_T = \widehat{\gamma}$ almost surely.

Intuitively, this says that 'any unbiased estimator can always be improved by a sufficient statistic. The last statement says that if some unbiased estimator ***cannot*** be improved by conditioning on a sufficient statistics, then the estimator is essentially a function of the statistic.

Remark. Notice that the Rao-Blackwell theorem implies that if an unbiased estimator is a function of a minimal sufficient statistic, then it is MVUE. Why is that? Notice that a minimal sufficient statistic T can be written as a function of any other sufficient statistic T'; therefore if $\widehat{\gamma}$ is an unbiased estimator, we have that $\widehat{\gamma}_T$ is a function of T and therefore a function of T'. Therefore $\mathbb{E}[\widehat{\gamma}_T | T'] = \widehat{\gamma}_T$ and we are in the equality case of the Rao-Blackwell theorem.

Proof of theorem. We prove the three parts in order:

1. Since T is sufficient, $f(x \mid \theta, T)$ is independent of θ, so

$$\widehat{\gamma}_T = \mathbb{E}_\theta[\widehat{\gamma} \mid T = t] = \int_{\mathcal{X}} \widehat{\gamma}(x) f(x \mid t, \theta) \, \mathrm{d}x = \int_{\mathcal{X}} \widehat{\gamma}(x) f(x \mid t) \, \mathrm{d}x$$

   which does not depend on θ.

2. By the unbiasedness of $\widehat{\gamma}$ and the tower property of expectations,

$$\mathbb{E}_\theta[\widehat{\gamma}_T] = \mathbb{E}_\theta[\mathbb{E}_\theta[\widehat{\gamma} \mid T]] = \mathbb{E}_\theta[\widehat{\gamma}] = \gamma.$$

3. For $k = 1$, the result is fairly straightforward:

$$\begin{aligned}
&(6.1)\\
&\mathrm{Var}_\theta(\widehat{\gamma}) = \mathbb{E}_\theta[(\widehat{\gamma} - \gamma)^2] = \mathbb{E}_\theta[(\widehat{\gamma} - \widehat{\gamma}_T + \widehat{\gamma}_T - \gamma)^2] \\
&(6.2) \qquad = \mathbb{E}_\theta[\mathbb{E}_\theta[(\widehat{\gamma} - \widehat{\gamma}_T + \widehat{\gamma}_T - \gamma)^2 \mid T]] \\
&(6.3) \qquad = \mathbb{E}_\theta[\mathbb{E}_\theta[(\widehat{\gamma} - \widehat{\gamma}_T)^2 \mid T] - 2\,\mathbb{E}_\theta[(\widehat{\gamma} - \widehat{\gamma}_T)(\widehat{\gamma}_T - \gamma) \mid T] + \mathbb{E}_\theta[(\widehat{\gamma}_T - \gamma)^2 \mid T]] \\
&(6.4) \qquad = \mathbb{E}_\theta[\mathbb{E}_\theta[(\widehat{\gamma} - \widehat{\gamma}_T)^2 \mid T]] - 0 + \mathbb{E}_\theta[\mathbb{E}_\theta[(\widehat{\gamma}_T - \gamma)^2 \mid T]] \\
&(6.5) \qquad = \mathbb{E}_\theta[\mathrm{Var}_\theta(\widehat{\gamma} \mid T)] + \mathrm{Var}_\theta(\widehat{\gamma}_T) \\
&(6.6) \qquad \geqslant \mathrm{Var}_\theta(\widehat{\gamma}_T).
\end{aligned}$$

   For $k > 1$, we can instead do:

$$\begin{aligned}
&(6.7) \quad \mathrm{Cov}_\theta[\widehat{\gamma}] = \mathbb{E}_\theta[(\widehat{\gamma} - \gamma)(\widehat{\gamma} - \gamma)^t] \\
&(6.8) \qquad = \mathbb{E}_\theta[(\widehat{\gamma} - \widehat{\gamma}_T)(\widehat{\gamma} - \widehat{\gamma}_T)^t] + \mathbb{E}_\theta[(\widehat{\gamma}_T - \gamma)(\widehat{\gamma}_T - \gamma)^t] - 2\,\mathbb{E}_\theta[(\widehat{\gamma} - \widehat{\gamma}_T)(\widehat{\gamma}_T - \gamma)^t] \\
&(6.9) \qquad = \mathbb{E}_\theta[(\widehat{\gamma} - \widehat{\gamma}_T)(\widehat{\gamma} - \widehat{\gamma}_T)^t] + \mathrm{Cov}_\theta(\widehat{\gamma}_T) + 2\,\mathbb{E}_\theta[(\widehat{\gamma} - \widehat{\gamma}_T)(\widehat{\gamma}_T - \gamma)^t].
\end{aligned}$$

   The first term here is clearly nonnegative, and it isn't too hard to see that the third term is equal to zero. The result follows.

The proof of the equality case is left as an exercise. □

Example. Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{Ber}(\theta)$ random variables. Note that $\widehat{\theta} = X_1$ is unbiased for θ, and that $T = \sum_{i=1}^n X_i$ is sufficient for θ.

In this case,

$$\begin{aligned}
&(6.10) \qquad \widehat{\theta}_T = \mathbb{E}_\theta[X_1 \mid T = t] = \mathbb{P}_\theta(X_1 = 1 \mid T = t) = \frac{\mathbb{P}_\theta(X_1 = 1, T = t)}{\mathbb{P}_\theta(T = t)} \\
&(6.11) \qquad = \frac{\mathbb{P}_\theta\left(X_1 = 1, \sum_{i=2}^n X_i = t - 1\right)}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&(6.12) \qquad = \frac{\theta\binom{n-1}{t-1}\theta^{t-1}(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{t}{n}
\end{aligned}$$

so $\widehat{\theta}_T = T/n$.

Definition 6.2. A statistical model $\{P_\theta : \theta \in \Theta\}$ is called ***complete*** if for any $h : \mathcal{X} \to \mathbb{R}$,

$$\mathbb{E}_\theta[h(X)] = 0 \ \forall \theta \in \Theta \implies \mathbb{P}_\theta(h(X) = 0) = 1 \ \ \forall \theta \in \Theta.$$

A statistic T is called ***complete*** if the model $\{P_\theta^T : \theta \in \Theta\}$ is complete, i.e.

$$\mathbb{E}_\theta[h(T)] = 0 \ \forall \theta \in \Theta \implies \mathbb{P}_\theta(h(T) = 0) = 1 \ \forall \theta \in \Theta.$$

Examples.

1. Suppose the statistical model consists only of the two distributions $\mathcal{N}(1, 2)$ and $\mathcal{N}(0, 1)$. This model is *not* complete: take $h(x) = (x-1)^2 - 2$. For both distributions, $\mathbb{E}[h(x)] = 0$, but $h(x) \neq 0 \ \forall x \neq \sqrt{2} + 1, 1 - \sqrt{2}$.
2. The statistical model $\{\mathcal{U}(0, \theta), \theta \in \mathbb{R}_+\}$ *is* complete. Indeed, suppose $0 = \mathbb{E}_\theta[h(X)] =$

$\int_0^\theta \frac{1}{\theta} h(x)\,\mathrm{d}x$ for all $\theta > 0$. Then

$$\frac{\partial}{\partial\theta}\int_0^\theta h(x)\,\mathrm{d}x = 0 \ \forall\theta > 0.$$

But $\frac{\partial}{\partial\theta}\int_0^\theta h(x)\,\mathrm{d}x = h(\theta)$ almost everywhere (by Lebesgue differentiation Theorem), we conclude that $h(x) = 0$ almost everywhere. Since the variable $X \sim \mathbb{P}_\theta$ has a density we conclude that $\mathbb{P}_\theta(h(X) = 0) = 1$.

3. If $X_1, \ldots, X_n$ are i.i.d. $\mathcal{U}(0,\theta)$ then $X_{\max}$ is a complete statistic. Indeed, the density of $X_{\max}$ is

$$f_\theta(t) = \frac{nt^{n-1}}{\theta^n}\mathbb{1}_{t\in[0,\theta]}.$$

Then if $0 = \mathbb{E}_\theta[h(X_{\max})] = \int_{-\infty}^{\infty} h(t)f_\theta(t)\,\mathrm{d}t = \frac{n}{\theta^n}\int_0^\theta h(t)t^{n-1}\,\mathrm{d}t$ for all $\theta \in \Theta$, and thus $0 = \int_0^\theta h(t)t^{n-1}\,\mathrm{d}t$ for all $\theta \in \Theta$. Thus, again using Lebesgue differentiation Theorem (which says that as a function of θ this is differentiable and equal to $g(\theta)$ almost everywhere in θ), we have that $g = 0$ almost everywhere and we conclude as above.

The following fact will be useful:

Lemma 6.3. *Let*

$$p(x;\theta) = \exp\big[\sum_{i=1}^k \eta_i(\theta)T_i(x) - B(\theta)\big]h(x), \theta \in \Theta$$

be a strictly k-parameter exponential family. The joint distribution of the natural observation vector $T(X) = (T_1(X), \ldots, T_k(X))$, with $X \sim p(\cdot;\theta)$, belongs to a strictly k parameter exponential family with natural parameters $\eta_1(\theta), \ldots, \eta_k(\theta)$.

Proof. We only deal with the discrete case. Fix some vector $y \in \mathbb{R}^k$ and let $\mathcal{T}_y = \{x : T(x) = y\}$. Then

$$\mathbb{P}_\theta(T = y) = \sum_{x\in\mathcal{T}_y} \mathbb{P}_\theta(X = x) \tag{6.13}$$

$$= \sum_{x\in\mathcal{T}_y} h(x)\exp\big[\sum_{i=1}^k \eta_i(\theta)y_i - B(\theta)\big] \tag{6.14}$$

$$= \exp\big[\sum_{i=1}^k \eta_i(\theta)y_i - B(\theta)\big]\sum_{x\in\mathcal{T}_y} h(x) \tag{6.15}$$

$$= \exp\big[\sum_{i=1}^k \eta_i(\theta)y_i - B(\theta)\big]h_0(y) \tag{6.16}$$

□

Theorem 6.4 (Completeness for exponential families). *If $\mathcal{P}$ is a full-rank strictly k-parameter exponential family then the natural observation $T(x) = (T_1(x), \ldots, T_k(x))$ is sufficient and complete.*

(Sketch). The idea is to use uniquess of Laplace transforms, or something closer to home, uniqueness of the Moment Generating Function.

Suppose that the model is

$$p(x,\theta) = h(x)\exp\left\{\sum_{i=1}^{k}\eta_i(\theta)T_i(x) - B(\theta)\right\}.$$

Hence assuming $\mathbb{E}_\theta[\Phi(T)] = 0$ for all θ means that

$$\int_{\mathbb{R}^k}\phi(t)h_0(t)\exp\left\{\sum_{i=1}^{k}\eta_i(\theta)t_i\right\}\mathrm{d}t = 0, \quad \forall\theta\in\Theta. \tag{6.17}$$

Since the model is full rank we know that $\eta(\Theta)$ contains a k-dimensional interval. Notice that by redefining h_0 we can shift the parameter vector $\eta(\theta)$ by any fixed vector without changing the exponential family; thus we can assume w.l.o.g. that for some $a > 0$ we have $[-a,a]^k \subset \eta(\Theta)$. In particular $0 \in \eta(\Theta)$ and therefore $\int\phi(t)h_0(t)\,\mathrm{d}t = 0$. We decompose $\phi = \phi^+ - \phi^-$ into its positive and negative parts, that is $\phi^+ = \max\{\phi,0\}$ and $\phi^- = -\min\{\phi,0\}$. Notice that this decomposes the range of the statistic $\mathcal{T}$ into two disjoint sets, $\mathcal{T}^+ = \{x : \phi(t) \geq 0\}$ and $\mathcal{T}^- = \{t : \phi(t) < 0\}$. Notice that ϕ^+ is supported on $\mathcal{T}^+$ and ϕ^- on $\mathcal{T}^-$. Then we have that

$$\int\phi^+(t)h_0(t)\,\mathrm{d}t = \int\phi^-(t)h_0(t)\,\mathrm{d}t = w.$$

Notice $w < \infty$ since otherwise the integral would not be defined.

Suppose that $w > 0$. Then we can define the probability measures

$$\lambda^+(\mathrm{d}t) = \frac{\phi^+(t)h_0(t)\mathrm{d}t}{w}\mathbb{1}_{\mathcal{T}^+}(t), \qquad \lambda^-(\mathrm{d}t) = \frac{\phi^-(t)h_0(t)\mathrm{d}t}{w}\mathbb{1}_{\mathcal{T}^-}(t).$$

By (6.17) we thus have that the MGFs of λ^+, λ^- are equal for all $\eta \in [-a,a]^k$, i.e. that

$$\mathbb{E}_{T\sim\lambda^+}\left\{\exp\left[\eta\cdot T\right]\right\} = \mathbb{E}_{T\sim\lambda^-}\left\{\exp\left[\eta\cdot T\right]\right\},$$

for all $\eta \in [-a,a]^k$, which by uniqueness of MGFs implies that $\lambda^+ = \lambda^-$. But this is impossible since $\lambda^+(\mathcal{T}^+) = 1$ whereas $\lambda^-(\mathcal{T}^+) = 0$. Therefore it must be that $w = 0$, in which case we conclude that $\phi^+, \phi^- \equiv 0$ on the support of the exponential family and we are done. □

We may then ask the question: ***does a complete sufficient statistic always exist?*** The answer is no as the following example shows.

Example. Suppose $X \sim \mathcal{U}(\theta,\theta+1)$, $\theta\in\mathbb{R}$. Then $T(X) = X$ is clearly a sufficient statistic. By considering the likelihood ratios, it is not difficult to establish that T is also minimal.

Now suppose that $\tilde{T}$ is a sufficient statistic. By minimality of T we know that there exists a function g such that $x = T(x) = g\circ\tilde{T}(x)$. Consider then the function $t \mapsto \sin\big(2\pi g(t)\big)$. We have

$$\begin{aligned}\mathbb{E}_\theta\left[\sin\left(2\pi g(\tilde{T})\right)\right] &= \int_\theta^{\theta+1}\sin\left(2\pi g\circ\tilde{T}(x)\right)\mathrm{d}x \\ &= \int_\theta^{\theta+1}\sin\left(2\pi x\right)\mathrm{d}x = 0\end{aligned}$$

for all θ. However

$$\mathbb{P}_\theta\left[\sin\left(2\pi g(\tilde{T})\right) = 0\right] = \mathbb{P}_\theta\left[\sin(2\pi X) = 0\right] = 0,$$

and thus $\tilde{T}$ cannot be complete. Since $\tilde{T}$ was any sufficient statistic, we conclude that the model admits no sufficient complete statistics.

The previous example shows that we are not always guaranteed to find sufficient complete statistics. When we do have access to one however, the following theorem says we can use it to construct MVUEs.

Theorem 6.5 (Lehmann-Scheffé Theorem). *Let T be a sufficient and complete statistic for the statistical model $\mathcal{P}$ and let $\widehat{\gamma}$ be an unbiased estimator for $\gamma = g(\theta) \in \mathbb{R}^k$.*

Then $\widehat{\gamma}_T = \mathbb{E}_\theta[\widehat{\gamma} \mid T]$ is an MVUE for γ.

Proof. Let $\tilde{\gamma}$ be another unbiased estimator of γ. Then $\tilde{\gamma}_T = \mathbb{E}_\theta[\tilde{\gamma} \mid T]$ is also an unbiased estimator. By definition $\tilde{\gamma}_T, \widehat{\gamma}_T$ are both functions of T, independent of θ by sufficiency, so we can define $f(T) := \tilde{\gamma}_T - \widehat{\gamma}_T$. Since both are unbiased estimators of γ we have that $\mathbb{E}_\theta[f(T)] = 0$ for all θ and since T is complete we have that $f(T) = 0$ P_θ almost surely for all θ. This proves that $\tilde{\gamma}_T = \widehat{\gamma}_T$ a.s. and therefore that for all θ

$$\mathrm{Cov}_\theta(\widehat{\gamma}_T) = \mathrm{Cov}_\theta(\tilde{\gamma}_T) \preceq \mathrm{Cov}_\theta(\tilde{\gamma}),$$

where the inequality follows from the Rao-Blackwell theorem. Since $\tilde{\gamma}$ was arbitrary the result follows. □

Examples.

1. **Uniform.** Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{U}[0, \theta]$ random variables. Recall that $\mathbb{E}_\theta[X_{\max}] = \frac{n}{n+1}\theta$.

   We have seen that $X_{\max}$ is complete and sufficient; hence $\widehat{\theta} = \frac{n+1}{n} X_{\max}$ is the MVUE. Note the CRLB does not apply as the distribution is not regular enough; e.g. Reg 1 is violated as the support depends on the parameter.

2. **Normal.** Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$ random variables. We know this is a strictly 2-parameter exponential family, so $T = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is complete and sufficient. As $(\bar{X}, S^2)$ is unbiased and a function of T, it is the MVUE. (Here $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 := \frac{1}{n-1}(X_i - \bar{X})^2$.)

   Remember that for S^2 the Cramer-Rao bound is not attained.

3. **Poisson 1.** Let $X = (X_1, \ldots, X_n)$ be a sample of i.i.d. $Po(\lambda)$ random variables. Recall that

   $$\widehat{\lambda}_{\mathrm{MME}} = \widehat{m}_1 = \frac{1}{n}\sum_{i=1}^n X_i \qquad \text{and} \qquad \tilde{\lambda}_{\mathrm{MME}} = \widehat{m}_2 - \widehat{m}_1^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$$

   are two moment estimators for λ.

   The Poisson family is a strictly 1-parameter exponential family with canonical observation $T(X) = \bar{X}$) (for the joint distribution). Thus $\bar{X}$ is a sufficient and complete statistic.

   Hence the Lehman-Scheffé Theorem tells us that $\widehat{\lambda}_{\mathrm{MME}}$ is the MVUE.

   What is the Cramer-Rao bound? For a single observation, $S(x, \lambda) = \frac{x}{\lambda} - 1$ and $i_X(\lambda) = \lambda^{-1}$. Thus the CR lower bound is λ/n. Since we can also check that $\mathrm{Var}(\bar{X}) = \lambda/n$, we confirm that $\widehat{\lambda}_{\mathrm{MME}} = \bar{X}$ is efficient (it achieves the CRLB).

4. **Poisson 2.** What about the other estimator above, $\tilde{\lambda}_{\mathrm{MME}}$? Well, doing a little calculation reveals that $X_i \mid \{\sum_{j=1}^n X_j = k\} \sim \mathrm{Bin}(k, 1/n)$. So, using Rao-Blackwell to 'improve' the

unbiased estimator $S^2 = \frac{n}{n-1}\tilde{\lambda}_{\text{MME}}$ by the sufficient statistic $\bar{X}$, we get

$$\mathbb{E}_\lambda\left[S^2 \mid \sum_{j=1}^n X_j = k\right] = \frac{n}{n-1}\left\{\mathbb{E}_\lambda\left[X_1^2 \mid \sum_{j=1}^n X_j = k\right] - \frac{k^2}{n^2}\right\} \tag{6.18}$$

$$= \frac{n}{n-1}\left\{\frac{k}{n}\left(1-\frac{1}{n}\right) + \frac{k^2}{n^2} - \frac{k^2}{n^2}\right\} \tag{6.19}$$

$$= \frac{k}{n}. \tag{6.20}$$

So starting from S^2 as an unbiased estimator for λ we arrive at $\bar{X}$ by Rao-Blackwell using $\sum X_i$.

The Lehmann-Scheffé theorem allows us to use complete sufficient statistics to construct MVUEs. We have also seen in Example 6 that there are situations in which no sufficient, complete statistic exists. The question remains then, whether an MVUE always exists. The following theorem says that a necessary and sufficient condition for an estimator to be MVUE is that it is uncorrelated with all unbiased estimators of 0.

Theorem 6.6 (NASC for MVUE). *Suppose that $\mathbb{P} = \{\mathcal{P}_\theta : \theta \in \Theta\}$ be a family of distributions on $\mathcal{X}$ and let $\mathcal{U}$ be the set of unbiased estimators of 0 with finite variance, that is*

$$\mathcal{U} := \left\{h : \mathcal{X} \mapsto \mathbb{R} : \mathbb{E}_\theta[h(X)] = 0,\ \mathbb{E}_\theta[h^2(X)] < \infty\right\}.$$

Then $\widehat{\gamma}$ is an unbiased estimator of $\gamma = g(\theta)$ if and only if $\mathbb{E}_\theta[\widehat{\gamma}U] = 0$ for all $U \in \mathcal{U}$.

Proof. "⇒" Suppose $\widehat{\gamma}$ is MVUE. Then for $U \in \mathcal{U}$ $c \in \mathbb{R}$ define $\widehat{\gamma}_c = \widehat{\gamma} + cU$. Notice that $\widehat{\gamma}_c$ is also unbiased and therefore

$$\operatorname{Var}_\theta[\widehat{\gamma}_c] \geq \operatorname{Var}_\theta[\widehat{\gamma}],$$

or equivalently that

$$c^2 \operatorname{Var}_\theta[U] + 2c \operatorname{Cov}_\theta[\widehat{\gamma}, U] \geq 0.$$

Viewing the LHS above as a quadratic function in c, the inequality implies that the discriminant is non-positive, ie that

$$4 \operatorname{Cov}_\theta[\widehat{\gamma}, U]^2 \leq 0,$$

which implies that $\operatorname{Cov}_\theta[\widehat{\gamma}, U] = 0$.

"⇐" Let $\tilde{\gamma}$ be another unbiased estimator of γ with finite variane. Then $U := \widehat{\gamma} - \tilde{\gamma} \in \mathcal{U}$. Therefore by assumption we have that

$$0 = \mathbb{E}_\theta[\widehat{\gamma}U] = \mathbb{E}_\theta[\widehat{\gamma}^2] - \mathbb{E}_\theta[\widehat{\gamma}\tilde{\gamma}].$$

Since $\mathbb{E}_\theta[\widehat{\gamma}] = \mathbb{E}_\theta[\tilde{\gamma}]$ the above implies that

$$\operatorname{Var}_\theta[\widehat{\gamma}]^2 = \operatorname{Cov}_\theta[\widehat{\gamma}, \tilde{\gamma}]^2 \leq \operatorname{Var}_\theta[\widehat{\gamma}] \operatorname{Var}_\theta[\tilde{\gamma}],$$

whence we conclude that $\operatorname{Var}_\theta[\widehat{\gamma}] \leq \operatorname{Var}_\theta[\tilde{\gamma}]$. □

Example (Example 6 continued.). The above theorem allows us to establish that in Example 6 there exists no MVUE. Suppose that $\widehat{\gamma} = T(X)$ is an unbiased estimator of $\gamma = g(\theta)$. Let $H \in \mathcal{U}$ be any unbiased estimator of 0. Then we have

$$\int_\theta^{\theta+1} H(x)\mathrm{d}x = 0,$$

and differentiating both sides

$$H(\theta+1) = H(\theta), \quad \text{a.e.}$$

By Theorem 6.6 we have that $\mathbb{E}_\theta[\widehat{\gamma}H] = 0$ and therefore that

$$\int_\theta^{\theta+1} T(x)H(x)\mathrm{d}x = 0,$$

and differentiating both sides

$$T(\theta+1)H(\theta+1) = T(\theta)H(\theta), \quad \text{a.e.}$$

which along with the fact that $H(\theta+1) = H(\theta)$ a.e. allows us to conclude that $T(\theta+1) = T(\theta)$ for a.e. θ.

Finally since $\widehat{\gamma}$ is unbiased we have

$$\int_\theta^{\theta+1} T(x)\mathrm{d}x = g(\theta),$$

and differentiating both sides

$$T(\theta+1) - T(\theta) = g'(\theta).$$

Combining everything, the only functions g for which an MVUE may exist are those that have $g' = 0$.

Chapter 7

Bayesian Inference: Conjugacy and Improper Priors

We turn now, in this second half of the course, to the Bayesian view of statistical inference, and look at how we may develop further the theory from Part A.

7.1 Recap of fundamentals

Recall that in Bayesian statistics, parameters are treated as random variables too (rather than having an unknown true value, as in frequentist statistics). At the core of this approach is of course Bayes' Theorem, which you have met several times over the last two years. In our setting it most commonly reads as follows:

Theorem 7.1 (Bayes' Theorem). *Given a* **likelihood** $L(\theta, x)$ *and a* **prior** $\pi(\theta)$ *for* θ*, the* **posterior** *distribution for* θ *(the conditional distribution of* θ *given the data* X*) is given by*

$$\pi(\theta \mid x) = \frac{L(\theta, x)\pi(\theta)}{\int L(\theta', x)\pi(\theta')\,\mathrm{d}\theta'}.$$

(If π *is a mass function replace the integral with a sum.)*

We will often simply write

$$\pi(\theta \mid x) \propto L(\theta, x)\pi(\theta),$$

i.e. **posterior** $\propto$ **likelihood** $\cdot$ **prior**. *The quantity* $p(x) = \int L(\theta', x)\pi(\theta')\,\mathrm{d}\theta'$ *is called the marginal distribution of* X*.*

Proof. Prelims/Part A probability and statistics. □

Remark. The denominator $p(x) = \int L(\theta', x)\pi(\theta')\,\mathrm{d}\theta'$ in the theorem above is called the ***marginal likelihood*** in this context.

Example. Suppose $X \sim \mathrm{Bin}(n, \theta)$, and that our prior distribution for θ is $\mathrm{Beta}(a, b)$, i.e.

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)b - 1}{B(a, b)}, \qquad 0 < \theta < 1.$$

The likelihood function is $L(\theta, x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ for $x = 0, \ldots, n$. So by Bayes' Theorem the

posterior distribution is

$$\pi(\theta \mid x) \propto \text{likelihood} \cdot \text{prior} \tag{7.1}$$

$$\propto \theta^x(1-\theta)^{n-x} \cdot \theta^{a-1}(1-\theta)^{b-1} \tag{7.2}$$

$$= \theta^{a+x-1}(1-\theta)^{n-x+b-1}. \tag{7.3}$$

This is again (up to normalisation) a Beta distribution, with updated parameters $a+x, b+n-x$. This is an example of ***conjugacy***, which we will meet next.

Suppose we choose a, b here such that $\mathbb{E}[\theta] = 0.7$ and $\text{Var}(\theta) = 0.1$. Suppose we then observe:

- $X = 3$ for a number of trials $n = 10$; or alternatively
- $X = 30$ for a number of trials $n = 100$.

In the first case our posterior will have a mean of about 0.5 to 0.6, and in the second case our posterior will have a mean of less than 0.4.

As n increases, the likelihood increasingly overwhelms the prior. This captures the intuition that the second observation seems to be much stronger evidence than the first case that θ is in fact near to 0.3.

Remark. This example illustrates the general effect at play in Bayesian inference: as we gather more (relevant) data, effectively the information we have about the unknown parameter increases and we revise our beliefs accordingly.

7.2 Conjugate priors

We start off now by introducing the notion of ***conjugacy***.

Definition 7.2. Consider a model $(L(\theta, x))_{\theta\in\Theta, x\in\mathcal{X}}$. We say that a family of prior distributions $(\pi_\gamma)_{\gamma\in\Gamma}$ is ***conjugate*** if for all $\gamma \in \Gamma$ and $x \in \mathcal{X}$, there exists $\gamma(x) \in \Theta$ such that $\pi_\gamma(\cdot \mid x) = \pi_{\gamma(x)}(\cdot)$.

We say the prior and the posterior are ***conjugate distributions***, and the prior is a ***conjugate prior*** for the likelihood L.

In other words, a conjugate prior is a prior which, when combined with the likelihood, produces a posterior distribution in the same family as the prior.

Examples. (Normal conjugacy) Consider the model $X_i \sim N(\mu, \sigma^2)$ i.i.d. and let $\tau = \sigma^{-2}$, $\theta = (\tau, \mu)$. We are going to assume a prior of the following form for θ: $\tau \sim \text{Gamma}(\alpha, \beta)$ and conditionally on τ we have $\mu \mid \tau \sim N(\nu, (k\tau)^{-1})$ where $k > 0, \nu \in \mathbb{R}$ are given parameters. In other words

$$\pi(\tau, \mu) = \pi(\tau)\pi(\mu \mid \tau) = \frac{\beta^\alpha}{\Gamma(\alpha)}\tau^{\alpha-1}e^{-\beta\tau}\frac{1}{\sqrt{2\pi}}\sqrt{k\tau}\exp\left\{-\frac{k\tau}{2}(\mu-\nu)^2\right\} \tag{7.4}$$

$$\propto \tau^{\alpha-1/2}\exp\left\{-\tau\left[\beta + \frac{k}{2}(\mu-\nu)^2\right]\right\}. \tag{7.5}$$

Since $L(\theta, \mathbf{x}) = (2\pi)^{-n/2}\tau^{n/2}\exp\left\{-\frac{\tau}{2}\sum_{i=1}^n(x_i-\mu)^2\right\}$ we have that

$$\pi(\theta \mid \mathbf{x}) \propto \tau^{\alpha+\frac{n-1}{2}}\exp\left\{-\tau\left[\beta + \frac{k}{2}(\mu-\nu)^2 + \frac{1}{2}\sum_{i=1}^n(x_i-\mu)^2\right]\right\}. \tag{7.6}$$

The expression in the square bracket is quadratic in μ so that

$$\left[\beta + \frac{k}{2}(\mu-\nu)^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right] = \left[\beta' + \frac{k'}{2}(\mu-\nu')^2\right]$$

where β', k', ν' depend on $n, k, \nu, \mathbf{x}$ but not on μ or τ.

By completing the squares first in μ and then in ν it is a good exercise to check that

$$\left[\beta + \frac{k}{2}(\mu-\nu)^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right] = \frac{1}{2}(k+n)\left(\mu - \frac{k\nu+n\bar{x}}{k+n}\right)^2 + \frac{1}{2}\frac{nk}{n+k}(\bar{x}-\nu)^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \beta$$

where each term can be easily identified. This shows that the prior is conjugate for this likelihood since

$$\pi(\tau, \mu \mid \mathbf{x}) \propto \tau^{\alpha'-\frac{1}{2}} \exp\left\{-\tau\left[\beta' + \frac{k'}{2}(\nu'-\mu)^2\right]\right\}$$

where $\alpha' = \alpha + \frac{n}{2}$. Thus, this tells us that $\tau \mid \mathbf{x} \sim \text{Gamma}(\alpha', \beta')$ and that $\mu \mid \tau, \mathbf{x} \sim N(\nu', (k'\tau)^{-1})$.

It turns out that conjugacy is a general phenomenon for exponential families!

Proposition 7.3 (Conjugate priors for exponential families). *Suppose*

$$L(\theta, x) = h(x)\exp\left\{\sum_{i=1}^{k}\eta_i(\theta)T_i(x) - B(\theta)\right\}$$

defines a k-parameter exponential family. Then the distributions of the form

$$\pi_\gamma(\theta) \propto \exp\left\{\gamma_0 B(\theta) + \sum_{i=1}^{k}\gamma_i\eta_i(\theta)\right\},$$

for parameters $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_k)$ are a conjugate prior family.

Proof. Exercise. □

Example. Let $X = (X_1, \ldots, X_n)$ be a sample of i.i.d. Poi(θ) random variables, so the (joint) likelihood is

$$L(\theta, x) \propto \exp(-n\theta + T(x)\log\theta)$$

where $T(x) = \sum_{i=1}^{n} x_i$. So the natural conjugate prior is of the form

$$\pi(\theta) \propto \exp(\gamma_0\theta + \gamma_1\log\theta).$$

(Note this is normalisable iff $\gamma_0 < 0$ and $\gamma_1 > -1$.)

Writing $\beta = -\gamma_0$ and $\alpha = \gamma_1 + 1$, we have $\pi(\theta) \propto \theta^{\alpha-1}\mathrm{e}^{-\beta\theta}$ which is the pdf of a $\Gamma(\alpha, \beta)$ distribution.

We can easily see that the posterior distribution is $\Gamma(\alpha + T(x), \beta + n)$. So indeed the Gamma distribution is a conjugate prior (for the Poisson likelihood).

Example (Multinomial Distribution and Dirichlet Prior). Consider a multinomial distribution with N trials and K levels with likelihood

$$p(x^{1:K}|\theta) = \theta_1^{x^1}\theta_2^{x^2}\cdots\theta_K^{x^K}.$$

Conjugate priors take the form

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1} \cdots \theta_K^{\alpha_K}, \qquad \sum \theta_i = 1, \theta_i \geq 0.$$

The above defines a proper prior when $\alpha_1, \ldots, \alpha_K > -1$, so it's more natural to parameterise as

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1}, \qquad \sum \theta_i = 1, \theta_i \geq 0,$$

with $\alpha_i > 0$. This is called the Dirichlet distribution denoted $\mathsf{Dirichlet}(\alpha_1, \ldots, \alpha_K)$.

If $(X_1, \ldots, X_n)$ are i.i.d. samples from $p(\cdot|\theta)$, where for $i = 1, \ldots, n$, $x_i = (x_i^1, \ldots, x_i^K)$,

$$p(x_{1:N}|\theta) = \theta_1^{\sum x_j^1} \theta_2^{\sum x_j^2} \cdots \theta_K^{\sum x_j^K},$$

it can be easily seen that the posterior is also Dirichlet

$$p(\theta|x_{1:n}, \alpha) = \mathsf{Dirichlet}\left(\sum x_j^1 + \alpha_1 - 1, \ldots, \sum x_j^K + \alpha_K - 1\right).$$

7.3 Improper priors

So far both the prior and the posterior functions have been probability densities (or mass functions). This is natural given the origin in Bayes' Theorem, but in fact we do not require that the prior be a 'real' probability distribution for the posterior to exist and be well-defined.

Definition 7.4. We say that a pdf/pmf π is an ***improper prior*** if it has infinite mass:

$$\int_\Theta \pi(\theta)\, \mathrm{d}\theta = \infty, \quad \pi(\theta) \geqslant 0 \ \forall \theta \in \Theta$$

(as usual replacing integrals with sums if necessary).

A posterior distribution $\pi(\theta \mid x)$ can be defined as usual as long as

$$\int_\Theta f(x, \theta)\pi(\theta)\, \mathrm{d}\theta < \infty.$$

Examples.

1. Likelihood $X \mid \mu \sim \mathcal{N}(\mu, 1)$ and prior $\pi(\mu) = 1 \ \forall \mu \in \mathbb{R}$. In this case $\log \pi(\mu \mid x) = -\frac{1}{2}(x - \mu)^2 + \text{constant}$, i.e. the posterior distribution is $\mathcal{N}(x, 1)$.
2. Likelihood $X \mid p \sim \mathrm{Bin}(n, p)$ and prior $\pi(p) = [p(1-p)]^{-1}$ (this is the ***Haldane prior***). The posterior is $\pi(p \mid x) \propto p^{x-1}(1-p)^{n-x-1}$ which is improper iff $x = 0$ or $x = n$; so the posterior is not always well-defined.

Exercise. If X is discrete and can take only finitely many values, say $\{z_1, \ldots, z_N\} = \mathcal{X}$, show that we *can't* use an improper prior.

Hint: try proving that the marginal likelihood cannot be finite for all $i = 1, \ldots, N$.

Does this argument work for $\mathcal{X}$ countably infinite? (Try $X \sim \mathrm{Po}(\lambda), \pi(\lambda) = \lambda^{-1}$.)

7.4 Predictive Distributions

Let us briefly touch on how we can make predictions for new datapoints.

Definition 7.5. If $X_1, \ldots, X_n, X_{n+1}$ are i.i.d. obsevations from the distribution $f(x, \theta)$, with prior $\pi(\theta)$, then the ***posterior predictive distribution*** is

$$f(x_{n+1} \mid x) = \int_\Theta f(x_{n+1}, \theta)\pi(\theta \mid x)\,\mathrm{d}\theta$$

where here $x = (x_1, \ldots, x_n)$.

Thus the predictive distribution describes the distribution of a new observation given all the observations we've already made.

Examples.

1. **Poisson likelihood, Gamma prior.** Suppose $Y \sim \mathrm{Poi}(\theta)$ and that our prior for θ is a $\Gamma(\alpha, \beta)$ distribution.

   The marginal likelihood for this model is

   $$m(y) = \int_0^\infty \mathrm{e}^{-\lambda}\frac{\lambda^y}{y!}\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\mathrm{e}^{-\beta\lambda}\,\mathrm{d}\lambda.$$

   On the other hand, we can use that $\pi(\theta \mid y) = \frac{f(y,\theta)\pi(\theta)}{m(y)}$, so $m(y) = \frac{f(y,\theta)\pi(\theta)}{\pi(\theta|y)}$. We have seen previously that in this setting the posterior is $\pi(\theta \mid y) \sim \Gamma(\alpha + y, \beta + 1)$. Hence

   $$m(y) = \frac{\left(\frac{\mathrm{e}^{-\theta}\theta^y}{y!}\right)\left(\frac{\beta^\alpha \mathrm{e}^{-\beta\theta}\theta^{\alpha-1}}{\Gamma(\alpha)}\right)}{\left(\frac{(\beta+1)^{\alpha+y}\theta^{\alpha+y-1}\mathrm{e}^{-(\beta+1)\theta}}{\Gamma(\alpha+y)}\right)} \tag{7.7}$$

   $$= \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)y!}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^y \tag{7.8}$$

   which is the pmf of a $\mathrm{NegBin}(\alpha, \beta)$ distribution.

   Thus we have shown that the densities/masses of the Poisson, Gamma and negative binomial distributions are related by

   $$p_{\mathrm{NegBin}}(y; \alpha, \beta) = \int_0^\infty p_{\mathrm{Po}}(y; \theta) \cdot p_\Gamma(\theta; \alpha, \beta)\,\mathrm{d}\theta.$$

   Hence the predictive distribution has pmf

   $$\pi(y_{n+1} \mid y) = \int_0^\infty p_{\mathrm{Po}}(y_{n+1}; \theta)p_\Gamma(\theta; \alpha + \Sigma y_i, \beta + n)\,\mathrm{d}\theta = p_{\mathrm{NegBin}}(y_{n+1}; \alpha + \Sigma y_i, \beta + n),$$

   so is a negative binomial distribution with parameters $\alpha + \sum_{i=1}^n y_i$ and $\beta + n$.

2. **Gaussian with known variance.** Suppose now that $X_1, \ldots, X_{n+1}$ are i.i.d. $\mathcal{N}(\theta, \sigma^2)$ random variables, where σ^2 is known, and that our prior distribution for the mean is $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$. We want to predict X_{n+1}, having seen $X_1, \ldots, X_n$.

   The posterior after the first n observations is

   $$\pi(\theta \mid x) \propto \pi(\theta)p(x \mid \theta) \propto \exp\left[-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right]\prod_{i=1}^n \exp\left[-\frac{1}{2\sigma^2}(x_i - \theta)^2\right] \tag{7.9}$$

   $$\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_0^2}(\theta - \mu)^2 - \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right]\right) \tag{7.10}$$

   $$\propto \exp\left[-\frac{1}{2\sigma_n^2}(\theta - \mu_n)^2\right] \tag{7.11}$$

where, by completing the square, we find that $\mu_n = \frac{\sigma_0^{-2}\mu_0 + \sigma^{-2}\sum_{i=1}^n x_i}{\sigma_0^{-2} + n\sigma^{-2}}$ and $\sigma_n^{-2} = \sigma_0^{-2} + n\sigma^{-2}$.

(Observe that if $\sigma^2 = \sigma_0^2$ then the prior has the same weight as that of a single extra observation.)

So $\theta \mid X \sim \mathcal{N}(\mu_n, \sigma_n^2)$ and $X_{n+1} \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$. We can rewrite these two facts as

$$\theta = \mu_n + \sigma_n Z, \quad X_{n+1} = \theta + \sigma Y$$

for some independent $Y, Z \sim \mathcal{N}(0,1)$, and so $X_{n+1} = \mu_n + \sigma_n Z + \sigma Y$. Thus $X_{n+1} \mid X \sim \mathcal{N}(\mu_n, \sigma^2 + \sigma_n^2)$.

(We could also have arrived at this last result by directly integrating the densities; our method was just an equivalent and simpler approach in this case.)

Chapter 8

Non-Informative Priors

We've just seen that priors don't always have to be probability distributions. When may we want to make use of this?

We're used to the notion of a ***subjective prior***, a distribution representing our ***prior knowledge*** about the parameter before any data is collected. With this approach, we may try different priors representing different 'points of view'.

This is in contrast to the concept of an ***objective prior*** (a ***non-informative prior***) which we'll explore in this chapter. This is a prior which is somehow 'automatic', reflecting the lack of any initial knowledge about the parameter — and crucially may have no probabilistic interpretation, and so doesn't have to be a valid probability distribution. Non-informative priors can be used when little no reliable information is available.

There are several approaches for defining a non-informative prior, three of which we'll mention here.

8.1 Uniform priors

Definition 8.1. The ***uniform prior*** or ***flat prior*** is the prior $\pi(\theta) \propto 1$.

This is the obvious, naive representation of lack of information; every value being equally likely. Under this prior, the posterior is

$$\pi(\theta \mid x) = \frac{L(\theta, x)}{\int_\Theta L(\theta, x)\,\mathrm{d}\theta},$$

which is well defined as long as $\int_\Theta L(\theta, x)\,\mathrm{d}\theta < \infty$.

Example. Let $X \sim \mathrm{Exp}(\theta)$ and $\pi(\theta) = 1$. The marginal likelihood is $\int_0^\infty \mathrm{e}^{-\theta x}\theta\,\mathrm{d}\theta$ which is finite for all $x > 0$, so the posterior is well-defined. But does it have nice properties?

Let $\eta = \log\theta$. Then the prior for η is

$$\tilde{\pi}(\eta) = \pi(\theta(\eta))\,\frac{\mathrm{d}\theta}{\mathrm{d}\eta} = \frac{\mathrm{d}\theta}{\mathrm{d}\eta} = \mathrm{e}^\eta \neq 1.$$

We see that after reparameterisation the prior is not flat anymore; in fact, as a prior in η, $\tilde{\pi}$ is very informative (large values are much more likely than small ones).

8.2 Jeffrey's prior

The last example motivates the construction of a prior that does not depend on the parameterisation.

Definition 8.2. In the one-dimensional case ***Jeffrey's prior*** is given by

$$\pi(\theta) \propto \sqrt{I_\theta}$$

where $I_\theta = \mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2}\ell(\theta, x)\right]$ is the Fisher information.

Remark. Why does this work? If $\theta = g(\psi)$ for some one-to-one differentiable function g then the reparameterised prior is

$$\tilde{\pi}(\psi) \propto \pi(g(\psi))|g'(\psi)| = \sqrt{I_\theta}|g'(\psi)|.$$

Recall that $I_\psi = (g'(\psi))^2 I_\theta$, so $\sqrt{I_\psi} = \sqrt{I_\theta}|g'(\psi)|$. Hence $\tilde{\pi}(\psi) \propto \sqrt{I_\psi}$.

So indeed Jeffrey's prior is invariant under reparametrisation.

8.2.1 Jeffrey's prior in higher dimensions

This definition generalises naturally to higher dimensions:

Definition 8.3. The ***k-dimensional Jeffrey's prior*** is given by

$$\pi(\theta) \propto |I_\theta|^{1/2},$$

where $|I_\theta| = \det I_\theta$ and I_θ is the Fisher information matrix, so under the standard regularity assumptions $(I_\theta)_{ij} = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta, x)\right]$.

It is easy to check that this is indeed invariant under one-to-one reparametrisation.

Example. Suppose $X \sim \text{Po}(\lambda)$, so that $f(x, \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$ for $x = 0, 1, 2, \ldots$.

Then Jeffrey's prior is

$$\pi(\lambda) \propto \sqrt{I_X(\lambda)} = \sqrt{\mathbb{E}[(\ell'(\lambda, X))^2]} \tag{8.1}$$

$$= \sqrt{\mathbb{E}\left[\left(\frac{x}{\lambda} - 1\right)^2\right]} \tag{8.2}$$

$$= \sqrt{\sum_{x=0}^{\infty} f(x, \lambda)\left(\frac{x-\lambda}{\lambda}\right)^2} \tag{8.3}$$

$$= \sqrt{e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!}\left(\frac{x^2}{\lambda^2} - \frac{2x}{\lambda} + 1\right)} \tag{8.4}$$

$$= \sqrt{\frac{1}{\lambda^2}\,\mathbb{E}\left[\mathbb{E}[(X-\lambda)^2]\right]} \tag{8.5}$$

$$= \lambda^{-1/2}. \tag{8.6}$$

Note this is an improper prior.

8.3 Maximum entropy prior

Another possible approach for constructing a non-informative prior is inspired by information theory.

Definition 8.4. The ***entropy*** of a pdf/pmf π is defined as

$$\mathrm{Ent}[\pi] = -\int_\Theta \pi(\theta)\log\pi(\theta)\,\mathrm{d}\theta.$$

As always, replace the integral with a sum if π is a pmf.

Remark. In the continuous case, entropy is often referred to as the ***differential entropy***.

A maximum entropy probability distribution has entropy that is at least as great as that of all other members of a specified class of probability distributions. According to the principle of maximum entropy, if nothing is known about a distribution except that it belongs to a certain class (usually defined in terms of specified properties or measures), then the distribution with the largest entropy should be chosen as the least-informative default.

Since maximizing entropy minimizes the amount of prior information built into the distribution it makes sense to pick the prior that *maximises the entropy* subject to any relevant constraints (e.g. a fixed mean).

Example. Suppose we wish to find the distribution π which maximises $\mathrm{Ent}[\pi]$ on $\Theta = \mathbb{R}$ subject to the constraints

$$\int_{-\infty}^{\infty}\pi(\theta)\,\mathrm{d}\theta = 1, \quad \int_{-\infty}^{\infty}\theta\pi(\theta)\,\mathrm{d}\theta = \mu \quad \text{and} \quad \int_{-\infty}^{\infty}(\theta-\mu)^2\pi(\theta)\,\mathrm{d}\theta = \sigma^2$$

for fixed μ, σ^2.

The solution is $\pi(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-(\theta-\mu)^2/2\sigma^2}$. This can be shown using variational calculus or using information-theoretic techniques (a proof is seen on a problem sheet in the Information Theory course).

Thus the Gaussian distribution is the maximum-entropy distribution for the real line when we fix the mean and variance.

Remark. The maximum entropy distribution does not always exist (for example the class of distributions may have unbounded entropy).

The previous example leads us to a more general theorem, which we shall not prove:

Theorem 8.5. *Let*

$$\pi(\theta) = \exp\Big[\sum_{i=1}^{p}\lambda_i T_i(\theta) - B(\lambda)\Big]\ \forall\theta\in\Theta,$$

be a probability density function and suppose that

$$\int T_i(x)\pi(\theta)\mathrm{d}x = t_i, \quad j = 1,\dots,p. \tag{8.7}$$

Then π uniquely maximises $\mathrm{Ent}[\pi]$ *among all densities satisfying the constraint* (8.7).

Proof. Let Π be the class of distributions satisfying the constraints.

Recall that for two distributions $\tau_1 \ll \tau_2$ the Kullback-Leibler divergence $\mathsf{KL}(\tau_1\|\tau_2)$ is defined through

$$\mathsf{KL}(\tau_1\|\tau_2) = \int \tau_1(\mathrm{d}x)\log\left(\frac{\mathrm{d}\tau_1}{\mathrm{d}\tau_2}\right),$$

where $\mathrm{d}\tau_1/\mathrm{d}\tau_2$ is the Radon-Nikodym derivative. If τ_1 is not absolutely continuous w.r.t. τ_2 we set

$\mathsf{KL}(\tau_1 \| \tau_2) = +\infty$. It is a simple application of Jensen's inequality to check that $\mathsf{KL}(\tau_1 \| \tau_2) \geq 0$.

We now have the tools to prove the result. Let π' be any element of Π. Then

$$\begin{aligned}
\mathrm{Ent}[\pi'] &= -\int \pi'(x) \log \pi'(x) \mathrm{d}x \\
&= -\int \pi'(x) \log \left(\frac{\pi'(x)}{\pi(x)} \right) \mathrm{d}x - \int \pi'(x) \log \pi(x) \mathrm{d}x \\
&= -\mathsf{KL}(\pi' \| \pi) - \int \pi'(x) \left[\sum_{i=1}^{p} \lambda_i T_i(x) \right] \mathrm{d}x + B(\lambda)
\end{aligned}$$

and since $\pi[T_i] = \pi'[T_i]$ for all i

$$\begin{aligned}
&= -\mathsf{KL}(\pi' \| \pi) + B(\lambda) - \int \pi(x) \left[\sum_{i=1}^{p} \lambda_i T_i(x) \right] \mathrm{d}x \\
&= -\mathsf{KL}(\pi' \| \pi) + \mathrm{Ent}[\pi].
\end{aligned}$$

Rearranging we obtain

$$\mathrm{Ent}[\pi] - \mathrm{Ent}[\pi'] = \mathsf{KL}(\pi' \| \pi) \geq 0. \qquad \square$$

Example (continued). In the example above, our two constraints were $\mathbb{E}[T_1(\theta)] = \mu$ and $\mathbb{E}[T_2(\theta)] = \sigma^2$, where $T_1(\theta) = \theta$ and $T_2(\theta) = (\theta - \mu)^2$.

The above theorem then gives that the maximum-entropy prior is of the form $\pi(\theta) \propto \exp(\lambda_1 \theta + \lambda_2 (\theta - \mu)^2)$. The two constraints then imply that $\lambda_1 = 0$ and $\lambda_2 = -\frac{1}{2\sigma^2}$, thus giving the Gaussian distribution we just saw.

Example. Suppose $a_0 \leqslant a_1 \leqslant \cdots \leqslant a_p$ and $\theta \in (a_0, a_p)$.

Consider the constraints $\pi(\theta \in (a_{j-1}, a_j]) = \phi_j$ for $j = 1, \ldots, p$. This is equivalent to requiring $\mathbb{E}[T_j(\theta)] = \phi_j$ for $j = 1, \ldots, p$, where $T_j(\theta) = \mathbb{1}_{\{a_{j-1} < \theta \leqslant a_j\}}$.

Under these conditions the maximum-entropy distribution is of the form

$$\pi(\theta) \propto \exp \left[\sum_{j=1}^{p} \lambda_j \mathbb{1}_{\{a_{j-1} < \theta \leqslant a_j\}} \right], \quad a_0 \leqslant \theta \leqslant a_p.$$

Hence π_θ is piecewise constant on the intervals $(a_i, a_{i+1}]$.

Chapter 9

Hierarchical Models

Chapter 5 p101 in Gelman, Carlin et al. Bayesian Data analysis. Section 7.7 p170 in Garthwaite, Joliffe, and Jones Statistical Inference. p253 Lehmann and Casella Theory of point estimation

9.1 Example

In certain situations, the data we are modelling has a natural ***hierarchical*** structure. We illustrate this first with an extended example.

Example (Study of cardiac treatment across different hospitals). Consider the dataset in fig. 9.1 consisting of mortality rates in infant cardiac surgery across $I = 12$ hospitals. Each hospital i conducts n_i surgeries, Y_i of which result in death. We use the natural model for the number of deaths at each hospital as $Y_i \sim \text{Bin}(n_i, \theta_i)$, where θ_i is an unknown parameter.

How do we model the mean mortality rates $\theta = (\theta_1, \ldots, \theta_{12})$?

Three broad approaches come to mind:

- **Identical parameters.** We assume all the θ_i are identical. This ignores the structure of the problem and pools all the data. In this case this means we're assuming the surgery success rate doesn't depend on which hospital conducts the surgery.
- **Independent parameters.** We assume all the θ_i are independent, i.e. entirely unrelated. The results from each unit can be analysed independently. In this case this means we're assuming there is nothing similar about the surgery at different hospitals, and the failure rates at different hospitals don't depend on each other in any way.
- **Exchangeable parameters.** We assume the θ_i are similar; no one hospital is *a priori* any better than another. More on this later.

Let's see how the first two approaches can work in this situation, where relevant examining our estimates for hospitals A and H in particular:

- **All θ_i equal (frequentist approach).** The model is $Y_i \sim \text{Bin}(n_i, \theta)$ for each i, so $\sum Y_i \sim \text{Bin}(\sum n_i, \theta)$. Thus the MLE for θ is $\widehat{\theta} = \frac{\sum y_i}{\sum n_i} = 0.0739$.
- **Independent θ_i (frequentist approach).** The model is $Y_i \sim \text{Bin}(n_i, \theta_i)$ independently for each i. The MLE for each θ_i is $\widehat{\theta}_i = \frac{y_i}{n_i}$. So in particular $\widehat{\theta}_A = 0$ and $\widehat{\theta}_H = 0.1442$.
- **All θ_i equal (Bayesian approach).** The model is $Y_i \mid \theta \sim \text{Bin}(n_i, \theta)$ for each i, and we'll use the prior $\theta \sim \text{Beta}(a, b)$ with $a = 4$ and $b = 46$. (We choose the Beta distribution since

	A	B	C	D	E	F	G	H	I	J	K	L	$\sum$
y_i	0	18	8	46	8	13	9	31	14	8	29	24	208
n_i	47	148	119	810	211	196	148	215	207	97	256	360	2814

Figure 9.1: *Number of infant cardiac surgeries and number of mortalities across 12 hospitals.*

it's a conjugate prior for the binomial distribution; and the choice of parameters a, b will be discussed later.) The posterior mean of θ is then $\frac{\sum y_i+\alpha}{\sum n_i+\alpha+\beta} = 0.0740$.

- **Independent θ_i (Bayesian approach).** The model is $Y_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$ independently for each i, with i.i.d. priors $\theta_i \sim \text{Beta}(a, b)$. The posterior mean for each θ_i is then $\frac{y_i+\alpha}{n_i+\alpha+\beta}$ which takes value 0.0412 for hospital A and 0.1321 for hospital H.

The first method (frequentist, equal parameters) gives some pretty unlikely results (e.g. the observed death rate for hospital H is very unlikely given our estimated θ), and the second method (frequentist, independent parameters) totally ignores data from other hospitals when estimating θ_i for a particular hospital; but this is the same medical procedure, so this is unnatural.

The third method (Bayesian, equal parameters) has the same problem as in the frequentist setting, but *the last method (Bayesian, independent parameters drawn from the same distribution) seems to address these issues*; the parameters are different for each hospital, but are all drawn from the same distribution, whose parameters can be inferred from the entire dataset.

This is what we mean by a ***natural hierarchical structure***.

How can we estimate the parameters, then, of the shared prior distribution?

Example (continued). In the example above, the approach we settled on models the θ_i as drawn independently from a Beta(a, b) distribution. How do we estimate the parameters (a, b)?

- **Approximate empirical Bayes approach.** The most obvious way to estimate (a, b) is to use a standard frequentist technique; the method of moments. In this context, this means we pick (a, b) so that the prior distribution has the same mean and variance as the sample mean and sample variance of the observed maximum likelihood estimates for the parameters θ_i.

  Specifically, we calculate $r_i = y_i/n_i$ for each hospital (this is the observed mortality rate; the MLE for θ_i) and we calculate the sample mean and sample variance of the set $\{r_1, \ldots, r_{12}\}$; and then solve for $\widehat{a}, \widehat{b}$ such that Beta$(\widehat{a}, \widehat{b})$ has the same mean and variance.

  (Then we use Beta$(\widehat{a}, \widehat{b})$ as our shared prior for the θ_i, to obtain the posterior distribution $\pi(\theta_i \mid \widehat{a}, \widehat{b}, y_i)$ for each θ_i as described above.)

This approach is reasonable, but we have the problem that we are using the same data twice — once to pick $\widehat{a}, \widehat{b}$) and once to find the individual posteriors for the θ_i. This leads to overconfidence in the posterior distributions! Moreover, we're making a fixed choice of $(\widehat{a}, \widehat{b})$ and working with that choice, so the posterior distributions we derive will not reflect the inherent uncertainty in the values of the parameters (a, b).

This motivates a more subtle approach that is Bayesian throughout.

- **hierarchical Bayesian model.** We may instead assume a joint probability model for (θ, a, b). In other words θ, a and b are *all* treated as random variables.

  As before (except now treating these explicitly as conditional distributions) we say $\theta_i \mid (a, b) \sim$

$\text{Beta}(a, b)$ independently for each i, and we now also model the marginal distribution of (a, b) as $(a, b) \sim p(a, b)$. This is effectively a prior distribution for (a, b); we call it the ***hyperprior***.

In summary, our hierarchical model has three layers:

- **Level 1:** $Y_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$ independently for each i;
- **Level 2:** $\theta_i \mid (a, b) \sim \text{Beta}(a, b)$ independently for each i;
- **Level 3:** $(a, b) \sim p(a, b)$ for some ***hyperprior*** distribution $p(a, b)$.

Note that the θ_i are now *not* independent, but they are *conditionally* independent given a, b.

9.2 Definition

The empirical Bayes approach will be discussed in more detail later in the course; the hierarchical Bayes approach can be defined in generality as follows:

Definition 9.1. The building blocks of a ***hierarchical Bayesian model*** for the observations $Y_1, \ldots, Y_n$ with parameters $\theta_1, \ldots, \theta_n$ and ***hyperparameter*** ϕ are

- $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ a family of probability distributions on $\mathcal{A}$. We write $p(y|\theta)$ for the pmf/pdf of P_θ.
- $\{\pi_\phi, \phi \in \Phi\}$ a family of probability distributions on Θ (the parametrized priors). We write $p(\theta|\phi)$ for the pdf/pmf of π_ϕ.
- and P be a distribution on Φ (the ***hyperprior distribution***). We write $p(\phi)$ for its pdf/pmf.

Then the corresponding hierarchical model is the following joint distribution of the Y_j, θ_i and ϕ.

I: $y_j|\theta_j, \phi \sim p(y_j|\theta_j)$ independently for each j, (note this does not depend on ϕ)

II: $\theta_j|\phi \sim p(\theta_j|\phi)$

III: $\phi \sim p(\phi)$

The ***joint prior*** distribution is $p(\theta, \phi) = p(\theta|\phi)p(\phi)$ and the ***joint posterior*** distribution is $p(\theta, \phi|y) \propto p(y|\theta, \phi)p(\theta, \phi) = p(y|\theta)p(\theta|\phi)p(\phi)$.

Example (continued). In the case of the hospital data, recall that our model is $Y_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$ independently for each i, with i.i.d. priors $\theta_i \sim \text{Beta}(a, b)$ and a hyper-prior $p(a, b)$. Since conditionally on (a, b) the (y_i, θ_i) are independent, the joint posterior distribution is

$$p(\theta, a, b \mid y) \propto p(y \mid \theta)p(\theta \mid a, b)p(a, b) \tag{9.1}$$

$$= \left(\prod_{i=1}^{I} p(y_i \mid \theta_i)\right)\left(\prod_{i=1}^{I} p(\theta_i \mid a, b)\right) p(a, b) \tag{9.2}$$

$$\propto \left(\prod_{i=1}^{I} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i}\right)\left(\prod_{i=1}^{I} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta_i^{a-1}(1-\theta_i)^{b-1}\right) p(a, b). \tag{9.3}$$

Thus we have

$$p(\theta \mid a, b, y) \propto \prod_{i=1}^{I} \theta_i^{a+y_i-1}(1-\theta_i)^{b+n_i-y_i-1} \propto \prod_{i=1}^{I} p(\theta_i|a, b, y_i)$$

(what the $\propto$ symbol means is that we only keep the terms that involve θ). This shows that, given a, b, y, the θ_i have independent beta posteriors.

On the other hand, the posterior for (a, b) is $p(a, b \mid y) \propto p(a, b)p(y \mid a, b)$. Observe first that

$p(y \mid a,b) = \prod_i p(y_i \mid a,b)$ by conditional independence given a,b. Let us call $G_{a,b} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ the normalisation constant of the Beta(a,b) distribution.

$$p(y_i \mid a,b) = \int p(y_i|\theta_i)p(\theta_i|a,b)d\theta \tag{9.4}$$

$$= \int \binom{n_i}{y_i}\theta^{y_i}(1-\theta)^{n_i-y_i}G(a,b)\theta^{a-1}(1-\theta)^{b-1}d\theta \tag{9.5}$$

$$\propto G(a,b)\int \theta^{y_i+a-1}(1-\theta)^{n_i-y_i+b-1}G(a,b)d\theta \tag{9.6}$$

$$= G(a,b)G(y_i+a, n_i-y_i+b) \tag{9.7}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(b+n_i-y_i)\Gamma(a+y_i)}{\Gamma(a+b+n_i)}. \tag{9.8}$$

Thus

$$p(a,b|y) \propto p(a,b)\prod_{i=1}^{I}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(b+n_i-y_i)\Gamma(a+y_i)}{\Gamma(a+b+n_i)}.$$

Remark. How can we draw from the joint posterior $p(\theta,\phi|y)$ in general?

1. Draw $\phi \sim p(\phi|y)$.
2. Draw $\theta \sim p(\theta|\phi,y)$.
3. If needed draw predictive values $\tilde{y}$ from $p(y|\theta)$.

9.3 Exchangeability

In the model we've seen, the parameters θ_i were conditionally independent given the hyperparameter vector ϕ.

This is a special case of a property that is in general desirable:

Definition 9.2. The distribution of a random vector $\theta = (\theta_1,\dots,\theta_I)$ is ***symmetric***, or ***exchangeable***, if

$$(\theta_1,\dots,\theta_I) \overset{d}{=} (\theta_{\sigma(1)},\dots,\theta_{\sigma(I)})$$

for any permutation σ.

Intuitively, this says that 'no one parameter is *a priori* to be treated differently from any of the other parameters'.

Let us see that conditional independence indeed satisfies this property:

Proposition 9.3. *If $\theta = (\theta_1,\dots,\theta_I)$ has (prior) distribution*

$$p(\theta) = \int \left[\prod_{i=1}^{I}\pi(\theta_i \mid \psi)\right] g(\psi)\,\mathrm{d}\psi$$

for some ψ with distribution $g(\psi)$, i.e. the θ_i are conditionally independent given ψ, then the distribution of θ is exchangeable (symmetric).

Proof. Exercise. □

In fact, this is sufficient:

Theorem 9.4 (De Finetti). *All exchangeable sequences are of the above form in the large sample limit.*

Proof. Omitted. □

9.4 Gaussian data example

This section is devoted to a single example of conjugate normal hierarchy. Similar examples can be found on p113 in *Bayesian data analysis* by Gelman, Carlin, et. al. and on page 171 of Garthwaite Joliffe and Jones.

Let us start by describing our model:

- For each $j = 1, \dots, J$, the $X_{i,j}, i = 1, \dots, n_j$ are i.i.d. $N(\theta_j, \sigma^2)$. The $X_{i,j}$ are also independent across different j's.
- The θ_i are i.i.d $N(\mu, \tau^2)$ (conditionally on $\phi = (\mu, \tau^2)$).
- The hyperparameter ϕ has the improper prior distribution $p(\phi) \propto$ constant.

We are going to use the notation $\mathbf{x} = \{x_{i,j}\}$ for the whole data set, $\mathbf{x}_j = \{x_{i,j}, i = 1, \dots n_j\}$ for the observations in group j, and

$$\bar{x}_{\cdot,j} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{i,j}$$

for the mean of $\mathbf{x}_j$.

We start by observing that

$$p(\mathbf{x}_j \mid \theta_j) \propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n_j} (x_{i,j} - \theta_j)^2 \right\} \tag{9.9}$$

$$= \exp\left\{ -\frac{1}{2\sigma^2} n_j (\theta_j^2 - 2\theta_j \bar{x}_{\cdot,j}) \right\} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n_j} x_{i,j}^2 \right\} \tag{9.10}$$

$$= \exp\left\{ -\frac{n_j}{2\sigma^2} (\theta_j - \bar{x}_{\cdot,j})^2 \right\} \exp\left\{ \frac{n_j}{2\sigma^2} \bar{x}_{\cdot,j}^2 \right\} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n_j} x_{i,j}^2 \right\} \tag{9.11}$$

$$\propto \exp\left\{ -\frac{n_j}{2\sigma^2} (\theta_j - \bar{x}_{\cdot,j})^2 \right\} \propto N(\bar{x}_{\cdot,j} \mid \theta_j, \sigma_j^2) \tag{9.12}$$

where $\sigma_j^2 = \sigma^2 / n_j$ is the variance of $\bar{x}_{\cdot,j}$. (this is just a fancy way of saying that $\bar{x}_{\cdot,j}$ is sufficient for θ_j),

Let us start by writing the ***joint posterior distribution***. Using the usual posterior $\propto$ prior $\times$ likelihood formula we have

$$p(\theta, \phi \mid \mathbf{x}) \propto p(\phi) p(\theta \mid \phi) p(\mathbf{x} \mid \theta) \tag{9.13}$$

$$\propto p(\phi) \prod_{j=1}^{J} N(\theta_j \mid \mu, \tau^2) \prod_{j=1}^{J} p(\mathbf{x}_j \mid \theta_j) \tag{9.14}$$

$$\propto p(\phi) \prod_{j=1}^{J} N(\theta_j \mid \mu, \tau^2) \prod_{j=1}^{J} N(\bar{x}_{\cdot,j} \mid \theta_j, \sigma_j^2) \tag{9.15}$$

$$\propto p(\phi) \left[\prod_{j=1}^{J} \exp\left\{ -\frac{1}{2\sigma_j^2} (\bar{x}_{\cdot,j} - \theta_i)^2 \right\} \right] \left[\prod_{j=1}^{J} \tau^{-1} \exp\left\{ -\frac{1}{2\tau^2} (\theta_i - \mu)^2 \right\} \right]. \tag{9.16}$$

Let us now determine the ***conditional posterior distribution of θ given ϕ***. In a hierarchical model, once the hyperparameter is given, the parameters θ_j are independent. Thus

$$p(\theta \mid \phi, \mathbf{x}) = \prod_{j=1}^{J} p(\theta_j \mid \phi, \mathbf{x}_j).$$

Conditionally on ϕ, we simply have J independent unknown normal means given a normal prior distribution. For each j, we thus have a simple Gaussian conjugate model: the observations $\mathbf{x}_j$ are i.i.d $N(\theta_j, \sigma^2)$, σ^2 known and $\theta_j \sim N(\mu, \tau^2)$. It can easily be checked that the posterior is still Gaussian

$$\theta_j \mid \mu, \tau^2, \mathbf{x}_j \sim N(\widehat{\theta}_j, V_j)$$

where

$$\widehat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\bar{x}_{\cdot,j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}.$$

Observe that the posterior mean is a weighted average of the prior mean μ and of the sample mean $\bar{x}_{\cdot,j}$ of group j.

We can now move on to the ***marginal posterior distribution*** of the hyperparameter ϕ. Once we have obtained the joint posterior $p(\theta, \phi \mid \mathbf{x})$ we can obtain the marginal posterio of the hyperparameter $p(\phi \mid \mathbf{x})$ by integrating out the parameter θ. But for the hierarchical normal model we can simply consider directly

$$p(\phi|\mathbf{x}) \propto p(\phi)p(\mathbf{x} \mid' phi).$$

Usually, this decomposition is no help because the *marginal likelihood* term $p(\mathbf{x} \mid \phi)$ cannot generally be written in closed form. But in the present case there is a particularly simple form to this marginal likelihood. The key observation is that, conditionally on $\phi = (\mu, \tau^2)$ the $\bar{x}_{\cdot,j}$ are independent and

$$\bar{x}_{\cdot,j} \mid \phi \sim N(\mu, \sigma_j^2 + \tau^2) \tag{9.17}$$

(take the time to think about it).

Thus we can write

$$p(\phi \mid \mathbf{x}) \propto p(\phi) \prod_{j=1}^{J} N(\bar{x}_{\cdot,j} \mid \mu, \sigma_j^2 + \tau^2).$$

Now we are going to make some assumptions about the hyperparameter distribution $p(\phi)$. We are going to assume a non-informative flat prior for μ given τ^2:

$$p(\mu, \tau^2) = p(\mu \mid \tau^2)p(\tau^2) \propto p(\tau^2).$$

We can decompose the posterior into

$$p(\phi \mid \mathbf{x}) = p(\mu \mid \mathbf{x}, \tau^2)p(\tau^2 \mid \mathbf{x})$$

(this is just the usual Bayes formula for pmf/pdf's for the variables μ and τ^2 under their conditional distribution given $\mathbf{x}$).

Thus we see that

$$p(\mu \mid \mathbf{x}) = \frac{p(\phi \mid \mathbf{x})}{p(\tau \mid \mathbf{x})} \propto p(\phi \mid \mathbf{x})$$

(the $\propto$ here is as a function of μ).

Pugging in our particular form of hyperprior into (9.17), we see that taking a log will give us a quadratic expression in μ

$$-\log p(\mu, \tau^2 \mid \mathbf{x}) = \sum_{j=1}^{J} \frac{1}{2(\sigma_j^2 + \tau^2)}(\mu - \bar{x}_{\cdot,j})^2 + \text{cstt not depending on } \mu.$$

Thus $\mu \mid \tau^2, \mathbf{x} \sim N(\widehat{\mu}, V_\mu)$ where we only need to find the mean $\widehat{\mu}$ and variance V_μ. It can be checked by "completing the squares" that

$$V_\mu^{-1} = \sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2} \quad \text{and} \quad \widehat{\mu} = V_\mu \sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2} \bar{x}_{\cdot,j}.$$

We thus have a proper posterior for μ given τ^2. Finally we want the ***posterior distribution of*** τ^2.

$$p(\tau^2 \mid \mathbf{x}) = \frac{p(\phi \mid \mathbf{x})}{p(\mu \mid, \tau^2 \mathbf{x})} \tag{9.18}$$

$$\propto \frac{p(\tau^2) \prod_{j=1}^{J} N(\bar{x}_{\cdot,j} \mid \mu, \sigma_j^2 + \tau^2}{N(\mu \mid \widehat{\mu}, V_\mu)}. \tag{9.19}$$

Observe that the left hand-side does not depend on μ so that the right-hand side cannot depend on μ as well. We can thus choose to evaluate it at $\mu = \widehat{\mu}$ for simplicity and get

$$p(\tau^2 \mid \mathbf{x}) \propto \frac{p(\tau^2) \prod_{j=1}^{J} N(\bar{x}_{\cdot,j} \mid \widehat{\mu}, \sigma_j^2 + \tau^2}{N(\widehat{\mu} \mid \widehat{\mu}, V_\mu)} \tag{9.20}$$

$$\propto p(\tau^2) V_\mu^{1/2} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(-\frac{(\bar{x}_{\cdot,j} - \widehat{\mu})^2}{2(\sigma_j^2 + \tau^2)} \right). \tag{9.21}$$

Both $\widehat{\mu}$ and V_μ also depends on τ^2 so this is a complicated function of τ^2.

Chapter 10

Decision Theory

- Garthwaite, Joliffe, and Jones *Statistical Inference* Chapter 6 p114
- Lehmann and Casella *Theory of point estimation* Chapter 4 p225 and Chapter 5 p309.
- Young, Smith, et al. *Essential of statistical inference* Chapter 2 p4

Throughout this course we have been exploring ways of estimating parameters, predicting new values, or inferring probability distributions. In the past we have come across hypothesis testing (which we'll explore again at the end of this course). All of these are examples of making decisions based on data. In this section we develop this into a formal theory.

10.1 Basic framework and risk function

As usual, we will assume a data ***model*** $X \mid \theta \sim f(x;\theta)$ from some parametric family $\{f(x,\theta) : \theta \in \Theta\}$, where Θ is our ***parameter space***.

We need the following objects.

- An ***action (or decision) space*** $\mathcal{A}$. Typical examples include $\mathcal{A} = \{0,1\}$ for selecting a hypothesis, or $\mathcal{A} = g(\Theta)$ for estimating a function $g(\theta)$ of a parameter.
- A ***loss function*** $L : \Theta \times \mathcal{A} \to \mathbb{R}_+$. Given an action $a \in \mathcal{A}$, if the true parameter is $\theta \in \Theta$ we incur loss $L(\theta, a)$ (whether this is a loss function or a).
- A ***set of decision rules*** $\mathcal{D} \subseteq \{\Delta : \mathcal{X} \to \mathcal{A}\}$. A decision rule Δ specifies which action we take given observation $x \in \mathcal{X}$.

With these in mind, we define our first measure of 'how bad' a decision rule is:

Definition 10.1. For a given rule $\Delta \in \mathcal{D}$ and parameter $\theta \in \Theta$, the ***(frequentist) risk*** is

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

This is the expected loss assuming the true parameter is θ.

For a given rule Δ we think of the frequentist risk as a profile of risk across the different values of θ.

We also want to consider ***randomised decision rules***, that is a rule that selects among a collection of decision rules according to some probability distribution.

Definition 10.2 (Randomised decision rule). Suppose that the action space $\mathcal{A}$ is equipped with a σ-algebra $\mathfrak{A}$ such that $(\mathcal{A}, \mathfrak{A})$ is a measurable space and let $\mathcal{P}(\mathcal{A})$ be the space of probability measures on $\mathcal{A}$.

A ***randomised decision rule*** $\mathfrak{d}$ is a mapping from $\mathfrak{d} : \mathcal{X} \mapsto \mathfrak{d}_x \in \mathcal{P}(\mathcal{A})$ such that for each $A \in \mathcal{A}$, the mapping $x \mapsto \mathfrak{d}_x(A)$ is measurable.

The frequentist risk of the randomised decision rule $\mathfrak{d}$ is given by

$$R(\theta, \mathfrak{d}) := \int_{\mathcal{X}} \int_{\mathcal{A}} L(\theta, a) \mathfrak{d}_x(\mathrm{d}a) f(x; \theta) \mathrm{d}x.$$

Example. A (***deterministic***) decision rule $\Delta \in \mathcal{D}$ can be thought of as a randomised decision rule by defining $\mathfrak{d}_x = \delta_{\Delta(x)}$, i.e. the probability measure with a single atom located at $\Delta(x)$.

Given a collection $\{\Delta_1, \ldots, \Delta_k\}$ and a probability vector $(p_1, \ldots, p_k)$ we can define the randomised rule $\mathfrak{d} = \sum_{i=1}^k p_i \delta_{\Delta_i}$, which for each x selects the rule $\Delta_i(x)$ with probability p_i.

Examples.

- **Estimation:** $\Delta(x)$ is an estimator of $\theta \in \mathbb{R}^k$ and $L(\theta, a) = \|a - \theta\|^2$, so that $R(\theta, \Delta) = \mathbb{E}_\theta[\|\Delta(X) - \theta\|^2]$.
- **Testing:** we test $\theta \in \mathcal{H}_0$ against $\theta \in \mathcal{H}_1$. In this case $\mathcal{A} = \{0, 1\}$ and

$$L(\theta, a) = \begin{cases} 1 & \text{if } \theta \in \mathcal{H}_0, a = 1 \\ 1 & \text{if } \theta \in \mathcal{H}_1, a = 0, \\ 0 & \text{otherwise.} \end{cases}$$

  The risk is then just the probability of the wrong decision:

$$R(\theta, \Delta) = \begin{cases} \mathbb{P}_\theta(\Delta(X) = 0) & \text{if } \theta \in \mathcal{H}_1, \\ \mathbb{P}_\theta(\Delta(X) = 1) & \text{if } \theta \in \mathcal{H}_0. \end{cases}$$

  These are the Type I/II error probabilities respectively.

10.2 Admissibility

Let's see how we might compare decision rules.

Definition 10.3. We say that Δ_2 ***strictly dominates*** Δ_1 if

$$R(\theta, \Delta_1) \geqslant R(\theta, \Delta_2) \ \forall \theta \in \Theta$$

and $R(\theta, \Delta_1) > R(\theta, \Delta_2)$ for at least some θ.

A procedure Δ_1 is ***inadmissible*** if there exists Δ_2 such that Δ_2 strictly dominates Δ_1.

We define ***admissible*** to simply mean *not inadmissible*.

Example. Suppose $X \sim \mathcal{U}[0, \theta]$. Let $\mathcal{D} = \{\text{estimators of the form } \widehat{\theta}(x) = ax\}$ (so this is a family indexed by a).

Using the quadratic loss, the risk will in general be

$$R(\theta, \widehat{\theta}) = \int_0^\theta (ax - \theta)^2 \cdot \frac{1}{\theta} \, \mathrm{d}x = \left(\frac{a^2}{3} - a + 1 \right) \theta^2$$

which is minimised at $a = 3/2$. Thus $\widehat{\theta}(x) = ax$ is *inadmissible* for all $a \neq 3/2$.

So $a = 3/2$ is a necessary condition for $\widehat{\theta}$ to be admissible for quadratic loss. Note that we have shown that $\widehat{\theta}(x) = \frac{3}{2}x$ is admissible in $\mathcal{D}$ but not among the set of all possible estimators of the form $\widehat{\theta}(x) = f(x)$ for some function f.

Remark. Note that being admissible is a fairly weak requirement. We will later see that some natural estimators are in fact inadmissible (see chapter 11).

10.3 Minimax rules and Bayes rules

We now explore notions of 'best possible' decision rules.

Definition 10.4. A rule Δ is a ***minimax rule*** if

$$\sup_\theta R(\theta, \Delta) \leqslant \sup_\theta R(\theta, \Delta') \ \forall \Delta' \in \mathcal{D}.$$

It minimises the maximum risk:

$$\Delta^* = \operatorname{argmin}_{\Delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \Delta).$$

Intuitively, a minimax rule does best in the worst case scenario. This can often still mean poor performance on average.

Given a prior belief π about the parameter θ, it is also natural to consider the average risk of a rule.

Definition 10.5. The ***Bayes integrated risk*** (or simply ***Bayes risk***)for a decision rule Δ and a prior $\pi(\theta)$ is

$$r(\pi, \Delta) := \int_\Theta R(\theta, \Delta) \pi(\theta) \, \mathrm{d}\theta.$$

A decision rule Δ is said to be a ***Bayes rule*** w.r.t. π if it minimises the Bayes risk:

$$r(\pi, \Delta) = \inf_{\Delta' \in \mathcal{D}} r(\pi, \Delta') =: r_\pi.$$

In case Θ is discrete the integral should be replaced by a sum.

Remark. Do not confuse Bayes rules with randomised decision rules.

We will see that Bayes rules (or estimators) provide a tool to solve minimax problems. To make this idea precise we need the notion of ***least favorable prior***. Recall that r_π is the Bayes risk of the Bayes estimator Δ_{Bayes} associated to π (when one exists).

Definition 10.6. A prior distribution π is least favorable if $r_\pi \geqslant r_{\pi'}$ for all prior distributions π'.

The following Theorem provides a simple condition for a Bayes estimator Δ_{Bayes} to be minimax.

Theorem 10.7. *Suppose that π is a prior distribution on Θ and that Δ_{Bayes} is the Bayes estimator for π with*

$$r(\pi, \Delta_{Bayes}) = r_\pi.$$

If the rule Δ_0 satisfies

$$\sup_\theta R(\theta, \Delta_0) \leqslant r_\pi$$

then Δ_0 is minimax, and, furthermore, if Δ_{Bayes} is the unique Bayes estimator for π then Δ_0 is the unique minimax procedure.

Proof. Let Δ be any other rule. Then

$$\sup_\theta R(\theta, \Delta) \geqslant \int R(\theta, \Delta)\pi(\theta)d\theta \tag{10.1}$$

$$\geqslant \int R(\theta, \Delta_{\text{Bayes}})\pi(\theta)d\theta \tag{10.2}$$

$$= r_\pi \geqslant \sup_\theta R(\theta, \Delta_0). \tag{10.3}$$

The second inequality is strict if there is a unique Bayes estimator which gives the second point. □

Remark. It is interesting to note that in the Theorem above one must have that $R(\theta, \Delta_0) = r_\pi$ for π-almost all θ. Indeed otherwise we would have

$$\int R(\theta, \Delta_0)\pi(\theta)\,\mathrm{d}\theta < r_\pi$$

which contradicts the definition of r_π.

Theorem 10.8. *Let Δ_{Bayes} be the Bayes estimator for some prior π. If*

$$R(\theta, \Delta_{Bayes}) \leqslant r_\pi \qquad \text{for all } \theta$$

then Δ_{Bayes} is minimax and π is a least favorable prior.

Proof. The first part is simply an application of Theorem 10.7.

Let π' be some other distribution. Then, writing Δ'_{Bayes} for the Bayes estimator with respect to π' we have

$$r_{\pi'} = \int R(\theta, \Delta'_{\text{Bayes}})\pi'(\theta)d\theta \leqslant \int R(\theta, \Delta_{\text{Bayes}})\pi'(\theta)d\theta \leqslant \sup_\theta R(\theta, \Delta_{\text{Bayes}}) = r_\pi.$$

□

The following Corollary is often very useful.

Corollary 10.9. *If a Bayes rule Δ_{Bayes} has constant Risk, then it is minimax.*

In fact the risk only needs to be constant ***almost everywhere***.

Corollary 10.10. *Let $\omega_\pi \subset \Theta$ be the set of θ at which the risk function of Δ_{Bayes} achieves its maximum, i.e.*

$$\omega_\pi = \{\theta : R(\theta, \Delta_{Bayes}) = sup_{\theta'} R(\theta', \Delta_{Bayes})\}.$$

If $\pi(\omega_\pi) = 1$. then Δ_{Bayes} is minimax.

Remark. Bayes and almost surely constant risk is sufficient for minimax but not necessary. The result is stated wrongly in Lehmann and Casella as an if and only if statement.

Example. Suppose that $X \sim \text{Bin}(n, p)$ and we wish to estimate p with the square error loss function $L(p, \widehat{p}) = (p - \widehat{p})^2$. We chose $\widehat{p} = \frac{X}{n}$. the risk function is

$$R(p, \widehat{p}) = \mathbb{E}_p[(p - \widehat{p})^2] = p(1-p)/n.$$

It has a unique maximizer at $p = 1/2$. So to apply the Corollary above we would need $\pi(1/2) = 1$ for which the corrresponding Bayes estimator is $\Delta = 1/2$, not X/n. In fact it can be checked that X/n is not minimax.

To determine a minimax estimator by the method suggested by Theorem 10.7 let us try a Beta(a, b) prior distribution. In that case we will see that the Bayes estimator is the posterior mean (this is proved in Proposition 10.15, but you can try to prove this for yourself here!) , i.e. $\Delta(x) = \frac{a+x}{a+b+n}$ and the risk function is

$$R(p, \Delta) = \frac{1}{(a+b+n)^2}\left\{np(1-p) + [a(1-p) - bp]^2\right\}.$$

Can we find values of a, b such that this risk function is constant? Setting the coefficients of p^2 and p to 0 show that $R(p, \Delta)$ is constant in p iff

$$(a+b)^2 = n \quad \text{and} \quad 2a(a+b) = n$$

and since a, b are positive we find $a = b = \frac{1}{2}\sqrt{n}$. It follows that the estimator

$$\Delta = \frac{X + \frac{1}{2}\sqrt{n}}{n + \sqrt{n}} = \frac{X}{n}\frac{\sqrt{n}}{1+\sqrt{n}} + \frac{1}{2}\frac{1}{1+\sqrt{n}}$$

is contant risk Bayes and hence minimax. Because of the uniqueness of the Bayes estimator we see that this is the unique minimax estimator.

10.4 Bayes rule and posterior risk

Definition 10.11. The ***expected posterior loss*** of a rule Δ w.r.t. a prior π is

$$\Lambda(x, \Delta) = \mathbb{E}\left[L[\theta, \Delta(x)] \mid X = x\right] = \int_\Theta L(\theta, \Delta(x))\pi(\theta \mid x)\,\mathrm{d}\theta.$$

The following result says that ***a Bayes rule minimises the expected posterior loss.***

Theorem 10.12. *Suppose that $X \mid \theta \sim P_\theta$ and that $\theta \sim \pi$. Suppose in addition that the following hypothesis hold for the problem of estimating $g(\theta)$ with non-negative loss function $L(\theta, d)$.*

(a) There exists an estimator (a rule) Δ_0 with finite risk.

(b) For almost all x, there exists a value $c(x)$ which minimizes

$$y \mapsto \Lambda(x, y)$$

Then $\Delta(x) = c(x)$ is a Bayes estimator.

Proof. The Bayes risk is (using $\pi(\theta \mid x) = f(\theta, x)\pi(\theta)/h(x)$ where $h(x)$ is the marginal distribution

of X)

$$r(\pi, \Delta) = \int R(\theta, \Delta)\pi(\theta)\,\mathrm{d}\theta = \int\int L(\theta, \Delta(x))f(\theta, x)\pi(\theta)\,\mathrm{d}x\,\mathrm{d}\theta \tag{10.4}$$

$$= \int\int L(\theta, \Delta(x))\pi(\theta \mid x)h(x)\,\mathrm{d}x\,\mathrm{d}\theta \tag{10.5}$$

$$= \int h(x)\int L(\theta, \Delta(x))\pi(\theta \mid x)\,\mathrm{d}\theta\,\mathrm{d}x \tag{10.6}$$

$$= \int h(x)\Lambda(x, \Delta(x))\,\mathrm{d}x \tag{10.7}$$

$$\leq \int h(x)\Lambda(x, \Delta'(x))\,\mathrm{d}x \tag{10.8}$$

for any other rule Δ'. □

Proposition 10.13 (Bayes rules and admissibility). *Let Δ^π be a Bayes rule w.r.t. π with finite Bayes risk. Then*

1. *If Δ^π is unique then it is admissible.*
2. *If $\theta \mapsto R(\theta, \Delta)$ is continuous for all Δ and π has a positive density w.r.t. the Lebesgue measure, then Δ^π is admissible.*

Proof.

1. If Δ^π is not admissible then there is some Δ such that $R(\theta, \Delta) \leqslant R(\theta, \Delta^\pi)\ \forall\theta \in \Theta$ and $R(\theta, \Delta) < R(\theta, \Delta^\pi)$ for some θ. This implies $r(\pi, \Delta) \leqslant r(\pi, \Delta^\pi)$, so Δ must also be Bayes, so by uniqueness $\Delta = \Delta^\pi$, contradicting the definition of Δ. So Δ^π is admissible.
2. As above, if Δ^π is not admissible then there is some Δ such that $R(\theta, \Delta) \leqslant R(\theta, \Delta^\pi)\ \forall\theta \in \Theta$ and $A_\Delta \neq \emptyset$, where $A_\Delta := \{\theta : R(\theta, \Delta) < R(\theta, \Delta^\pi)\}$.

   Since $\theta \mapsto R(\theta, \Delta) - R(\theta, \Delta^\pi)$ is continuous, A_Δ must contain an open set. So $\pi(A_\Delta) > 0$ arriving at a contradiction.

□

10.5 Point estimation

In the setting of point estimation (coming up with a best guess for a parameter, as we've been doing a lot in this course) there are three common loss functions:

Definition 10.14. The ***zero-one loss*** is of the form $L(\theta, \widehat{\theta}) = \begin{cases} a & \text{if } |\theta - \widehat{\theta}| > b, \\ 0 & \text{otherwise} \end{cases}$ where a, b are positive constants.

The ***absolute error loss*** is of the form $L(\theta, \widehat{\theta}) = k|\widehat{\theta} - \theta|$ where k is a positive constant.

The ***quadratic loss*** is of the form $L(\theta, \widehat{\theta}) = k(\widehat{\theta} - \theta)^2$ where k is a positive constant.

Let us see what the Bayes estimate (Bayes rule) is for each of these losses, by minimising the expected posterior loss.

Proposition 10.15. *The Bayes estimate under the:*

1. *zero-one loss with interval radius b tends to the posterior mode as $b \to 0$ (assuming say*

continuous posterior density);

2. *absolute error loss is the posterior median;*
3. *quadratic loss is the posterior mean.*

Proof.

1. The expected posterior loss is

$$\Lambda(x) = \int \pi(\theta \mid x) L(\theta, \widehat{\theta}) \, \mathrm{d}\theta \tag{10.9}$$

$$= a \int_{\widehat{\theta}+b}^{\infty} \pi(\theta \mid x) \, \mathrm{d}\theta + a \int_{-\infty}^{\widehat{\theta}-b} \pi(\theta \mid x) \, \mathrm{d}\theta \tag{10.10}$$

$$\propto 1 - \int_{\widehat{\theta}-b}^{\widehat{\theta}+b} \pi(\theta \mid x) \, \mathrm{d}\theta. \tag{10.11}$$

To find the Bayes rule one has to minimise the above or equivalently maximise

$$\int_{\widehat{\theta}-b}^{\widehat{\theta}+b} \pi(\theta \mid x) \, \mathrm{d}\theta.$$

Differentiating w.r.t. to $\widehat{\theta}$ and setting the derivative equal to 0 we need to find $\widehat{\theta}$ such that

$$\pi\left(\widehat{\theta} + b | x\right) - \pi\left(\widehat{\theta} - b | x\right) = 0. \tag{10.12}$$

If the map $\theta \mapsto \pi(\theta|x)$ is continuous, the above is guaranteed to have a solution for b small enough. To see why let $\tilde{\theta}$ be a mode of $\pi(\theta|x)$; then

$$\pi\left(\tilde{\theta}|x\right) - \pi\left(\tilde{\theta} - 2b|x\right) \geq 0, \quad \widehat{\theta} = \tilde{\theta} - b$$
$$\pi\left(\tilde{\theta} + 2b|x\right) - \pi\left(\tilde{\theta} - b|x\right) \leq 0, \quad \widehat{\theta} = \tilde{\theta} + b,$$

for b small enough, since $\tilde{\theta}$ is a local maximum. Therefore by the intermediate value theorem there must be a solution $\widehat{\theta}$ in the interval $[\tilde{\theta} - b, \tilde{\theta} + b]$. To verify this is indeed a maximum, one may also do a second derivative test if $\theta \mapsto \pi(\theta|x)$ is differentiable to obtain

$$\pi'(\widehat{\theta} + b|x) - \pi'(\widehat{\theta} - b|x) \leq 0,$$

again by the fact that $\widehat{\theta}$ is a local maximum and thus the derivative must change sign. So the Bayes rule is to choose $\widehat{\theta}(x)$ so that (10.12) is satisfied, which can be achieved by some $\widehat{\theta} \in [\tilde{\theta} - b, \tilde{\theta} + b]$. So as $b \to 0$, $\widehat{\theta}$ tends towards the posterior mode.

2. The expected posterior loss is

$$\Lambda(x) = \int |\widehat{\theta} - \theta| \pi(\theta \mid x) \, \mathrm{d}\theta = \int_{-\infty}^{\widehat{\theta}} (\widehat{\theta} - \theta) \pi(\theta \mid x) \, \mathrm{d}\theta + \int_{\widehat{\theta}}^{\infty} (\theta - \widehat{\theta}) \pi(\theta \mid x) \, \mathrm{d}\theta$$

so that

$$\frac{\partial}{\partial \widehat{\theta}} \Lambda(x) = \int_{-\infty}^{\widehat{\theta}} \pi(\theta \mid x) \, \mathrm{d}\theta - \int_{\widehat{\theta}}^{\infty} \pi(\theta \mid x) \, \mathrm{d}\theta = \frac{\partial}{\partial \widehat{\theta}} \Lambda(x) = 2 \int_{-\infty}^{\widehat{\theta}} \pi(\theta \mid x) \, \mathrm{d}\theta - 1$$

so, setting this to zero, Λ is minimised(indeed second derivative is non-negative) when

$$\int_{-\infty}^{\widehat{\theta}} \pi(\theta \mid x)\,\mathrm{d}\theta = \int_{\widehat{\theta}}^{\infty} \pi(\theta \mid x)\,\mathrm{d}\theta,$$

i.e. $\widehat{\theta}$ is the median of $\pi(\theta \mid x)$.

3. The expected posterior loss is

$$\Lambda(x) = \mathbb{E}[(\widehat{\theta} - \theta)^2 \mid X = x] \tag{10.13}$$

$$= \mathbb{E}[(\widehat{\theta} - \mu_x + \mu_x - \theta)^2 \mid X = x] \text{ where } \mu_x \text{ is the posterior mean} \tag{10.14}$$

$$= (\widehat{\theta}^2 - \mu_x)^2 + 2(\widehat{\theta} - \mu_x)\,\mathbb{E}[\theta - \mu_x \mid X = x] + \mathbb{E}[(\theta - \mu_x)^2 \mid X = x] \tag{10.15}$$

$$= (\widehat{\theta} - \mu_x)^2 + \mathrm{Var}(\theta \mid X = x). \tag{10.16}$$

So Λ is minimised when $\widehat{\theta} = \mu_x$, the posterior mean.

□

Example. (Example 4.1.5 in Lehmann Casella p230) Suppose $X \sim \mathrm{Bin}(n, p)$ with a $\mathrm{Beta}(a, b)$ prior for p. As we have seen this is a conjugate prior and the posterior density of p is proportional to $p^{x+a-1}(1-p)^{n-x+b-1}$. Therefore, under the quadratic loss function, the Bayes estimator $\widehat{p}_{\text{Bayes}}$ of p is

$$\widehat{p}_{\text{Bayes}} = \mathbb{E}[p \mid x] = \frac{a+x}{a+b+n}.$$

It is interesting to compare this with the MLE (or UMVE) which is just X/n. In fact, before taking any observation, the estimator from the Bayesian approach would be the mean of the prior $a/(a+b)$. Once X has been observed, the standard non-Bayesian estimator is X/n. The Bayes estimator $\widehat{p}_{\text{Bayes}}$ lies between the two and in fact

$$\widehat{p}_{\text{Bayes}} = \left(\frac{a+b}{a+b+n}\right)\frac{a}{a+b} + \left(\frac{n}{a+b+n}\right)\frac{X}{n}$$

it is a weighted average of the two.

It is instructive to examine the cases:

1. $n \to \infty$ while a, b fixed
2. n fixed and $a, b \to \infty$ with a/b fixed.

10.6 Finite decision problems

Definition 10.16. A decision problem is said to be finite when Θ is finite. We write $\Theta = \{\theta_1, \ldots, \theta_k\}$.

In the case of a finite decision problem, the notions of admissibility, minimax and Bayes rules can be given geometric interpretations.

We now assume that the set of decision rules is also finite and contains l decision rules $\Delta_1, \ldots, \Delta_l$ and randomised decision rules that can be formed by their convex combinations, that is the decision set is the convex hull of $\{\Delta_1, \ldots, \Delta_l\}$.

$$\mathcal{D} := \left\{ \sum_{i=1}^{l} p_i \Delta_i : p_i \geq 0, \sum p_i = 1 \right\}.$$

Definition 10.17. The ***risk set*** $S \subseteq \mathbb{R}^k$ is the set of points $\{(R(\theta_1, \Delta), \ldots, R(\theta_k, \Delta)) : \Delta \in \mathcal{D}\}$.

It may also be the case that $\mathcal{D}$ is the convex hull of an infinite collection (or continuum) of non-randomized decision rules, see e.g. Figure 10.2.

Lemma 10.18. *S is a convex set.*

Proof. Let $\Delta_1, \Delta_2 \in \mathcal{D}$ be two rules. Take $\alpha \in (0, 1)$. Then define a randomized rule as follows:

$$\Delta'(x) = \begin{cases} \Delta_1(x) & \text{with prob } \alpha, \\ \Delta_2(x) & \text{with prob } 1 - \alpha. \end{cases}$$

Then $R(\theta, \Delta') = \alpha R(\theta, \Delta_1) + (1-\alpha) R(\theta, \Delta_2)$. So the convex combination is a valid decision rule. □

10.6.1 The case $k = 2$

The two-dimensional case (i.e. there are two possible parameters) is particularly interesting.

In Figure 10.1 we can see an example of a risk set when $\Theta = \{\theta_1, \theta_2\}$. The extreme points of the risk set are the ***deterministic rules***.

The thick line at the bottom defines the set of admissible rules. To see why, recall that a rule Δ is admissible if there is no other rule Δ' such that $R(\theta, \Delta') \leq R(\theta, \Delta)$ for all θ with strict inequality for at least one θ. In our scenario this means that a rule Δ is admissible if no other rule Δ' achieves a risk in the interior of the box

$$\left\{x \leq R(\theta_1, \Delta),\, y \leq R(\theta_2, \Delta)\right\}.$$

Note also that the minimax rule lies on the line $R_1 = R_2$; since the risk set intersects the line $R_1 = R_2$ this must be the case. Suppose that an admissible rule has constant risk, that is (R_1, R_2) with $R_1 = R_2$. Let (R_1', R_2') be the risk of any other admissible rule; it must satisfy $R_1' \geq R_1$ and $R_2' \leq R_2$ or $R_1' \leq R_1$ and $R_2' \geq R_2$. In either case we have that $\max\{R_1, R_2\} \leq \max\{R_1', R_2'\}$. The situation is similar in Figure 10.2.

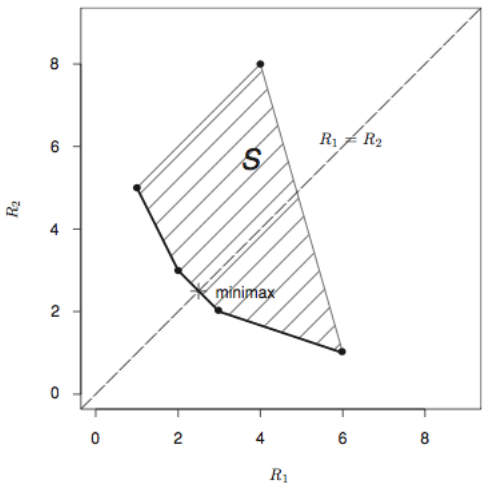

Figure 10.1: Riskset convex hull of five non-randomized rules. Minimax is randomized.

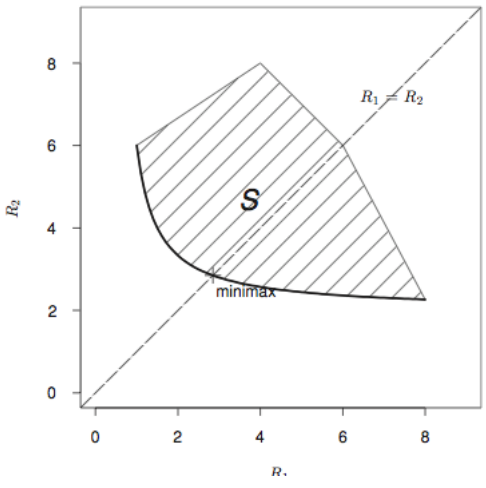

Figure 10.2: Riskset convex hull of continuum of non-randomized rules. Minimal non-randomized.

However in Figure 10.2 notice that the set of admissible rules consists of deterministic rules (in this case $\mathcal{D}$ cannot be expressed as the convex hull of a finite number of deterministic rules).

For Bayes rules, suppose that (π_1, π_2) is the prior. Then the lines $\pi_1 R_1 + \pi_2 R_2 = c$ represent decision

rules with the same Bayes risk c. We can see this in Figures 10.3,10.4. The slope of the lines are determined by the prior. We increase c looking for the line that just touches the risk set–this gives the Bayes rule.

In Figure 10.3 the Bayes rule is not unique, and the minimax is actually a non-randomized Bayes rule. In Figure 10.3 the Bayes rule is unique, and non-randomized. The minimax is actually a randomized rule distinct from the Bayes rule.

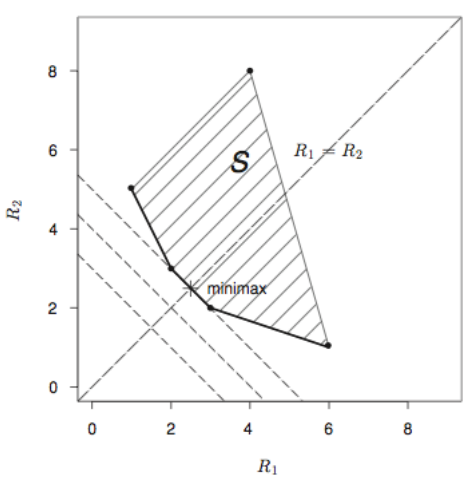

Figure 10.3: The minimax is a randomized Bayes rule; Bayes rule not unique. In fact there are non-randomized Bayes rules.

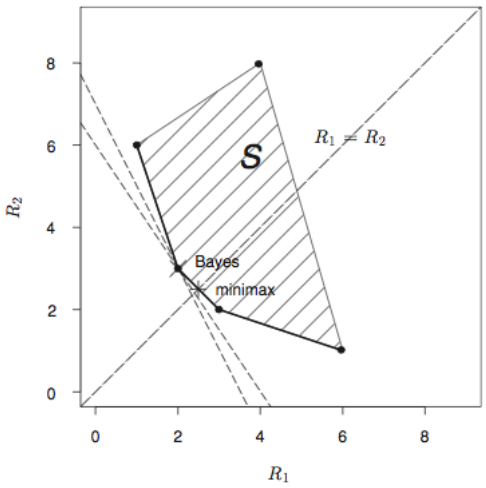

Figure 10.4: The minimax is no longer a Bayes rule. Minimax is still randomized although Bayes rule is non-randomized.

Chapter 11

The James-Stein Estimator

This chapter explores some rather counter-intuitive situations that can occur when one simultaneously estimates several parameters.

Assume that $X_i \sim \mathcal{N}(\mu_i, 1)$ are mutually independent unit-variance Gaussian random variables, and write $X = (X_1, \ldots, X_p)$ and $\mu = (\mu_1, \ldots, \mu_p)$. The goal is to estimate μ *from a single observation* X.

We know the maximum likelihood estimate is $\widehat{\mu}_{\mathrm{MLE}} = X$, and we have seen that this is the MVUE.

Is this estimate admissible (for, say, quadratic loss)? **For $p \geqslant 3$, the answer is *no*!**

Theorem 11.1 (Stein's Paradox). *The* **James-Stein estimator**

$$\widehat{\mu}_{\mathrm{JSE}} \coloneqq \left(1 - \frac{p-2}{\sum_{i=1}^{p} X_i^2}\right) X$$

strictly dominates $\widehat{\mu}_{\mathrm{MLE}}$ for quadratic loss.

(We will prove this shortly.)

Corollary 11.2. *If $p \geqslant 3$, $\widehat{\mu}_{\mathrm{MLE}}$ is inadmissible for quadratic loss.*

Remark. This is *very surprising*! For instance, suppose you take measurements to estimate:

1. The average weight K of a kiwi at Tesco;
2. The average height G of a blade of grass in University Parks;
3. The average speed S of a bike going down Cornmarket Street.

These are totally unrelated quantities; but Stein's paradox tells us that we get better estimates (on average) for the vector (K, G, S) by simultaneously using the three measurements![1]

Let's see how to prove this.

Lemma 11.3 (Stein's Lemma). *For independent Gaussian random variables $X = (X_1, \ldots, X_p)$*

[1] Or does it? Stein's paradox is about normal random variables. This can be justified by using the CLT.

with $X_i \sim \mathcal{N}(\mu_i, 1)$ for each i, then for each i and for any bounded differentiable function h,

$$\mathbb{E}[(X_i - \mu_i)h(X)] = \mathbb{E}\left[\frac{\partial h(X)}{\partial X_i}\right].$$

Proof. By the Tower Law,

$$\mathbb{E}[(X_i - \mu_i)h(X)] = \mathbb{E}\left[\mathbb{E}[(X_i - \mu_i)h(X) \mid \{X_j : j \neq i\}]\right].$$

Using integration by parts,

(11.1)
$$\mathbb{E}[(X_i - \mu_i)h(X) \mid \{X_j : j \neq i\}] = \int_{-\infty}^{\infty} (x_i - \mu_i)h(x)\mathrm{e}^{-(x_i-\mu_i)^2/2}\,\mathrm{d}x_i$$

(11.2)
$$= \left[-\mathrm{e}^{-(x_i-\mu_i)^2/2}h(x)\right]_{x_i=-\infty}^{x_i=\infty} + \int_{-\infty}^{\infty} \frac{\partial h(x)}{\partial x_i}\,\mathrm{e}^{-(x_i-\mu_i)^2/2}\,\mathrm{d}x_i$$

(11.3)
$$= 0 + \mathbb{E}\left[\frac{\partial h(X)}{\partial X_i} \mid X_j : j \neq i\right]$$

since h is bounded. Applying the tower property of conditional expectations again gives the result. □

Proof of Stein's Paradox. Consider the family of estimators $\widehat{\mu}_{\mathrm{JSE}} = \left(1 - \frac{a}{\sum X_i^2}\right) X$ indexed by the parameter a. These are called the ***James-Stein estimators***.

Recalling that $\widehat{\mu}_{\mathrm{MLE}} = X$, we get

$$R(\mu, \widehat{\mu}_{\mathrm{MLE}}) = \sum_{i=1}^{p} \mathbb{E}[(\mu_i - X_i)^2] = p$$

(since $\mathrm{Var}(X_i) = 1$).

On the other hand, writing $\widehat{\mu}_i := \left(1 - \frac{a}{\sum_j X_j^2}\right) X_i$,

(11.4)
$$R(\mu, \widehat{\mu}_{\mathrm{JSE}}) = \sum_{i=1}^{p} \mathbb{E}[(\mu_i - \widehat{\mu}_i)^2]$$

(11.5)
$$= \sum_{i=1}^{p} \left(\mathbb{E}[(\mu_i - X_i)^2] - 2a\,\mathbb{E}\left[\frac{(X_i - \mu_i)X_i}{\sum_j X_j^2}\right] + a^2\,\mathbb{E}\left[\frac{X_i^2}{\left(\sum_j X_j^2\right)^2}\right] \right).$$

Now the first term is just 1, since $\mathrm{Var}(X_i) = 1$, and by Stein's Lemma,

$$\mathbb{E}\left[\frac{(X_i - \mu_i)X_i}{\sum_j X_j^2}\right] = \mathbb{E}\left[\frac{\partial}{\partial X_i}\frac{X_i}{\sum_j X_j^2}\right] = \mathbb{E}\left[\frac{\sum_j X_j^2 - 2X_i^2}{\left(\sum_j X_j^2\right)^2}\right] = \mathbb{E}\left[\frac{1}{\sum_j X_j^2} - 2\frac{X_i^2}{\left(\sum_j X_j^2\right)^2}\right].$$

Putting this all together, we get

$$R(\mu, \widehat{\mu}_{\text{JSE}}) = p - (2ap - 4a)\,\mathbb{E}\left[\frac{1}{\sum X_j^2}\right] + a^2\,\mathbb{E}\left[\frac{1}{\sum X_j^2}\right] \tag{11.6}$$

$$= p - (2a(p-2) - a^2)\,\mathbb{E}\left[\frac{1}{\sum X_j^2}\right]. \tag{11.7}$$

This is minimised at $a = p - 2$, and is less than p for this value; this concludes the proof. □

Remark. The James-Stein estimator shrinks each component of X towards the origin. However, there is of course nothing special about the origin; a similar estimator $\widehat{\mu}_{\text{JSE}}^{(\mu_0)} = \mu_0 + \left(1 - \frac{p-2}{||X-\mu_0||^2}\right)(X - \mu_0)$ can be defined which shrinks X towards an arbitrary point μ_0, and it can easily be shown that this also strictly dominates $\widehat{\mu}_{\text{MLE}}$. (See the handwritten notes for the details.)

Exercise. Show that for some a the estimator $\bar{X}\mathbf{1}_p + \left(1 - \frac{a}{||X-\bar{X}\mathbf{1}_p||^2}\right)(X - \bar{X}\mathbf{1}_p)$ strictly dominates $\widehat{\mu}_{\text{JSE}}$, where $\mathbf{1}_p = (1, \ldots, 1)$.

Remark. Observe that when $||X - \mu_0||^2 < p - 2$, the shinkage factor becomes negative. To avoid this problem, we can define

$$\widehat{\mu}_{\text{JSE}+}^{(\mu_0)} = \mu_0 + \left(1 - \frac{p-2}{||X-\mu_0||^2}\right)^+ (X - \mu_0)$$

(where x^+ denotes the positive part), which strictly dominates $\widehat{\mu}_{\text{JSE}}^{(\mu_0)}$).

It is worth noting that neither $\widehat{\mu}_{\text{JSE}}^{(\mu_0)}$ nor $\widehat{\mu}_{\text{JSE}+}^{(\mu_0)}$ are admissible.

Example (Baseball example). Consider the dataset in fig. 11.1, taken from Young and Smith. It shows statistics from the 1998 baseball pre-season in the US for 17 top players. Our interest is in predicting the home run strike rate of each player in the full season.

For each player i, Y_i is the number of home runs out of n_i times at bat in the pre-season. We assume that home runs occur according to a binomial distribution, so that player i has probability p_i of hitting a home run each time at bat, independently of other at bats and other players. Thus $Y_i \sim \text{Bin}(n_i, p_i)$.

Here p_i is the true full-season strike rate (and Y_i/n is the strike rate in the pre-season); the actual values of p_i as well as the actual number AB_i of at bats of each player (in the full season) and the actual number of home runs HR_i are shown in the figure.

So, how might we estimate p_i given just the pre-season statistics Y_i and n_i for each player? Obviously the naÃ¯ve estimate is the MLE $\widehat{p}_i = Y_i/n_i$. These give rise to the estimated number of home runs $\widehat{HR}_i = \widehat{p}_i \cdot AB_i$ (assuming we know the actual number of at bats, which of course at the time we wouldn't have). These values are shown in the figure.

The Stein paradox tells us we may be able to do better.

First transform the data, setting $X_i = f_{n_i}(Y_i/n_i)$ where $f_n(y) := n^{1/2}\sin^{-1}(2y - 1)$. Then $X_i \sim \mathcal{N}(\mu_i, 1)$ for each i, with $\mu_i = f_{n_i}(p_i)$.

We can then use the James-Stein estimator to estimate the means μ_i. Using the 'improved version' we just encountered, we set

$$JS_i := \bar{X} + \left(1 - \frac{p-3}{V}\right)(X_i - \bar{X})$$

for each i, where $\bar{X} = \sum X_i/p$ and $V = \sum(X_i - \bar{X})^2$ (here $p = 17$).

	Y_i	n_i	p_i	AB	X_i	JS_i	μ_i	HR	$\hat{HR}$	$\hat{HR}_s$
McGwire	7	58	0.138	509	−6.56	−7.12	−6.18	70	61	50
Sosa	9	59	0.103	643	−5.90	−6.71	−7.06	66	98	75
Griffey	4	74	0.089	633	−9.48	−8.95	−8.32	56	34	43
Castilla	7	84	0.071	645	−9.03	−8.67	−9.44	46	54	61
Gonzalez	3	69	0.074	606	−9.56	−9.01	−8.46	45	26	35
Galaragga	6	63	0.079	555	−7.49	−7.71	−7.94	44	53	48
Palmeiro	2	60	0.070	619	−9.32	−8.86	−8.04	43	21	28
Vaughn	10	54	0.066	609	−5.01	−6.15	−7.73	40	113	78
Bonds	2	53	0.067	552	−8.59	−8.40	−7.62	37	21	24
Bagwell	2	60	0.063	540	−9.32	−8.86	−8.23	34	18	24
Piazza	4	66	0.057	561	−8.72	−8.48	−8.84	32	34	38
Thome	3	66	0.068	440	−9.27	−8.83	−8.47	30	20	25
Thomas	2	72	0.050	585	−10.49	−9.59	−9.52	29	16	28
T. Martinez	5	64	0.053	531	−8.03	−8.05	−8.86	28	41	41
Walker	3	42	0.051	454	−6.67	−7.19	−7.24	23	32	24
Burks	2	38	0.042	504	−6.83	−7.29	−7.15	21	27	19
Buhner	6	58	0.062	244	−6.98	−7.38	−8.15	15	25	21

Figure 11.1: *Data for 17 players in the 1998 baseball pre-season and full season taken from Young and Smith.*

These estimates of the μ_i are shown in the figure, and transforming back will give us estimates $\widehat{HR}_s$ for the number of home runs of each player, which are also shown.

We see that the James-Stein approach gives much better estimates on average! More precisely, the James-Stein estimator achieves a lower aggrigate risk than the naÃ¯ve estimator, but allows increased risk in estimation of individual components.

Chapter 12

Empirical Bayes Methods

We return now to our discussion of Bayes estimators (Bayes rules). While Bayes estimators have desirable properties (the posterior mean, the Bayes estimator under quadratic loss, is often admissible), they can be hard to calculate, in particular for the hierarchical models met in chapter 9.

This motivates the empirical Bayes approach.

12.1 Basic setup

Recall that a hierarchical Bayesian model consists of three 'layers': the likelihood $X \sim f(x, \theta)$ parametrised by θ, the prior $\theta \sim \pi(\theta, \psi)$ parametrised by ψ, and the hyperprior $\psi \sim g(\psi)$.

Definition 12.1. ***Empirical Bayes*** methods adapt the hierarchical Bayesian model by replacing the hyperparameter vector ψ with a point-estimate $\widehat{\psi}$ derived from the data.

So we now just have the likelihood $X \sim f(x, \theta)$ and the prior $\theta \sim \widehat{\psi}(\theta) = \pi(\theta, \widehat{\psi})$.

Remark. Empirical Bayes methods can be viewed as an approximation of a full hierarchical Bayes model that allows us to avoid doing ψ-integrals. One layer of the hierarchy has been 'chopped off'.

Recall that we met this idea briefly in chapter 9 before hierarchical models were introduced.

The reduced model has posterior

$$\widehat{\pi}(\theta \mid x) \propto L(\theta, x)\pi(\theta, \widehat{\psi})$$

and a ***Bayes estimator*** $\widehat{\theta}_{\text{EB}}$ can be calculated using $\widehat{\pi}(\theta \mid x)$. So for quadratic loss, we have $\widehat{\theta}_{\text{EB}} = \int \theta \widehat{\pi}(\theta \mid x) \, \mathrm{d}\theta$, the posterior mean.

Remark. In this setting, the Bayes estimator is called an ***empirical Bayes estimator***, or an ***EB estimator***.

12.2 Choice of point estimate

How can we choose our point estimate $\widehat{\psi}$ of the hyperparameter? We have all the classical frequentist techniques at our disposal. The two most obvious ways are:

- Use the MLE $\widehat{\psi} = \operatorname{argmax}_{\psi} p(x \mid \psi)$ where

$$p(x \mid \psi) = \int L(\theta, x)\pi(\theta, \psi) \, \mathrm{d}\theta$$

  is the marginal likelihood.

Previous experiments:

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

Current experiment:
4/14

Figure 12.1: *Data on tumor incidence in historical control groups and current group of rats, from Tarone 1982. The table displays the values y_j/n_j: (number of rats with tumors)/(total number of rats).*

- Use the method of moments: choose $\widehat{\psi}$ such that $\pi(\theta, \widehat{\psi})$ has the same mean and variance as the *sample mean* and *sample variance* of the MLEs of the θ_i.

Example (Meta-analysis of studies of tumors in rodents). The data in fig. 12.1 shows the number of rats with tumors, Y_i, and the total number of rats n_i in each of a number of previous experiments on tumor growth, as well as the results of a new experiment which we are interested in analysing.

As usual we'll assume each $Y_i \sim \text{Bin}(n_i, \theta_i)$ independently, for parameters θ_i which we want to estimate. As our prior distribution we assume that $\theta_i \sim \text{Beta}(\alpha, \beta)$ independently for each i, where α, β are hyperparameters. This choice of prior is natural as it is conjugate for the binomial distribution: the posterior distribution, after observing the new experiment (14 rats, 4 with tumors) will be $\pi(\theta \mid y) = \text{Beta}(\alpha + 4, \beta + 10)$.

Using an empirical Bayes approach with the method of moments goes as follows:

1. Compute the MLEs Y_i/n_i for the previous experiments $i = 1, \ldots, 70$.
2. Compute the sample mean and variance of these MLEs: $m = 0.136$ and $v = 0.0106$.
3. Pick $\widehat{\alpha}, \widehat{\beta}$ such that $\text{Beta}(\widehat{\alpha}, \widehat{\beta})$ has 'matched moments', i.e.

$$\frac{\widehat{\alpha}}{\widehat{\alpha} + \widehat{\beta}} = m, \quad \frac{\widehat{\alpha}\widehat{\beta}}{(\widehat{\alpha} + \widehat{\beta})^2(\widehat{\alpha} + \widehat{\beta} + 1)} = v.$$

   This solves to $\widehat{\alpha} = 1.4, \widehat{\beta} = 8.6$.
4. Calculate the Bayes estimate, which for the quadratic loss is the posterior mean. In this case the posterior is $\widehat{\pi}(\theta \mid y) = \text{Beta}(5.4, 18.6)$ so the posterior mean is 0.225

This estimate is less than the maximum-likelihood estimate of $\widehat{\theta}_{\text{MLE}} = 4/14$ we'd get based solely on the current experiment, not taking into account past experiments.

12.3 James-Stein and empirical Bayes

Suppose we have $X_1, \ldots, X_p \sim \mathcal{N}(\theta_i, 1)$ as in the setup for the James-Stein estimator. Given one observation x_i per parameter θ_i we wish to estimate the parameters θ_i.

Proposition 12.2. *The James-Stein estimator can be intepreted as an empirical Bayes estimator.*

(Specifically, for $a = p$ it's the EB estimator for quadratic loss when using a mean-zero Gaussian prior whose variance is estimated using maximum likelihood.)

Proof. We wish to construct an EB estimator for quadratic loss. There is some freedom of choice of prior, but we will assume as our prior that θ_i are drawn independently from a $\mathcal{N}(0, \tau^2)$ distribution.

Given τ, then, we have $\theta_i \mid (x_i, \tau^2) \sim \mathcal{N}\left(x_i \frac{\tau^2}{1+\tau^2}, \frac{\tau^2}{1+\tau^2}\right)$. This can be calculated by completing the square.

To estimate τ, then, we can compute the marginal likelihood of X_i given τ:

$$X_i \mid \tau^2 \sim \mathcal{N}(0, \tau^2 + 1) \text{ independently for each } i.$$

This is maximised by $\widehat{\tau}^2 = \frac{1}{p}\sum_{j=1}^{p}(X_j^2 - 1)$. (This is from the standard result for the MLE for the variance of a Gaussian distribution).

So the estimated posterior distribution is $\theta_i \mid x_i \sim \mathcal{N}\left(x_i \frac{\widehat{\tau}^2}{1+\widehat{\tau}^2}, \frac{\widehat{\tau}^2}{1+\widehat{\tau}^2}\right)$. Thus the Bayes estimator for quadratic loss, i.e. the posterior mean, is

$$\widehat{\theta}_{\mathrm{EB},i} = X_i \frac{\widehat{\tau}^2}{1+\widehat{\tau^2}} = X_i \frac{\left(\frac{1}{p}\sum_{j=1}^{p} X_j^2\right) - 1}{\frac{1}{p}\sum_{j=1}^{p} X_j^2} = X_i \left(1 - \frac{p}{\sum X_j^2}\right).$$

This is the James-Stein estimator with $a = p$. □

Remark. This is not the minimum James-Stein estimator (with $a = p - 2$) but it does strictly dominate the MLE for all θ. The James-Stein estimator with $a = p - 2$ can be recovered by using moment estimators (see Young and Smith section 3.5).

Example. Suppose that $X_i \sim \mathrm{Po}(\theta_i)$ independently for $i = 1, \ldots, p$.

The maximum-likelihood estimate for each θ_i would be simply x_i. Let's follow roughly the same empirical Bayes approach as above to find a better estimator (similar to the James-Stein estimator).

As a prior we assume that θ_i are i.i.d. $\mathrm{Exp}(\lambda)$, so that $\pi(\theta_i \mid \lambda) = \lambda \mathrm{e}^{-\lambda\theta_i}$ for each i and λ is a hyperparameter to be estimated.

The marginal likelihood for λ is, for a single data point i,

$$p(x_i \mid \lambda) = \int_0^\infty \frac{\mathrm{e}^{-\theta_i}\theta_i^{x_i}}{x_i!}\lambda \mathrm{e}^{-\lambda\theta_i}\,\mathrm{d}\theta_i = \left(\frac{1}{1+\lambda}\right)^{x_i}\frac{\lambda}{1+\lambda} \sim \mathrm{Geom}\frac{\lambda}{1+\lambda}.$$

So given λ the X_i are marginally i.i.d. $\mathrm{Geom}\left(\frac{\lambda}{1+\lambda}\right)$ with mean λ^{-1}.

So the maximum marginal likelihood estimator is $\widehat{\lambda} = \frac{1}{\bar{x}} = \frac{n}{\sum x_i}$.

Hence our empirical Bayes approximation gives marginal posterior

$$\widehat{\pi}(\theta \mid x) \propto L(\theta, x)\pi(\theta, \widehat{\lambda}) = \prod_{i=1}^{p} \mathrm{e}^{-\theta_i}\theta_i^{x_i}\widehat{\lambda}\mathrm{e}^{-\widehat{\lambda}\theta_i}.$$

We recognise from this expression that $\theta_i \mid x_i \sim \Gamma(x_i + 1, \widehat{\lambda} + 1)$ for each i. So the EB estimator is

the approximated posterior mean,

$$\widehat{\theta}_{\mathrm{EB},i} = \frac{\alpha}{\beta} = \frac{x_i + 1}{\widehat{\lambda} + 1} = \bar{x}\frac{x_i + 1}{\bar{x} + 1} = \bar{x}\frac{1}{\bar{x} + 1} + x_i\frac{\bar{x}}{\bar{x} + 1}.$$

This has the effect of shrinking the MLE estimates towards the mean $\bar{x}$.

Remark. We see that the empirical Bayes approach tends to pull the estimates towards the common mean. This is true in general for models with exchangeable parameters.

Note also that, as mentioned in chapter 9, one drawback of the empirical Bayes approach is that we're potentially using the same data twice, leading to overfitting.

12.4 Non-parametric empirical Bayes

So far we have estimated a hyperprior distribution by finding a point estimate for the hyperparameter. We could instead estimate the hyperprior (or marginal) distribution *directly* from the data. This is known as ***non-parametric empirical Bayes***. One such method is illustrated below.

Example. Suppose $Y_i \sim \mathrm{Po}(\theta_i)$ independently. Assume that the parameters θ_i are drawn independently from some distribution π whose form we do not know.

The posterior mean is

$$\widehat{\theta}_i = \mathbb{E}[\theta_i \mid Y_i] = \int \theta\pi(\theta \mid Y_i)\,\mathrm{d}\theta \tag{12.1}$$

$$= \frac{\int \left(\frac{\theta^{Y_i+1}\mathrm{e}^{-\theta}}{Y_i!}\right)\pi(\theta)\,\mathrm{d}\theta}{\int \left(\frac{\theta^{Y_i}\mathrm{e}^{-\theta}}{Y_i!}\right)\pi(\theta)\,\mathrm{d}\theta} \text{ by Bayes' Theorem} \tag{12.2}$$

$$= \frac{(Y_i + 1)p(Y_i + 1)}{p(Y_i)} \tag{12.3}$$

where $p(y)$ is the marginal pmf.

Robbin's method is then to approximate the marginal pmf $p(y)$ by the actual number of observed datapoints equal to y. So in this case

$$\widehat{\theta}_i = \frac{(y_i + 1)\widehat{p}(y_i + 1)}{\widehat{p}(y_i)} = \frac{(y_i + 1) \cdot |\{j : y_j = y_i + 1\}|}{|\{j : y_j = y_i\}|}.$$

Chapter 13

Hypothesis Tests

13.1 Recap from part A

13.1.1 General setup

Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta)$ where $\theta \in \Theta$ is a scalar or vector parameter. Suppose we are interested in testing

- The null hypotehsis $H_0 : \theta \in \Theta_0$
- against the alternative $H_1 : \theta \in \Theta_1$.

Unless specified otherwise we assume that $\Theta_0 \cap \Theta_1 = \emptyset$

If a hypothesis consists of a single point in Θ so that $\Theta_0 = \{\theta_0\}$ say, we say that it is a ***simple*** hypotehsis. Otherwise it is called a ***composite*** hypothesis.

In general a test consists of a ***critical region*** C such that we reject H_0 if and only if $X \in C$. We reformulate this slightly by introducing the concept of the ***test function*** $\phi : \mathcal{X} \mapsto \{0, 1\}$

$$\phi(x) = \begin{cases} 1 \text{ if } x \in C \\ 0 \text{ if } x \notin C \end{cases}$$

We will sometimes simply say ***the test*** ϕ. We will also sometimes need the notion of a randomized test. Suppose that $\mathcal{X} = C_1 \cup C_0 \cup C_=$ where $C_1, C_0, C_=$ are pairwise disjoint. Fix $\gamma \in [0, 1]$. Then the we generalize the notion of test function by saying that

$$\phi(x) = \begin{cases} 1 \text{ if } x \in C_1 \\ \gamma \text{ if } x \in C_= \\ 0 \text{ if } x \in C_0 \end{cases}$$

is the test where we ***reject*** H_0 when $x \in C_1$, ***accept*** H_0 when $x \in C_0$, and ***reject*** H_0 ***with probability*** γ if $x \in C_=$ (by flipping a coin). Such a test ϕ is called a ***randomized test***.

Definition 13.1.

- The ***power function*** of a test is defined to be

$$w(\theta) = \mathbb{P}_\theta(\text{Reject} H_0) = \mathbb{E}_\theta[\phi(X)].$$

- The ***size*** of a test is often denoted α and is defined to be

$$\alpha := \sup_{\theta \in \Theta_0} w(\theta).$$

The idea is this: a good test has a small size so that $\alpha \leqslant \alpha_0$ for some specified value α_0 and makes $w(\theta)$ as large as possible on Θ_1. Within this framework we can consider various classes of problems:

1. Simple H_0 vs simple H_1: here there is an elegant and complete theory which tells us exactly how to construct the best test given by Neyman-Pearson Theorem.
2. Simple H_0 vs composite H_1: In this case the obvious approach is to pick $\theta_1 \in \Theta_1$ and construct the Neyman-Pearson test of H_0 against H_1. In some cases, the critical region one obtains is the same for all θ_1. When that happens the test is said to be ***uniformly most powerful*** (or UMP). But there are many situations in which UMP tests do not exists, and then the problem is harder.
3. Composite H_0 vs composite H_1: In this case the problem is harder again.

13.1.2 Neyman-Pearson Theorem

Consider a test of a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$. Define the ***likelihood ratio***:

$$\Lambda(x) = \frac{f(x, \theta_1)}{f(x, \theta_0)}.$$

Theorem 13.2. *Define the critical region*

$$C = \{x : \Lambda(x) \geqslant k\}$$

and suppose that the constants k and α are such that $\mathbb{P}_{\theta_0}(X \in C) = \alpha$. Then among all tests of H_0 against H_1 of size α, the test with critical region C has **maximum power**.

The tests with critical regions such as C are called ***Neyman-Pearson test*** or ***likelihood ratio test*** (LRT).

13.1.3 Uniformly most powerful tests

Definition 13.3. A ***uniformly most powerful test*** or UMP test of size α is a test function ϕ_0 such that

1. $\mathbb{E}_\theta(\phi_0(X)) \leqslant \alpha$ for all $\theta \in \Theta_0$,
2. Given any other test ϕ for which $\mathbb{E}_\theta(\phi(X)) \leqslant \alpha$ for all $\theta \in \Theta_0$, we have $\mathbb{E}_\theta(\phi_0(X)) \geqslant \mathbb{E}_\theta(\phi(X))$ for all $\theta \in \Theta_1$.

Note that a UMP test does not necessarily exist. However, for one-sided testing problems involving a single parameter there is a wide class of parametric families that have a UMP test.

Definition 13.4. A family of densities $\{f(x, \theta), \theta \in \Theta \subseteq \mathbb{R}\}$ with real scalar parameter θ is said to be of ***monotone likelihood ratio*** or MLR for short if there exists a function $t(x)$ such that the likelihood ratio

$$x \mapsto \frac{f(x, \theta_2)}{f(x, \theta_1)}$$

is a non-decreasing function of $t(x)$ whenever $\theta_1 \leqslant \theta_2$.

Theorem 13.5. *Suppose that X has a distribution from a family which is MLR with respect to a statistic $t(X)$ and that we wish to test $H_0 : \theta \leqslant \theta_0$ against $H_1 : \theta > \theta_0$. Suppose that the distribution of $t(X)$ is continuous. Then*

1. *The test with critical region*
$$C = \{x : t(x) > t_0\}$$
is UMP among all test of size at most $\mathbb{P}_{\theta_0}(X \in C)$.
2. *Given α, there exists some t_0 such that the test above has size α.*

Proof. For any $\theta_1 > \theta_0$ the Neyman-Pearson test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ has a critical region of the form $C = \{x : t(x) > t_0\}$ for some t_0 which is chosen so that $\mathbb{P}_{\theta_0}(T(X) > t_0) = \alpha$. Note that t_0 does not depend on θ_1 and so the critical region C is the same for all values of θ_1. Thus, we see that this test is UMP for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

Next, we claim that for any critical region of the form $C = \{x : t(x) > t_0\}$ the map
$$\theta \mapsto \mathbb{P}_\theta(X \in C)$$
is non-decreasing. This can be seen using a argument involving randomized test procedures and the optimality of the LRT (see Young and Smith p72).

It follows that if $\mathbb{P}_{\theta_0}(X \in C) = \alpha$ then $\sup_{\theta \leqslant \theta_0} \mathbb{P}_\theta(X \in C) \leqslant \alpha$. Suppose that C' is another critical region such that $\sup_{\theta \leqslant \theta_0} \mathbb{P}_\theta(X \in C') \leqslant \alpha$ as well. This implies trivially that $\mathbb{P}_{\theta_0}(X \in C') \leqslant \alpha$ and thus by optimality of the LRT that for all $\theta_1 > \theta_0$ we have
$$\mathbb{P}_{\theta_1}(X \in C') \leqslant \mathbb{P}_{\theta_1}(X \in C)$$
This shows that C is UMP among all tests of its size.

The second statement in the Theorem is clear by continuity. □

Example. Suppose the $X_1, \ldots, X_n$ are i.i.d. from an exponential distribution with mean θ : $f(x, \theta) = \theta^{-1} e^{-x/\theta}, x > 0$ where $\theta \in (0, \infty)$. Let us say that we frist want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ (this is a simple-composite test).

Consider $\theta_1 > \theta_0$. We have
$$\Lambda(x) = \frac{f(x, \theta_1)}{f(x, \theta_0)} = \left(\frac{\theta_0}{\theta_1}\right)^n \exp\left\{\left(\frac{1}{\theta_0} - \frac{1}{\theta_1}\right) \sum x_i\right\}.$$
Since $\left(\frac{1}{\theta_0} - \frac{1}{\theta_1}\right) > 0$ we see that $\Lambda(x)$ is an increasing function of $t(x) = \sum x_i$. So the Neyman Pearson test will be *reject* H_0 if $t(x) > k_\alpha$ where k_α is chosen so that $\mathbb{P}_{\theta_0}(t(X) > k_\alpha) = \alpha$. Since $t(X) \sim \text{Gamma}(n, 1/\theta)$ and therefore we can determine k_α (from tables for Gamma cdf's or numerical computations) and you can see that k_α does not depend on θ_1. In other words the test is UMP for all $\theta \in \Theta_1$.

Suppose now that we want to test $H_0^* : \theta \leqslant \theta_0$ against $H_1 : \theta > \theta_0$. Note that
$$\mathbb{P}_\theta(\sum_i X_i > k) = \mathbb{P}_\theta\left(\frac{\sum_i X_i}{\theta} > \frac{k}{\theta}\right) \tag{13.1}$$
$$= \mathbb{P}(Y > k/\theta) \tag{13.2}$$
where Y is a Gamma-$(n, 1)$ r.v. This is a non decreasing function of θ. Therefore the test with critical region $C = \{x : t(x) > k_\alpha\}$ with size α for H_0 also has size α for H_0^*. Now let $\phi(X)$ be any other test of size $|alpha$ under H_0^*. Since H_0 is a smaller hypothesis than H_0^*, the test ϕ also

has size $\leqslant \alpha$ under H_0. But then by the Neyman-Pearson Theorem $\mathbb{E}_{\theta_1}\phi(X) \leqslant \mathbb{E}_{\theta_1}\phi_0(X)$ for all $\theta_1 > \theta_0$. Thus ϕ_0 is UMP.

Example. Suppose the $X_1, \ldots, X_n$ are i.i.d from the (one dimensional exponential) density

$$f(x, \theta) = h(x)e^{\theta T(x) - B(\theta)}.$$

Write $t(\mathbf{x}) = \sum_i T(x_i)$. Then

$$\frac{f(\mathbf{x}, \theta_2)}{f(\mathbf{x}, \theta_1)} = e^{n(B(\theta_1) - B(\theta_2))} \exp\{(\theta_2 - \theta_1)t(\mathbf{x})\}.$$

This is non-decreasing in $t(\mathbf{x})$ and so the family is MLR.

13.2 Bayes factors

13.2.1 Bayes factors for simple hypotheses

Consider first the case $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ (simple vs simple). We need to specify a prior probability for each hypothesis: let's call π_0 the prior probability of H_0 and π_1 that of H_1 with $\pi_0 + \pi_1 = 1$.

Bayes' rule tells us that (writing f_i for the density of X under H_i)

$$\mathbb{P}(H_0 \text{ is true} \mid X_x) = \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}$$

which can also be written as

$$\frac{\mathbb{P}(H_0 \text{ is true} \mid X = x)}{\mathbb{P}(H_1 \text{ is true} \mid X_x)} = \frac{\pi_0}{\pi_1}\frac{f_0(x)}{f_1(x)}.$$

In words this is often expressed as

$$\text{posterior odds} = \text{prior odds} \times \text{Bayes factor}.$$

Definition 13.6. We call $\frac{\pi_0}{\pi_1}$ the ***prior odds*** in favor of H_0 and $B = \frac{f_0(x)}{f_1(x)}$ is the ***Bayes factor***.

The first person to use Bayes factors extensively was Jeffreys, in his book *Theory of probability* (first edition 1939). This can be considered to be Jeffreys' main contribution to the theory of statistics. Following Jeffreys however, there were few methodological developments until the 80's. This is an active field of research today.

[insert table of interpretation of Bayes factors here]

From the point of view of decision theory, any Bayes rule will take the form $C = \{B < k\}$ [Make a theorem?] for some value k and is therefore an LRT. The class of Bayes Rules is exactly the class of Neyman-Pearson rules.

Remark. A rough guide to interpreting Bayes factors given by Adrian Raftery is as follows:

$\mathbb{P}(H_0 \mid x)$	$B_{0/1}$	$2\log(B_{0/1})$	evidence for H_0
< 0.5	< 1	< 0	negative (supports H_1)
0.5 to 0.75	1 to 3	0 to 2	barely worth mentioning
0.75 to 0.92	3 to 12	2 to 5	positive
0.92 to 0.99	12 to 150	5 to 10	strong
> 0.99	> 150	> 10	very strong

The value $2\log(B_{0/1})$ is sometimes reported because it's on the same scale as the familiar deviance and likelihood ratio test statistic.

13.2.2 Bayes factors for composite hypothesis

Suppose now that the hypothesis H_0 and/or H_1 are composite. Then, it is not enough to know the prior probabilities π_0 and π_1 that H_0 and H_1 are correct, we also need to know the full prior distribution of $\theta \in \Theta_0$ conditionally on H_0 being true and of $\theta \in \Theta_1$ conditionally on H_1 being true. Let's call those priors g_0 and g_1 respectively.

The hypothesis H_i is not just $\theta \in \Theta_i$ but rather the full prior model that

$$\theta \mid H_i \sim g_i(\theta), \theta \in \Theta_i.$$

Observe in particular that there is absolutely no need to suppose that the Θ_i are disjoints.

Definition 13.7. The ***Bayes factor*** in the composite-composite case is defined to be

$$B = \frac{\int_{\Theta_0} f(x,\theta) g_0(\theta)\,\mathrm{d}\theta}{\int_{\Theta_1} f(x,\theta) g_1(\theta)\,\mathrm{d}\theta}.$$

The ***Bayes factor*** in the simple-composite case is defined to be

$$B = \frac{f(x,\theta_0)}{\int_{\Theta_1} f(x,\theta) g_1(\theta)\,\mathrm{d}\theta}.$$

More generally, there is nothing here that requires the same parametrization under the two hypothesis. Suppose that we have two candidate parametric models M_1 and M_2 for data X, and the two models have respective parameter vectors θ_1 and θ_2. Under prior densities $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$, the marginal distribution for X under each models are found as

$$p(x \mid M_i) = \int f(x, \theta_i, M_i)\pi_i(\theta_i)\,\mathrm{d}\theta_i$$

and the ***Bayes factor*** is just their ratio

$$B = \frac{p(x \mid M_1)}{p(x \mid M_2)}.$$

Note that form this point of view, what we have is really a hierarchical Bayesian model where where the model correspond to the hyperparameter.

Example. Suppose the $X_1, \ldots, X_n$ are iid $\sim N(\theta, \sigma^2)$ with σ^2 known. Consider $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$. Also suppose that the prior g_1 under H_1 is $N(\mu, \tau^2)$. We have

$$B = \frac{p_1}{p_2}$$

where

$$p_1 = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum X_i^2\right),$$

and

$$p_2 = (2\pi\sigma^2)^{-n/2} \int_{\mathbb{R}} \exp\left\{-\frac{1}{2\sigma^2}\sum (X_i - \theta)^2\right\} (2\pi\tau^2)^{-1/2} \exp\left\{-\frac{1}{2\tau^2}(\theta - \mu)^2\right\} \mathrm{d}\theta.$$

Completing the square we see that for an arbitrary k we have

$$\sum_{i=1}^{n}(x_i - \theta)^2 + k(\theta - \mu)^2 = (n+k)(\theta - \widehat{\theta})^2 + \frac{nk}{n+k}(\bar{x} - \mu)^2 + \sum (x_i - \bar{x})^2$$

where $\widehat{\theta} = (n\bar{x} + k\mu)/(n+k)$. Thus:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2 + \frac{1}{\tau^2}(\theta-\mu)^2 = \frac{n\tau^2+\sigma^2}{\sigma^2\tau^2}(\theta-\widehat{\theta})^2 + \frac{n}{n\tau^2+\sigma^2}(\bar{x}-\mu)^2 + \frac{1}{\sigma^2}\sum(x_i-\bar{x})^2.$$

Using that

$$\int exp\left\{-\frac{n\tau^2+\sigma^2}{2\sigma^2\tau^2}(\theta-\widehat{\theta})^2\right\}\mathrm{d}\theta = \left(\frac{2\pi\sigma^2\tau^2}{n\tau^2+\sigma^2}\right)^{1/2}$$

we see that

$$p_2 = (2\pi\sigma^2)^{-n/2}\left(\frac{\sigma^2}{n\tau^2+\sigma^2}\right)^{1/2}\exp\left[-\frac{1}{2}\left(\frac{n}{n\tau^2+\sigma^2}(\bar{x}-\mu)^2 + \frac{1}{\sigma^2}\sum(x_i-\bar{x})^2\right)\right].$$

Hence, the Bayes factor is

$$B = \left(1+\frac{n\tau^2}{\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2}\left[\frac{n\bar{x}^2}{\sigma^2} - \frac{n}{n\tau^2+\sigma^2}(\bar{x}-\mu)^2\right]\right\}.$$

Writing $t == \sqrt{n}\bar{x}/\sigma, \eta = -\mu/\tau, \rho = \sigma/(\tau\sqrt{n})$ we can rewrite this

$$B = \left(1+\frac{1}{\rho^2}\right)^{1/2}\exp\left\{-\frac{1}{2}\left[-\frac{(t-\rho\eta)^2}{1+\rho^2} - \eta^2\right]\right\}.$$

This illustrate a difficulty with the Bayes factor approach. In general, many Bayesian solutions to point and interval estimation problems are close to the classical solutions when the prior is diffuse. However, here when we let $\tau^2 \to \infty$ we see that $\rho \to 0$ and thus $B \to \infty$. In other words, in the limit that the prior under H_1 is diffuse (infinite variance), then we have overwhelming support for H_0 no matter the observed data. This is an instance of ***Lindley's paradox***. One must therefore chose η, ρ to represent some reasonable judgement of where θ is likely to be when H_0 is false and there is no way to escape this by using some non-informative prior!

Example (Psychokinesis example). In 1987 Schmidt, Jahn and Radin ran an experiment where a subject with alleged psychokinetic ability tried to 'influence' a stream of quantum particles arriving at a quantum gate. Each particle would upon arrival at the gate either trigger a red light or a green light; the laws of quantum mechanics suggest a 50/50 ratio, and the subject tried to influence the particles to go to red.

Let X be the number of particles observed to go to red out of a total of n. We use the model $X \sim \mathrm{Bin}(n, \theta)$ where θ is unknown. In the experiment, $n = 104,490,000$ and the observed value of X was $x = 52263471$.

Has the subject influenced the particles?

Framing this as a hypothesis test, the natural choice of hypotheses is

$$H_0 : \theta = 1/2, \quad H_1 : \theta \neq 1/2.$$

The frequentist p-value is $\mathbb{P}_{\theta=1/2}(X \geqslant x) = 0.0003$. This suggests very strong evidence of paranormal ability?

Let's reframe this as a Bayesian test to see what's going on. Choose the mixed prior with $\pi_0 = 1/2$ and $g_1 = \mathbb{1}_{[0,1]}$ corresponding to a flat prior on $[0,1]$. Under this prior, the posterior probability of

H_0 is

$$\pi(H_0 \mid x) = \frac{\pi_0 f(x, 1/2)}{\pi_0 f(x, 1/2) + (\pi_1) \int_0^1 f(x, \theta)\, \mathrm{d}\theta} \tag{13.3}$$

$$= \frac{\binom{n}{x} 2^{-n}}{\binom{n}{x} 2^{-n} + \frac{1}{n+1}} \tag{13.4}$$

$$\approx 0.92 \tag{13.5}$$

in our casecorresponding to a Bayes factor of

$$B \equiv 11.5$$

This gives a very different conclusion from the one based on the p-value.

This reflects that we are reasonably sure before conducting the experiment that $\theta = 1/2$ is a more likely value than any other.

13.3 Hypothesis testing in the context of decision theory

13.3.1 Bayes tests for simple-simple hypothesis

Suppose we wish to test the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$ and consider the (non-random) test ϕ with critical region C

$$\phi(x) = \begin{cases} 1 \text{ if } x \in C \\ 0 \text{ if } x \notin C \end{cases}$$

A generic loss function can be written:

$$L(\theta, \phi(x)) = \begin{cases} a\phi(x) & \text{if } \theta = \theta_0 \\ b(1 - \phi(x)) & \text{if } \theta = \theta_1. \end{cases}$$

Lemma 13.8. *The rule ϕ has risk $R(\theta_0, \phi) = a\alpha$ and $R(\theta_1, \phi) = b\beta$ where $\beta = 1 - w(\theta_1)$.*

Proof. We have

$$R(\theta_0, \phi) = \mathbb{E}_{\theta_0}[a\phi(X)] = a\alpha \tag{13.6}$$

$$R(\theta_1, \phi) = \mathbb{E}_{\theta_1}[b(1 - \phi(X))] = b(1 - w(\theta_1). \tag{13.7}$$

□

To calculate the Bayes risk we need a prior π. Let $\pi(\theta_0) = p_0$ and $\pi(\theta_1) = p_1$ be the prior probabilities that H_0 and H_1 hold, respectively.

Lemma 13.9. *The Bayes risk for ϕ under the prior π is*

$$r(\pi, \Delta_C) = p_0 a\alpha(C) + p_1 b\beta(C).$$

Proof. Trivial, by calculating the expected risk. □

Remark. Note here that we write $\alpha = \alpha(C)$, $\beta = \beta(C)$ to emphasise that α, β depend on (and only on) our choice of critical region, whereas the other quantities are independent of it.

Definition 13.10. The ***Bayes test*** is the rule δ_C with the critical region C chosen to minimise the Bayes risk (under the loss function defined above).

Theorem 13.11 (Bayes test for simple hypotheses). *The critical region for the Bayes test with prior π and loss L is*

$$C = \left\{ x : \frac{f(x,\theta_1)}{f(x,\theta_0)} \geqslant A \right\}$$

where $A = \frac{p_0 a}{p_1 b}$.

Proof. The Bayes test minimises the Bayes risk

$$p_0 a\alpha + p_1 b\beta = p_0 a\,\mathbb{P}(X \in C \mid H_0) + p_1 b\,\mathbb{P}(X \in C' \mid H_1) \tag{13.8}$$

$$= p_0 a \int_C f(x,\theta_0)\,\mathrm{d}x + p_1 b \int_{C'} f(x,\theta_1)\,\mathrm{d}x \tag{13.9}$$

$$= p_0 a \int_C f(x,\theta_0)\,\mathrm{d}x + p_1 b \left[1 - \int_C f(x,\theta_1)\,\mathrm{d}x\right] \tag{13.10}$$

$$= p_1 b + \int_C \left[p_0 a f(x,\theta_0) - p_1 b f(x,\theta_1)\right]\mathrm{d}x. \tag{13.11}$$

So choose C such that $x \in C$ iff $p_0 a f(x,\theta_0) - p_1 b f(x,\theta_1) \leqslant 0$, i.e.

$$C = \left\{ x : \frac{f(x,\theta_1)}{f(x,\theta_0)} \geqslant \frac{p_0 a}{p_1 b} \right\}.$$

□

Corollary 13.12. *The Bayes test is a likelihood ratio test with $A = \frac{p_0 a}{p_1 b}$.*

Corollary 13.13. *Every likelihood ratio test is a Bayes test for some prior probabilities p_0, p_1.*

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known, and we want to test $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$, with $\mu_1 > \mu_0$.

The critical region for a likelihood ratio test becomes

$$C = \left\{ x \in \mathbb{R}^n : \frac{f(x,\mu_0)}{f(x,\mu_1)} \geqslant A \right\} \tag{13.12}$$

$$= \left\{ x \in \mathbb{R}^n : \bar{x} \geqslant \frac{\sigma^2 \log(A)}{n(\mu_1 - \mu_0)} + \frac{1}{2}(\mu_0 + \mu_1) \right\}. \tag{13.13}$$

For the Bayes test we need $A = \frac{p_0 a}{p_1 b}$, so we simply substitute into the above to find the critical region.

As an example, take $\mu_0 = 0, \mu_1 = 1, \sigma^2 = 1, n = 4, a = 2, b = 1, p_0 = \frac{1}{4}, p_1 = \frac{3}{4}$. Then

$$C = \left\{ x \in \mathbb{R}^n : \bar{x} \geqslant \frac{1}{4}\log\left(\frac{2}{3}\right) + \frac{1}{2} \right] = \{ x \in \mathbb{R}^n : \bar{x} \geqslant 0.3999 \}.$$

Using that $\bar{X} \sim \mathcal{N}(\mu, 1/4)$, this gives Type I/II error probabilties

$$\alpha = \mathbb{P}\left(\bar{X} \geqslant 0.3999 \mid \mu = 0, \frac{\sigma^2}{n} = \frac{1}{4}\right) = 0.212$$

and

$$\beta = \mathbb{P}\left(\bar{X} < 0.3999 \mid \mu = 1, \frac{\sigma^2}{n} = \frac{1}{4}\right) = 0.115.$$

The frequentist approach, fixing $\alpha = 0.05$, would give $\beta = 0.363$ (easy to check), so we see that in the Bayes test α is increased and β decreased relative to the frequentist test.

13.3.2 The case of the 0–1 loss function

In the case that L is the 0–1 loss, so $a = b = 1$ and

$$L(\theta, \delta_C(x)) = \begin{cases} 1 & \text{if } \theta = \theta_0 \text{ and } x \in C, \\ 1 & \text{if } \theta = \theta_1 \text{ and } x \notin C, \\ 0 & \text{otherwise,} \end{cases}$$

the Bayes test takes a particularly intuitive form.

Definition 13.14. The ***maximum a posteriori (MAP) test*** chooses the hypothesis with the highest posterior probability $\mathbb{P}(H_i \mid X = x)$.

Theorem 13.15. *The MAP test is the Bayes test under the 0–1 loss.*

Proof. Exercise. □

Let $m_i(x)$ be the marginal likelihood of x under hypothesis H_i. Thus is H_i is ***simple*** $H_i : \theta = \theta_i$ we have $m_i(x) = f(x, \theta_i)$. If H_i is ***composite*** $H_i : \theta \in \Theta_i$ we have

$$m_i(x) = \int_{\Theta_i} f(x, \theta) g_1(\theta) \, \mathrm{d}\theta.$$

Let π_0, π_1 be the prior probabilities of H_0, H_1.

Proposition 13.16. *The Bayes test for the 0–1 loss (i.e. the MAP test) rejects H_0 iff*

$$\frac{m_0(x)}{m_1(x)} < \frac{\pi_1}{\pi_0}.$$

Proof. This is just an application of Theorem 13.11 with $a = b = 1$.

Nevertheless let us just check that this is indeed the MAP test in the case where H_0 is simple and H_1 is composite. The marginal distribution for X under this (hierarchical) prior is

$$m(x) = \pi_1 \int_{\Theta_1} f(x, \theta) g_1(\theta) \, \mathrm{d}\theta + \pi_0 f(x, \theta_0).$$

Thus the posterior probability of H_0 is

$$\pi(H_0 \mid x) = \frac{\pi_0 f(x, \theta_0)}{\pi_1 \int_{\Theta_1} f(x, \theta) g_1(\theta) \, \mathrm{d}\theta + \pi_0 f(x, \theta_0)}.$$

The Bayes test for the 0–1 loss, i.e. the MAP test, rejects H_0 iff $\pi(H_0 \mid x) < \pi(H_1 \mid x)$, i.e. iff

$\pi(H_0 \mid x) < 1/2$. This occurs iff

$$2\pi_0 f(x,\theta_0) < \pi_1 \int_{\Theta_1} f(x,\theta) g_1(\theta)\,\mathrm{d}\theta + \pi_0 f(x,\theta_0) \tag{13.14}$$

$$\iff \pi_0 f(x,\theta_0) < \pi_1 \int_{\Theta_1} f(x,\theta) g_1(\theta)\,\mathrm{d}\theta \tag{13.15}$$

$$\iff \frac{f(x,\theta_0)}{\int_{\Theta_1} f(x,\theta) g_1(\theta)\,\mathrm{d}\theta} < \frac{\pi_1}{\pi_0}, \tag{13.16}$$

giving the result. □

Example. In a quality inspection program components are selected at random from a batch and tested. Let θ denote the failure probability. Suppose that we want to test the hypotheses

$$H_0 : \theta \leqslant 0.2, \quad H_1 : \theta > 0.2.$$

We use the following prior for θ: Let $g(\theta) = 30\theta(1-\theta)^4$ for $0 < \theta < 1$. We also set $\pi_0 = \int_0^{0.2} g(\theta)\,\mathrm{d}\theta \approx 0.345$ and $\pi_1 \approx 1 - 0.345$ with $g_0(\theta) = g(\theta \mid \theta \leqslant 0.2) = g(\theta)\mathbf{1}_{[0,0.2]}(\theta)/\pi_0, g_1(\theta) = g(\theta|\theta > 0.2)$.

Suppose n components are selected for independent testing. Modelling the number of failures X as $X \sim \mathrm{Bin}(n,\theta)$, the marginal likelihood for H_0 is

$$m_0(x) = \int_{\Theta_0} f(x,\theta) g_0(\theta)\,\mathrm{d}\theta \tag{13.17}$$

$$= \binom{n}{x} \int_0^{0.2} \theta^x (1-\theta)^{n-x} \frac{30\theta(1-\theta)^4}{\pi_0}\,\mathrm{d}\theta. \tag{13.18}$$

For one batch of size $n = 5$, the value $X = x = 0$ is observed. So

$$m_0(x) = \binom{5}{0} \int_0^{0.2} \frac{30\theta(1-\theta)^9}{\pi_0}\,\mathrm{d}\theta \approx \frac{0.185}{0.345} = 0.536.$$

Similarly $m_1(x) = \binom{5}{0} \int_{0.2}^1 \frac{30\theta(1-\theta)^9}{\pi_1}\,\mathrm{d}\theta \approx 0.134$.

So the Bayes factor is $B_{0/1} = \frac{m_0(x)}{m_1(x)} = \frac{0.536}{0.134} = 4 > \frac{\pi_1}{\pi_0} = 1.89$ so the Bayes test does not reject H_0.

Indeed, the overall marginal likelihood is $m(x) = m_0(x)\pi_0 + m_1(x)(1-\pi_0) \approx 0.273$, so the posterior probabilities for the hypotheses are $\pi(H_0 \mid x) = \frac{\pi(x|H_0)\pi_0}{m(x)} \approx \frac{0.185}{0.273} = 0.678$ and $\pi(H_1 \mid x) \approx 0.322$; we see that H_0 indeed maximises the posterior.

13.4 Exponential families

Good material here: Washington.edu

13.5 Two sided hypothesis tests

We now consider in more details situations in which $H_0 : \theta \in \Theta_0$ is either $\Theta_0 = [\theta_1, \theta_2]$ or $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \mathbb{R} \setminus \Theta_0$. In this situation we cannot expect to find a UMP test, even for nice families such as exponentials or MLR. The reason is obvious: if we construct a Neyman–Pearson test of say $\theta = \theta_0$ against $\theta = \theta_1$ for some $\theta_1 \neq \theta_0$, the test takes quite a different form when $\theta_1 > \theta_0$ from when $\theta_1 < \theta_0$. We simply cannot expect one test to be most powerful in both cases simultaneously. However, if we have an exponential family with natural statistic $T = t(X)$, or a family with MLR with respect to $t(X)$, we

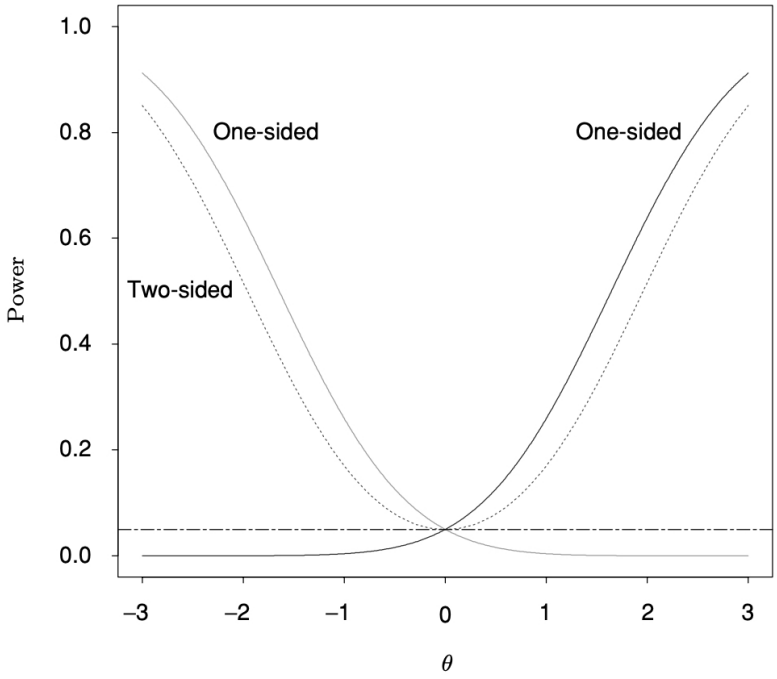

Figure 13.1: *Power functions for ones and two-sided tests (From Young and Smith).*

might still expect tests of the form

$$\phi(x) = \begin{cases} 1 \text{ if } x \in t(x) \notin [t_1, t_2] \\ \gamma(x) \text{ if } t(x) = t_1 \text{ or } t_2 \\ 0 \text{ if } x \in (t_1, t_2). \end{cases}$$

where $t_1 < t_2$ to have good properties. Such tests are called ***two sided tests*** based on T.

Definition 13.17. A test of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is called ***unbiased*** of size α if

$$\mathbb{P}_\theta(X \in C) \leqslant \alpha \quad \forall \theta \in \Theta_0 \qquad \text{but} \qquad \mathbb{P}_\theta(X \in C) \geqslant \alpha \quad \forall \theta \in \Theta_1.$$

A test which is uniformly most powerful amongst the class of all unbiased tests is called ***uniformly most powerful unbiased***, abbreviated UMPU.

The idea is illustrated by Figure 13.1, for the case $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ (In the figure, $\theta_0 = 0$.) The optimal UMP tests for the alternatives $H_1 : \theta > \theta_0$ and $H_1 : \theta, \theta_0$ each fails miserably to be unbiased, but there is a two-sided test whose power function is given by the dotted curve, and we may hope that such a test will be UMPU.

13.5.1 UMPU tests for one-parameter exponential families

Consider an exponential family of the form

$$f(x, \theta) = h(x) \exp\{\theta t(x) - B(\theta)\}$$

with $\theta \in \mathbb{R}$. Let $T = t(X)$ be the natural observation.

Remember that T itself also belongs to an exponential family with density form

$$f_T(t, \theta) = h_T(t) \exp\{\theta t - B(\theta)\}.$$

We shall assume that T is a continuous random variable with $h_T > 0$ on the open set that defines the range of T. This avoids the need for randomised tests and this makes our proofs less technical at the cost of very little loss of generality.

Theorem 13.18. *For any α there exists a UMPU test of size α which is of the two-sided form in T.*

We do not include a full proof of this result here. However, we mention that it starts with the following generalisation of Neyman-Pearson's Theorem:

Lemma 13.19. *Let $f_0, f_1, \ldots, f_m$ be $m+1$ probability densities, and let $\alpha_1, \ldots, \alpha_m$ be constants such that the class $\mathcal{C}$*

$$\mathcal{C} = \{\phi : \int \phi(x) f_i(x)\,\mathrm{d}x = \alpha_i,\ i = 1, \ldots, m\}$$

is non-empty. Then

1. *There is one member of $\mathcal{C}$ that maximizes $\int f_0(x)\phi(x)\,\mathrm{d}x$.*
2. *A necessary and sufficient condition for $\phi^* \in \mathcal{C}$ to be a maximizer is that there exists constants $k_1, \ldots, k_m$*

$$\phi(x) = \begin{cases} 1 \text{ if } f_0(x) > \sum_{i=1}^m k_i f_i(x) \\ 0 \text{ if } f_0(x) < \sum_{i=1}^m k_i f_i(x) \end{cases} . \tag{13.19}$$

3. *If $\phi \in \mathcal{C}$ satisfies (13.19) with $k_1, \ldots, k_m \geqslant 0$ then it maximises $\int f_0(x)\phi(x)\,\mathrm{d}x$ among all functions satisfying*

$$\int \phi(x) f_i(x)\,\mathrm{d}x \leqslant \alpha_i,\ i = 1, \ldots, m$$

Bibliography

Chang, Joseph T and David Pollard. "Conditioning as disintegration". In: ***Statistica Neerlandica*** 51.3 (1997), pp. 287–317.

Liero, Hannelore and Silvelyn Zwanzig. ***Introduction to the theory of statistical inference***. CRC press, 2016.