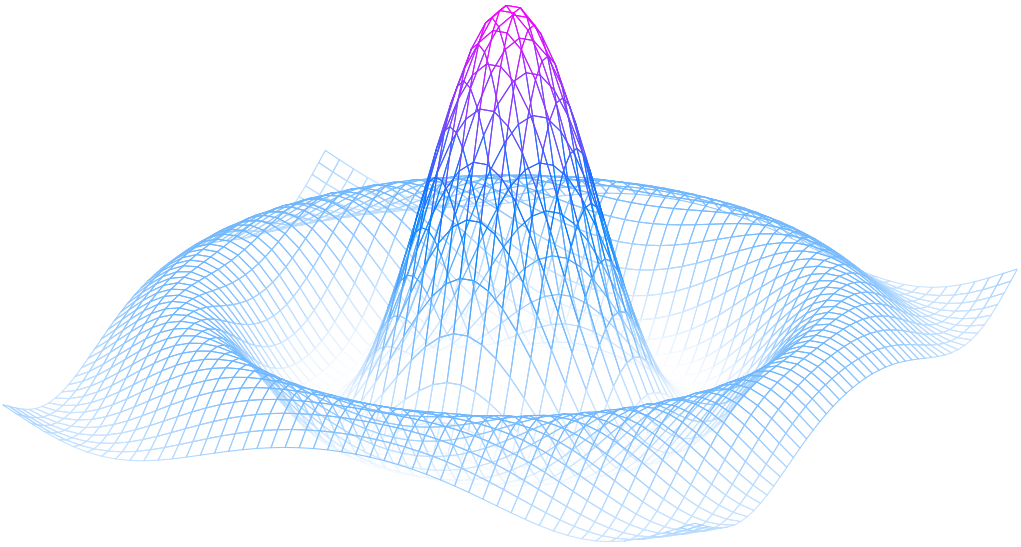
SB2.1

Foundations of Statistical Inference

Michaelmas Term 2020



Typeset and adapted by **Damon Falck**
from the handwritten lecture slides by **Julien Berestycki**

Updated 31 May 2021

Contents

Chapter 0

Notation

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, 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 will usually have a ***class*** of possible distributions $\mathcal{P}$, 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 X, is called a ***statistical model***. Here Θ is the ***parameter space***.

If P_θ is *continuous* we write $f(x, \theta)$ for its probability *density* function, whereas if P_θ is *discrete* we write $f(x, \theta)$ for its probability *mass* function. 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, and in most places I follow the notation of the handwritten notes these are based on, but do forgive me if it is confusing or even sickening at times.

Chapter 1

Exponential Families

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 $\{f(x;\theta) : \theta = (\theta_1, \theta_2, \ldots, \theta_k) \in \mathbb{R}^k\}$ of pdf/pmfs indexed by θ is a ***k-parameter exponential family*** if the pdf/pmfs $f(x;\theta)$ have the form

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

where the η_i and B are real-valued functions of θ, the T_i are real-valued ***statistics*** (i.e. functions of x), and x can be a vector or a scalar.

Important. In an exponential family $f(x;\theta)$ the support of $f(x;\theta)$ does *not* depend on θ. We will write $\mathcal{A}$ for the common support of the $f(x;\theta)$.

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

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

Since for all $\theta \in \Theta$

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

we can think of $\exp(-B(\theta))$ as a ***normalisation***. Observe that B only depends on $\eta(\theta)$.

Often, it is useful to write the model in its ***canonical form***,

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

(Note this is possible even if $\theta \mapsto \eta$ is not one-to-one.)

Remark. In general θ and x can be multidimensional.

Examples (Common 1-parameter exponential families).

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

$$\begin{aligned} f(x;\theta) &= \frac{1}{x!}\mathrm{e}^{-\theta + x\log\theta} \\ &= h(x)\exp(\eta(\theta)x - B(\theta)) \end{aligned}$$

  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 $\text{Bin}(n, p)$ distribution, considering n to be known and p to be the parameter, the mass may be written as

$$\begin{aligned} f(x;p) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} \exp\left[x(\log p - \log(1-p)) + n\log(1-p)\right] \end{aligned}$$

  (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 $\text{Gamma}(\alpha, \beta)$ distribution, with $\theta = (\alpha, \beta)$, we have mass function

$$\begin{aligned} f(x;\theta) &= \frac{\beta^\alpha x^{\alpha-1}\mathrm{e}^{-\beta x}}{\Gamma(\alpha)}\mathbb{1}_{x\geqslant 0} \\ &= \exp\Bigg[\underbrace{(\alpha-1)}_{\eta_1(\theta)}\underbrace{\log x}_{T_1(x)} - \underbrace{\beta}_{\eta_2(\theta)}\underbrace{x}_{T_2(x)} - \underbrace{(\log(\Gamma(\alpha)) - \alpha\log\beta)}_{B(\theta)}\Bigg]\underbrace{\mathbb{1}_{x\geqslant 0}}_{h(x)}. \end{aligned}$$

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

$$\begin{aligned} f(x;\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ &= \exp\Bigg[\underbrace{-\frac{1}{2\sigma^2}}_{\eta_1(\theta)}\underbrace{x^2}_{T_1(x)} + \underbrace{\frac{\mu}{\sigma^2}}_{\eta_2(\theta)}\underbrace{x}_{T_2(x)} - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)}_{B(\theta)}\Bigg]. \end{aligned}$$

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

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

Definition 1.2. If $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ with $d < k$ and $f(x;\theta) = h(x)\exp\left[\sum_{i=1}^{k} \eta_i(\theta)T_i(x) - B(\theta)\right]$ (where η_i is non-trivial for all i) then the family is said to be ***curved***.

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

$$\begin{aligned}\log f(x;\theta) &= -\frac{(x_1-\theta)^2}{2} - \frac{(x_2-\frac{1}{\theta})^2}{2} + \text{constant} \\ &= x_1\theta + x_2\frac{1}{\theta} - \frac{\theta^2}{2} - \frac{\theta^{-2}}{2} + \text{terms in } (x_1, x_2) \text{ alone},\end{aligned}$$

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.

Remark. In this course we normally assume exponential families not to be curved, i.e. in the above notation that $k = d$.

Observe that $\eta(\Theta) = \{(\theta, \frac{1}{\theta}) \in \mathbb{R}^2 : \theta \in \mathbb{R} \setminus \{0\}\}$ is a one-dimensional manifold — we can see from where the terminology 'curved' originates.

Definition 1.3. The ***parameter space*** is defined to be

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

i.e. the set of θ for which $f(x;\theta)$ can be defined.

Definition 1.4. The ***natural parameter space*** is defined to be

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

i.e. the set of η for which $f(x;\eta)$ can be defined (this is really an abuse of notation).

Observe that you can have $\eta(\Theta) \neq \Xi$ (but $\eta(\Theta) \subseteq \Xi$).

Proposition 1.5. *Ξ is convex.*

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

$$\begin{aligned}B(\alpha\eta + (1-\alpha)\eta') &= \log \int \exp(\alpha \textstyle\sum_i \eta_i T_i(x) + (1-\alpha)\sum_i \eta_i' T_i(x))h(x)\,\mathrm{d}x \\ &= \log \int \left[\exp(\textstyle\sum_i \eta_i T_i(x))h(x)\right]^{\alpha}\left[\exp(\textstyle\sum_i \eta_i' T_i(x))h(x)\right]^{1-\alpha}\mathrm{d}x \\ &\qquad\qquad \text{(using } h = h^{\alpha}h^{1-\alpha}\text{)} \\ &\leqslant \log\left(\int \exp(\textstyle\sum_i \eta_i T_i(x))h(x)\,\mathrm{d}x\right)^{\alpha}\left(\int \exp(\textstyle\sum_i \eta_i' T_i(x))h(x)\,\mathrm{d}x\right)^{1-\alpha} \\ &\qquad\qquad \text{by Hölder's inequality} \\ &= \alpha B(\eta) + (1-\alpha)B(\eta') < \infty.\end{aligned}$$

□

Definition 1.6. A family such that Ξ is open and non-empty is called ***regular***.

Definition 1.7. The functions $T_1, \dots, T_n$ are called ***$\mathcal{P}$-affine independent*** if for any $c_0, \dots, c_n \in \mathbb{R}$,

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

If $X \sim f(x;\eta)$, then $T = (T_1(X), \dots, T_N(X))$ is a random vector. Let $\mathrm{Cov}_\eta(T)$ be its covariance matrix under $f(x;\eta)$.

Proposition 1.8. *The functions T_i are $\mathcal{P}$-affine independent if $\mathrm{Cov}_\eta(T)$ is positive definite for all $\eta \in \Xi$.*

Proof. Omitted from lectures; see Liero & Zwanzig p.17. □

Definition 1.9. A family is ***strictly k-dimensional*** if the functions $\eta_i(\theta)$ are linearly independent and the T_i are $\mathcal{P}$-affine independent.

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

$$\begin{aligned} p(x;\theta) &= p_1^{I_1(x)} p_2^{I_2(x)} p_3^{I_3(x)} \qquad \text{where } I_i(x) \coloneqq \mathbb{1}_{x=i} \\ &= \exp(I_1(x)\log(p_1) + I_2(x)\log(p_2) + I_3(x)\log(p_3)), \end{aligned}$$

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.

Theorem 1.10. *The natural parameter space Ξ of a strictly k-parameter exponential family is convex and contains a non-empty k-dimensional interval.*

Proof. Omitted. □

Write $T = (T_1, \dots, T_k)$ for the vector of natural observations.

Theorem 1.11. *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, i.e.

$$\mathbb{E}_\eta[|T(X)|^k] < \infty \text{ for all } k \geqslant 1;$$

(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. See handwritten notes (lecture 1). □

Chapter 2

Sufficiency and Factorisation

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

$$\begin{aligned} f(x \mid t, p) = \mathbb{P}(X = x \mid T = t, p) &= \frac{\mathbb{P}(X = x, T = t \mid p)}{\mathbb{P}(T = t \mid p)} \\ &= \frac{\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} \\ &= \frac{p^t(1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} = \binom{n}{t}^{-1}, \end{aligned}$$

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 (Factorisation Criterion). *Suppose $X \sim f(x;\theta)$ and let $T(X)$ be any statistic. 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 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 $\text{Ber}(p)$ 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?

Definition 2.3. A statistic is ***minimal sufficient*** if it can be expressed as a function of any other sufficient statistic.

Remark (Partition induced by T). A statistic T induces a partition on $\mathcal{X}$ (the set of possible outcomes for X) via the equivalence relation $x \sim y \iff T(x) = T(y)$.

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

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.

We can think of a minimal statistic as one inducing the *coursest (least fine) possible partition* (i.e. conveying the least possible information).

Theorem 2.4 (Lehman-Scheffé Criterion). *A statistic T is minimal sufficient if and only if*

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

Example (continued). In the coin-tossing example, first consider T_2. With x = TTH and y = 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 = HTH and y = 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,

$$\begin{aligned} f(x \mid t, \theta) = \mathbb{P}_\theta(X = x \mid T = t) = \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}_\theta(T = t)} &= \frac{f(x;\theta)}{\sum_{y:T(y)=t} f(y;\theta)} \\ &= \frac{f(x;\theta)}{\sum_{y:T(y)=t} f(x;\theta)k(x,y)} \\ &= \left(\sum_{y:T(y)=t} k(x,y) \right)^{-1} \end{aligned}$$

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

Minimality. Now suppose 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, $T(x) = T(y)$. Thus $U(x) = U(y) \implies T(x) = T(y)$, i.e. the partition of U is *finer* than that of T. So T is a function of U. Hence T is minimal sufficient.

($\Longrightarrow$) 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, 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. But 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 $x \sim y \implies T(x) = T(y)$. □

Let's turn to the case of exponential families.

Theorem 2.5. *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* $X = (X_1, \ldots, X_n)$ *be a sample of i.i.d. random variables with distribution* $f(x,\theta)$*. Then:*

1. $T_{(n)} = \left(\sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_k(X_i)\right)$ *is minimal sufficient; and*
2. *the distribution of* $T_{(n)}(x)$ *belongs to a* k*-parameter exponential family.*

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

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

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 $j = 1, \ldots, k$.

The proof of the second point is left as an exercise. □

Examples.

1. **Bernoulli.** Let $X_1, \ldots, X_n$ 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

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

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

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

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

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

   Exercise: is it minimal sufficient?

3. **Normal.** Let $X = (X_1, \ldots, X_n)$ 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

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

   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.

Chapter 3

The Fisher Information and Point Estimation

3.1 The Fisher information

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 use partial derivatives and 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.

3.1.1 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* $\dfrac{\partial f(x, \theta)}{\partial \theta}$ *exists and is finite.*

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 $\theta \in \Theta$.

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

Theorem 3.6. *Under Regs 1–4,*

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

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,

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

and thus

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

The discrete case is similar. □

Proposition 3.7 (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 $Z=(X_1,\dots,X_n)$ 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. Omitted from lectures (does not imply off-syllabus). □

3.1.2 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 (that the support is independent of θ) 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′. *The log-likelihood ℓ has all its second partial derivatives, and these can all be commuted with summation/integration over $\mathcal{A}$.*

We can now generalise our definitions:

Definition 3.8. When Regs 1, 2′, 3′ are satisfied, we define the ***score function*** to be

$$S(\theta,x)=\nabla_\theta\ell(\theta,x)=\left(\frac{\partial}{\partial\theta_1}\ell(\theta,x),\dots,\frac{\partial}{\partial\theta_k}\ell(\theta,x)\right)^t.$$

Definition 3.9. 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

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

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 3.10. *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). □

3.2 Point estimation

Definition 3.11. 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 3.12. 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 $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ is an unbiased estimator for μ, and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \hat{\mu})^2$ is an unbiased estimator for σ^2

(Exercise: prove this.)

3.2.1 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 3.13. For eack $k = 1, \ldots, r$ let $\hat{m}_k = \frac{1}{n}\sum_{i=1}^n X_i^k$. Then the ***moment estimator*** for γ is defined as

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

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 $\hat{m}_1 = \frac{1}{n}\sum_{i=1}^n X_i$, so that

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

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

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

Which estimator is "better"?

3.2.2 Maximum likelihood estimators

Definition 3.14. 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 $\hat{\theta}_{MLE}$.

Theorem 3.15 (The Invariance Property). *If $\gamma = g(\theta)$ and g is bijective, then $\hat{\theta}$ is a MLE for θ if and only if $\hat{\gamma} = g(\hat{\theta})$ is a MLE for γ.*

Proof. Part A statistics. □

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

Theorem 3.16. *If $L(\theta, x)$ is differentiable (in θ) and has a unique maximum in $\mathrm{int}(\Theta)$, then $\hat{\theta}_{MLE}$ is the unique solution of $\frac{\partial}{\partial\theta} L(\theta, x) = 0$.*

Proof. Prelims analysis. □

3.2.3 Variance and mean squared error

Definition 3.17. 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 3.18. *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 $X = (X_1, \ldots, X_n)$ be a sample of i.i.d. $\mathcal{U}(0, \theta)$ random variables. Then $\hat{\theta}_{MLE} = X_{\max} = \max\{X_i : i = 1, \ldots, n\}$.

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 $\hat{\theta} = \frac{n+1}{n}X_{\max}$ is unbiased, and indeed

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

Chapter 4

MVUEs and the Cramer-Rao Lower Bound

Now that we've (among other things) developed some different techniques for estimating a parameter, it is natural to seek 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 4.1. 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 $\hat{\theta} = \theta_0$, then $\mathrm{MSE}_{\theta_0}(\hat{\theta}) = 0$. Hence no other estimator can be uniformly better!

Let's restrict ourselves now to unbiased estimators. We start with the univariate case.

4.1 The one-dimensional case

Definition 4.2. $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is the ***minimum variance unbiased estimator (MVUE)*** for θ (or $g(\theta)$) if

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

Theorem 4.3 (Cramer-Rao Lower Bound (CRLB) in 1 dimension). *Suppose Regs 1–4 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 an **unbiased** *estimator of γ. Then*

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

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

$$\begin{aligned}
\mathrm{Cov}_\theta(T, S(\theta, X)) &= \mathbb{E}_\theta[TS(\theta, X)] \qquad \text{since } \mathbb{E}_\theta(S(\theta, X)) = 0 \\
&= \int_{\mathcal{X}} T(x)\, \frac{\partial \log p(x, \theta)}{\partial \theta}\, p(x, \theta)\, \mathrm{d}x \\
&= \int_{\mathcal{X}} T(x)\, \frac{\partial p(x, \theta)}{\partial \theta}\, \mathrm{d}x \\
&= \frac{\partial}{\partial \theta} \int_{\mathcal{X}} T(x) p(x, \theta)\, \mathrm{d}\theta \qquad \text{(note this step strictly requires an additional hypothesis on } T\text{)} \\
&= \frac{\partial}{\partial \theta}\, \mathbb{E}_\theta[T] = g'(\theta).
\end{aligned}$$

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

$$\begin{aligned}
0 \leqslant \mathrm{Var}_\theta(T - c(\theta)S(\theta, X)) &= \mathrm{Var}_\theta\, T + c^2(\theta)\, \mathrm{Var}_\theta(S(\theta, X)) - 2c(\theta)\, \mathrm{Cov}_\theta(T, S(\theta, X)) \\
&= \mathrm{Var}_\theta\, T + c^2(\theta) I_X(\theta) - 2c(\theta) g'(\theta) \\
&= \mathrm{Var}_\theta(T) - \frac{|g'(\theta)|^2}{I_X(\theta)}
\end{aligned}$$

which is the CRLB. We have inequality 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 $g(\theta)$:

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

□

Example. Suppose $X \sim \mathrm{Bin}(n, \theta)$, 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

$$\begin{aligned} I_X(\theta) &= -\mathbb{E}_\theta\left[\tfrac{\partial}{\partial\theta}S(\theta,X)\right] \\ &= \frac{n-\mathbb{E}_\theta[X]}{1-\theta^2}+\frac{\mathbb{E}_\theta[X]}{\theta^2} = \frac{n}{(1-\theta)\theta}. \end{aligned}$$

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 $\mathrm{Var}_\theta(T)=\frac{\theta}{n}-\frac{\theta^2(5n-7)-4\theta^3(2n-3)+\theta^4(4n-6)}{n(n-1)}$ which is larger than the CRLB of $\frac{(1-2\theta)^2\theta(1-\theta)}{n}$.

Remark. We sometimes say that a statistic T is an ***efficient*** estimator for γ if it is unbiased for γ and attains the Cramer-Rao lower bound.

4.2 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 4.4. 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 4.5 (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. Omitted. □

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.

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

Theorem 4.6. *Under Regs 1, 2′, 3′, 4′, if $\hat{\theta}_{MLE}$ is the MLE for θ and if there exists $\tilde{\theta}$ which is unbiased and attains the CRLB, then $\tilde{\theta} = \hat{\theta}_{MLE}$ almost surely.*

Proof. Omitted. □

4.3 Exponential families and the CRLB

We conclude by returning to the case of an exponential family:

Theorem 4.7. *Suppose $X = (X_1, \ldots, X_n)$ belongs to a one-parameter exponential family in η and T. Then the sufficient statistic T is efficient (attains the CRLB) for $\gamma = g(\theta) = \mathbb{E}_\theta[T]$.*

Proof. Note that $p(x, \theta) = h(x) \exp[T(x)\eta(\theta) - B(\theta)]$. So $S(\theta, x) = \dfrac{\partial}{\partial\theta}\,\ell(\theta, x) = -B'(\theta) + \eta'(\theta)T(x)$. This means $S(\theta, x)$ and $T(x)$ are linearly related, which implies

$$\mathrm{Cov}_\theta(S(\theta, X), T(X))^2 = \mathrm{Var}_\theta(S(\theta, X))\,\mathrm{Var}_\theta(T(X)).$$

Since $\mathrm{Cov}_\theta(S(\theta, X), T(X)) = g'(\theta)$ and $\mathrm{Var}_\theta(S(\theta, X)) = I_X(\theta)$, we conclude that $\mathrm{Var}_\theta(T(X)) = \frac{[g'(\theta)]^2}{I_X(\theta)}$, so indeed T attains the CRLB. □

Chapter 5

Completeness and the Rao-Blackwell Theorem

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

Theorem 5.1 (Rao-Blackwell Theorem). *Let $X \sim P_\theta$ and let T be a sufficient statistic. Let $\hat{\gamma}$ be an unbiased estimator for $\gamma = g(\theta)$.*

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

1. *$\hat{\gamma}_T$ is a function of T alone and does not depend on θ,*
2. *$\mathbb{E}_\theta[\hat{\gamma}_T] = \gamma \ \forall \theta \in \Theta$ ($\hat{\gamma}_T$ is unbiased), and*
3. *$\operatorname{Var}_\theta(\hat{\gamma}_T) \leqslant \operatorname{Var}_\theta(\hat{\gamma})$, or $\operatorname{Cov}_\theta(\hat{\gamma}_T) \preceq \operatorname{Cov}_\theta(\hat{\gamma})$ in the case $\theta \in \mathbb{R}^k$.*

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

Intuitively, this says that 'any unbiased estimator can always be (weakly) improved by a sufficient statistic' — our best guess for the value of a particular unbiased estimator, given that we already know some sufficient statistic, is at least as good as knowing the real thing.

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
$$\hat{\gamma}_T = \mathbb{E}_\theta[\hat{\gamma} \mid T = t] = \int_{\mathcal{X}} \hat{\gamma}(x) f(x \mid t, \theta) \, \mathrm{d}x = \int_{\mathcal{X}} \hat{\gamma}(x) f(x \mid t) \, \mathrm{d}x$$
which does not depend on θ.
2. By the unbiasedness of $\hat{\gamma}$ and the tower property of expectations,
$$\mathbb{E}_\theta[\hat{\gamma}_T] = \mathbb{E}_\theta[\mathbb{E}_\theta[\hat{\gamma} \mid T]] = \mathbb{E}_\theta[\hat{\gamma}] = \gamma.$$

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

$$\begin{aligned}\mathrm{Var}_\theta(\hat{\gamma}) = \mathbb{E}_\theta[(\hat{\gamma}-\gamma)^2] &= \mathbb{E}_\theta[(\hat{\gamma}-\hat{\gamma}_T+\hat{\gamma}_T-\gamma)^2] \\ &= \mathbb{E}_\theta[\mathbb{E}_\theta[(\hat{\gamma}-\hat{\gamma}_T+\hat{\gamma}_T-\gamma)^2 \mid T]] \\ &= \mathbb{E}_\theta[\mathbb{E}_\theta[(\hat{\gamma}-\hat{\gamma}_T)^2 \mid T] - 2\,\mathbb{E}_\theta[(\hat{\gamma}-\hat{\gamma}_T)(\hat{\gamma}_T-\gamma) \mid T] + \mathbb{E}_\theta[(\hat{\gamma}_T-\gamma)^2 \mid T]] \\ &= \mathbb{E}_\theta[\mathbb{E}_\theta[(\hat{\gamma}-\hat{\gamma}_T)^2 \mid T]] - 0 + \mathbb{E}_\theta[\mathbb{E}_\theta[(\hat{\gamma}_T-\gamma)^2 \mid T]] \\ &= \mathbb{E}_\theta[\mathrm{Var}_\theta(\hat{\gamma} \mid T)] + \mathrm{Var}_\theta(\hat{\gamma}_T) \\ &\geqslant \mathrm{Var}_\theta(\hat{\gamma}_T).\end{aligned}$$

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

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

   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 fact about trace is left as an exercise. □

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

In this case,

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

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

Definition 5.2. A statistical model $\{P_\theta : \theta \in \Theta\}$ is called ***complete*** if for any $h : \mathcal{X} \to \mathbb{R}$,

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

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

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

Remark. This definition is motivated by the following consequence: if T is complete and $g(T)$ is unbiased for θ, then $g(T)$ is the *unique* (up to a.s.) unbiased estimator for θ that is a function of T. The proof of this is a simple application of the definition of completeness, and is left as an exercise.

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, so $h(x) = 0$ almost surely.

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 $0 = \mathbb{E}_\theta[h(X_{\max})] = \int_{-\infty}^{\infty} h(t) f_\theta(t)\,\mathrm{d}t = \frac{n}{\theta^n}\int_0^\infty h(t)t^{n-1}\,\mathrm{d}t$ for all $\theta \in \Theta$, we must have

$$\int_0^\infty h^-(t)t^{n-1}\,\mathrm{d}t = \int_0^\infty h^+(t)t^{n-1}\,\mathrm{d}t \;\forall\theta\in\Theta,$$

where $h^\pm$ are the positive/negative parts of h. This implies $h^+(t) = h^-(t)$ and therefore $h(t) = 0$ (almost surely).

Theorem 5.3 (Completeness for exponential families). *Assume $\mathcal{P}$ is a k-parameter exponential family with natural parameters $\eta = (\eta_1, \ldots, \eta_k)$ and that the natural parameter space Ξ contains a non-empty k-dimensional interval.*

Then $T(x) = (T_1(x), \ldots, T_k(x))$ is sufficient and complete.

Proof. Exercise. □

Corollary 5.4. *If P_θ belongs to a strictly k-parameter exponential family, then the vector of natural observations $T(x)$ is sufficient and complete.*

Proof. Immediate from theorem, since strictly k-parameter implies Ξ contains a non-empty k-dimensional interval (otherwise one of the natural parameters would be fixed by the others). □

Theorem 5.5 (Lehman-Scheffé Theorem). *Let T be a sufficient and complete statistic for the statistical model $\mathcal{P}$ and let $\hat\gamma$ be an unbiased estimator for $\gamma = g(\theta) \in \mathbb{R}^k$.*

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

Remark. In particular, any unbiased estimator that is a function of a complete sufficient statistic is an MVUE. This is how the Lehman-Scheffé Theorem will often be used.

Proof of theorem. By contradiction. Suppose there exists an unbiased estimator $\tilde\gamma$ with $\mathrm{Cov}_{\theta_0}\tilde\gamma \prec \mathrm{Cov}_{\theta_0}\hat\gamma_T$ for some $\theta_0 \in \Theta$.

The Rao-Blackwell Theorem implies, for $\tilde\gamma_T \coloneqq \mathbb{E}_\theta[\tilde\gamma \mid T]$, that

$$\mathrm{Cov}_{\theta_0}\tilde\gamma_T \preceq \mathrm{Cov}_{\theta_0}\tilde\gamma \prec \mathrm{Cov}_{\theta_0}\hat\gamma_T.$$

On the other hand, $\tilde\gamma_T$ and $\hat\gamma_T$ are both unbiased estimators which are functions of the complete statistic T. Hence, by completeness (see the remark after the definition) $\hat\gamma_T = \tilde\gamma_T$ a.s. and so $\mathrm{Cov}_{\theta_0}\tilde\gamma_T = \mathrm{Cov}_{\theta_0}\hat\gamma_T$, yielding the contradiction. □

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 $\hat{\theta} = \frac{n+1}{n} X_{\max}$ is the MVUE. (Note the CRLB does not apply.)

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 $\bar{X} \coloneqq \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 \coloneqq \frac{1}{n-1}(X_i - \bar{X})^2$.)

   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

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

   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 λ_{MME} is the MVUE.

   What is the Cramer-Rao bound? For a single observation, $S(x, \lambda) = \frac{x}{\lambda} - 1$ and $I_X(\lambda) = \lambda^{-1}$, so the lower bound is λ/n. Since also $\text{Var}(\bar{X})$, we conclude that $\lambda_{\text{MME}} = \bar{X}$ is efficient (it achieves the CRLB).

4. **Poisson 2.** What about the other estimator above, $\hat{\lambda}_{\text{MME}}$? Well, doing a little calculation (see the lecture slides for details) reveals that $X_i \mid \{\sum_{j=1}^n X_j = k\} \sim \text{Bin}(k, 1/n)$. 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

$$\begin{aligned}
\mathbb{E}_\lambda\left[S^2 \mid \sum_{j=1}^n X_j = k\right] &= \frac{n}{n-1}\left\{\mathbb{E}_\lambda\left[X_1^2 \mid \sum_{j=1}^n X_j = k\right] - \frac{k^2}{n^2}\right\} \\
&= \frac{n}{n-1}\left\{\frac{k}{n}\left(1 - \frac{1}{n}\right) + \frac{k^2}{n^2} - \frac{k^2}{n^2}\right\} \\
&= \frac{k}{n}.
\end{aligned}$$

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

Chapter 6

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.

6.1 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 we have met variously over the last two years. In our setting it most commonly reads as follows:

Theorem 6.1 (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**.

Proof. Prelims/Part A probability and statistics. □

Remark. It is worth emphasising here that in Bayesian statistics the likelihood $L(\theta, x)$ is the *conditional pmf* of X given the random variable θ; or the conditional probability, in the case of a discrete distribution for the data. This is in contrast to the frequentist setting of the first half of the course, where $L(\theta, x)$ was just the pmf/pdf of X, parameterised by the value θ (which had a fixed unknown 'true' value).

Remark. The denominator 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's Theorem the

posterior distribution is

$$\begin{aligned}\pi(\theta \mid x) \propto \text{likelihood} \cdot \text{prior} & \\ &\propto \theta^x(1-\theta)^{n-x} \cdot \theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{a+x-1}(1-\theta)^{n-x+b-1}.\end{aligned}$$

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 make more observations of random variables dependent on our unknown parameter — as we gather more data, effectively — the information we have about the unknown parameter and we revise our beliefs accordingly.

6.2 Conjugate priors

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

Definition 6.2. 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

$$\forall \gamma \in \Gamma, x \in \mathcal{X}\ \exists \tilde{\gamma}(x) \text{ s.t. } \pi_\gamma(\cdot \mid x) = \pi_{\tilde{\gamma}(x)}(\cdot),$$

i.e. all posteriors also belong to the family.

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. See the handwritten notes for two example of conjugate priors; the first on the Gamma prior for the Gaussian distribution, and the second on the Beta prior for the binomial distribution.

It turns out exponential families have precisely this property!

Proposition 6.3 (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. $\text{Po}(\theta)$ 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).

6.3 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 6.4. 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 soon as

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

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 $X \mid p \sim \text{Bin}(n, p)$ 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 $X \sim \text{Po}(\lambda), \pi(\lambda) = \lambda^{-1}$.)

Chapter 7

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, so doesn't have to be a valid probability distribution. Non-informative priors can be used when little or no reliable information is available.

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

7.1 Uniform priors

Definition 7.1. The ***uniform prior*** is the prior $\pi(\theta) = 1 \ \forall \theta$.

Remark. Note this is just the Lebesgue measure on Θ (in the continuous case).

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

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

so is defined as long as $\int_\Theta L(\theta, x)\,\mathrm{d}\theta < \infty$ almost surely in x.

Example. Let $X \sim \mathrm{Exp}(\theta)$ and $\pi(\theta) = 1$. The marginal likelihood is $\int_0^\theta \mathrm{e}^{-\theta x}\theta\,\mathrm{d}\theta$ 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 reparametrising means the prior isn't flat anymore; in fact, as a prior in η, $\tilde{\pi}$ is very informative (large values are much more likely than small ones).

7.2 Jeffrey's prior

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

Definition 7.2. ***Jeffrey's prior*** is given, in the one-dimensional case, by

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

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

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

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

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.

7.2.1 Jeffrey's prior in higher dimensions

This definition generalises naturally to higher dimensions:

Definition 7.3. 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

$$\begin{aligned}
\pi(\lambda) \propto \sqrt{I_X(\lambda)} &= \sqrt{\mathbb{E}[(\ell'(\lambda, X))^2]} \\
&= \sqrt{\mathbb{E}\left[\left(\frac{x}{\lambda} - 1\right)^2\right]} \\
&= \sqrt{\sum_{x=0}^{\infty} f(x, \lambda)\left(\frac{x-\lambda}{\lambda}\right)^2} \\
&= \sqrt{e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!}\left(\frac{x^2}{\lambda^2} - \frac{2x}{\lambda} + 1\right)} \\
&= \sqrt{\frac{1}{\lambda^2}\,\mathbb{E}\left[\mathbb{E}[(X-\lambda)^2]\right]} \\
&= \lambda^{-1/2}.
\end{aligned}$$

Note this is an improper prior.

7.3 Maximum entropy prior

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

Definition 7.4. 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***.

Intuitively, entropy is a measure of the uncertainty of a distribution. A large entropy means the space is well-explored at all scales.

For a non-informative prior, then, it makes sense to pick the function 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.

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 7.5. *The density $\pi(\theta)$ that maximises* $\mathrm{Ent}[\pi]$ *subject to $\mathbb{E}[T_j(\theta)] = t_j$ for $j = 1, \ldots, p$ takes the p-parameter exponential family form*

$$\pi(\theta) \propto \exp\left[\sum_{i=1}^{p}\lambda_i T_i(\theta)\right] \quad \forall\theta\in\Theta,$$

where $\lambda_1, \ldots, \lambda_p$ are determined by the constraints.

Proof. Omitted; see Leonard and Hsu for a proof. □

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 8

Predictive Distributions

We move on now towards applications of Bayesian inference. Let us briefly touch on how we can make predictions for new datapoints.

Definition 8.1. If $X_1, \dots, X_n, X_{n+1}$ 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 $x = (x_1, \dots, x_n)$.

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 $Y \sim \mathrm{Po}(\theta)$ 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

   $$\begin{aligned} m(y) &= \frac{\left(\frac{\mathrm{e}^{-\theta}\theta^y}{y!}\right)\left(\frac{\beta^\alpha \mathrm{e}^{-\beta\theta}\theta^{\alpha-1}}{\Gamma(\alpha)}\right)}{\left(\frac{(\beta+1)^{\alpha+y}\theta^{\alpha+y-1}\mathrm{e}^{-(\beta+1)\theta}}{\Gamma(\alpha+y)}\right)} \\ &= \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)y!}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^y \end{aligned}$$

   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

$$\begin{aligned}\pi(\theta \mid x) \propto \pi(\theta)p(x\mid\theta) &\propto \exp\left[-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right]\prod_{i=1}^n \exp\left[-\frac{1}{2\sigma^2}(x_i-\theta)^2\right]\\ &\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_0^2}(\theta-\mu)^2-\frac{1}{\sigma^2}\sum_{i=1}^n(x_i-\theta)^2\right]\right)\\ &\propto \exp\left[-\frac{1}{2\sigma_n^2}(\theta-\mu_n)^2\right]\end{aligned}$$

   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 9

Heirarchical Models

In certain situations, the data we are modelling has a natural ***heirarchical*** 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. We'll discuss this more 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 $\hat{\theta} = \frac{\sum y_i}{\sum n_i} = 0.0739$.
- **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 $\hat{\theta}_i = \frac{y_i}{n_i}$. So in particular $\hat{\theta}_A = 0$ and $\hat{\theta}_B = 0.1442$.
- **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 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.

	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.*

The first method (frequentist, equal parameters) gives some pretty unlikely results (e.g. the observed death rate for hospital H is not probably 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 heirarchical 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 $\hat{a}, \hat{b}$ such that Beta$(\hat{a}, \hat{b})$ has the same mean and variance.

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

This approach is reasonable, but we have the problem that we're using the same data twice — once to pick $\hat{a}, \hat{b}$) 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 $(\hat{a}, \hat{b})$ 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 through and through!

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

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

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

Definition 9.1. A ***heirarchical Bayesian model*** introduces a vector ϕ of ***hyperparameters*** with a ***hyperprior*** distribution $p(\phi)$; the vector θ of parameters we are interested in is modelled as having conditionally independent entries given ϕ.

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

Example (continued). In the case of the hospital data, the joint posterior distribution is

$$\begin{aligned} p(\theta, a, b \mid y) &\propto p(y \mid \theta)p(\theta \mid a, b)p(a, b) \\ &= \left(\prod_{i=1}^{I} p(y_i \mid \theta_i)\right)\left(\prod_{i=1}^{I} p(\theta_i \mid a, b)\right) p(a, b) \\ &\propto \left(\prod_{i=1}^{I} \theta_i^{y_i}(1-\theta_i)^{n_i - y_i}\right)\left(\prod_{i=1}^{I} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta_i^{a-1}(1-\theta_i)^{b-1}\right) p(a, b). \end{aligned}$$

Thus we have

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

(all we did here was drop factors that depend only on a, b).

This shows that, given a, b, the θ_i have independent beta posteriors.

On the other hand, the posterior for (a, b) is

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

See the handwritten lecture slides for plots of these distributions.

Remark. How can we generate new datapoints using the joint posterior $p(\theta, \phi \mid y)$ in general?

We can use the existing data to first draw possible parameters from the current posterior and then draw new datapoints given the chosen parameters:

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

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 $\theta = (\theta_1, \ldots, \theta_I)$ is ***symmetric***, or ***exchange-***

able, if

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

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's see that conditional independence indeed satisfies this property:

Proposition 9.3. *If $\theta = (\theta_1, \ldots, \theta_I)$ 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. □

9.1 Gaussian data example

See the handwritten course slides for an extended example of heirarchical modelling with Gaussian-distributed data.

Chapter 10

Decision Theory

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 admissibility

As usual, we will assume a data ***model*** $X \mid \theta \sim f(x, \theta)$ for some parametric family $\{f(x, \theta) : \theta \in \Theta\}$, where Θ is our ***parameter space***.

We will introduce additionally now:

- 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 $L(\theta, a)$.
- 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.

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.

10.1.1 Admissibility

Let's see how we might compare decision rules.

Definition 10.2. 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 $\mathcal{D} = \{\text{estimators of the form } \hat{\theta}(x) = ax\}$ (so this is a family indexed by a).

Using the quadratic loss, the risk will in general be

$$R(\theta, \hat{\theta}) = \int_0^\infty (ax - \theta)^2 \cdot \frac{1}{\theta} \, \mathrm{d}x = (\frac{a^3}{3} - a + 1)\theta^2$$

which is minimised at $a = 3/2$. Thus $\hat{\theta}(x) = ax$ is *inadmissible* for all $a \neq 3/2$.

So $a = 3/2$ is a necessary condition for $\hat{\theta}$ to be admissible for quadratic loss; but we have *not* shown that $\hat{\theta}(x) = \frac{3}{2}x$ is admissible!

Remark. Note that being admissible is a fairly weak requirement; it is simply the absence of another property.

Remark. We will later see that some natural estimators are in fact inadmissible (see chapter 11).

10.2 Minimax rules and Bayes rules

We now further explore notions of 'best possible' decision rules.

Definition 10.3. 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; see the handwritten notes (lecture 10.2) for some diagrams showing why this might be the case.

Given a prior belief about the parameter θ, a more natural choice of rule emerges.

Definition 10.4. The ***Bayes integrated 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:

$$r(\pi, \delta) = \inf_{\delta' \in \mathcal{D}} r(\pi, \delta') =: m_\pi.$$

In the case that the infimum is not attained, we define the following:

Definition 10.5. Given $\varepsilon > 0$, if a decision rule δ_ε is such that

$$r(\pi, \delta_\varepsilon) < m_\pi + \varepsilon,$$

then δ_ε is said to be an ***ε-Bayes rule*** w.r.t. π.

A rule δ is said to be an ***extended Bayes rule*** if for all $\varepsilon > 0$ there is some prior π with respect to which it is ε-Bayes.

Let's see another way of looking at Bayes rules.

Definition 10.6. The ***expected posterior loss*** of a rule δ w.r.t. a prior π is

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

Proposition 10.7. *A Bayes rule minimises the expected posterior loss.*

Proof. The Bayes risk is

$$\begin{aligned} r(\pi, \delta) = \int R(\theta, \delta)\pi(\theta)\,\mathrm{d}\theta &= \int\int L(\theta, \delta(x)) f(\theta, x)\pi(\theta)\,\mathrm{d}x\,\mathrm{d}\theta \\ &= \int\int L(\theta, \delta(x))\pi(\theta \mid x)h(x)\,\mathrm{d}x\,\mathrm{d}\theta \\ &= \int h(x)\int L(\theta, \delta(x))\pi(\theta \mid x)\,\mathrm{d}\theta\,\mathrm{d}x \\ &= \int h(x)\Lambda(x)\,\mathrm{d}x \end{aligned}$$

so to minimise $r(\pi, \delta)$, for each x pick $\delta(x)$ to minimise $\Lambda(x)$. □

Now we turn to a version of admissibility that takes into account our prior:

Definition 10.8. A rule δ^* is said to be ***π-admissible*** if for all rules δ,

$$R(\theta, \delta) \leqslant R(\theta, \delta^*)\ \forall \theta \in \Theta \implies \pi(\{\theta : r(\theta, \delta) < R(\theta, \delta^*)\}) = 0.$$

In other words, we now don't care if there is a rule whose risk is less under some parameter unless that parameter could actually occur, with positive probability, according to our prior.

Theorem 10.9. *A Bayes rule w.r.t. π is π-admissible.*

Proof. By contradiction. If a Bayes rule δ^* is not π-admissible, there is some δ s.t. $R(\theta, \delta) \leqslant R(\theta, \delta^*)\ \forall \theta$ and $\pi(A_\delta) > 0$, where $A_\delta := \{\theta : R(\theta, \delta) < R(\theta, \delta^*)\}$, and so

$$\begin{aligned} r(\pi, \delta) - r(\pi, \delta^*) &= \int_{A_\delta} [R(\theta, \delta) - R(\theta, \delta^*)]\pi(\theta)\,\mathrm{d}\theta + \int_{A_\delta^c} [R(\theta, \delta) - R(\theta, \delta^*)]\pi(\theta)\,\mathrm{d}\theta \\ &= \int_{A_\delta} [R(\theta, \delta) - R(\theta, \delta^*)]\pi(\theta)\,\mathrm{d}\theta < 0 \end{aligned}$$

(since the integrand is negative). This contradicts that δ^* is Bayes.

(Note this argument requires a little measure-theoretic justification.) □

Proposition 10.10 (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 $\pi(A_\delta) > 0$. A contradiction!

□

10.3 Finite decision problems

Definition 10.11. 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.

Definition 10.12. The ***risk set*** $S \subseteq \mathbb{R}^k$ is the set of points $\{(R(\theta_1, \delta), \ldots, R(\theta_k, \delta)) : \delta \in \mathcal{D}\}$.

Lemma 10.13. *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. □

10.3.1 The case $k = 2$

The two-dimensional case (i.e. there are two possible parameters) is particularly interesting. See the handwritten notes for some exciting diagrams.

10.4 Relating Bayes to minimax

Theorem 10.14. *If δ is a Bayes rule w.r.t. π with $r(\pi, \delta) = c$ and δ_0 is a rule such that $\max_\theta R(\theta, \delta_0) = c$, then δ_0 is minimax.*

Proof. If for some other rule δ' we have $\max_\theta R(\theta, \delta') = c - \varepsilon$ for some $\varepsilon > 0$ (so δ_0 is not minimax) then

$$\begin{aligned} r(\pi, \delta') &= \int R(\theta, \delta')\pi(\theta)\,\mathrm{d}\theta \\ &\leqslant \int (c-\varepsilon)\pi(\theta)\,\mathrm{d}\theta \\ &= c - \varepsilon < r(\pi, \delta) \end{aligned}$$

so δ is not a Bayes rule. □

Theorem 10.15. *If δ is a Bayes rule w.r.t. π such that $R(\theta, \delta)$ does not depend on θ, then δ is minimax.*

Proof. Let $R(\theta, \delta) = c \ \forall \theta$. Then $r(\pi, \delta) = \int c\pi(\theta)\,\mathrm{d}\theta = c$.

If there exists δ' with $\max_\theta R(\theta, \delta') = c - \varepsilon$ for some $\varepsilon > 0$ (so δ is not minimax) then $r(\pi, \delta') \leqslant c - \varepsilon < c = r(\pi, \delta)$, giving us our contradiction. □

Remark. In other words, *the Bayes estimator with constant risk is minimax.*

Example (Minimax estimator for quadratic loss.). Suppose $X \sim \mathrm{Bin}(n, \theta)$ and $\pi(\theta) \sim \mathrm{Beta}(\alpha, \beta)$.

The Bayes estimator is $\hat{\theta} = \frac{\alpha + X}{\alpha + \beta + n}$ (this is the posterior mean). This gives risk

$$\begin{aligned} R(\hat{\theta}, \theta) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] &= \mathrm{MSE}(\hat{\theta}) \\ &= (\mathrm{Bias}(\hat{\theta}))^2 + \mathrm{Var}(\hat{\theta}) \\ &= \left[\theta - \mathbb{E}_\theta\left(\frac{\alpha + X}{\alpha + \beta + n}\right)\right]^2 + \mathrm{Var}\left(\frac{\alpha + X}{\alpha + \beta + n}\right) \\ &= \left[\theta - \frac{\alpha + n\theta}{\alpha + \beta + n}\right]^2 + \frac{n\theta(1-\theta)}{(\alpha + \beta + n)^2} \\ &= \frac{[\theta(\alpha + \beta) - \alpha]^2 + n\theta(1-\theta)}{[\alpha + \beta + n]^2}. \end{aligned}$$

We can see that if $\alpha = \beta = \sqrt{n}/2$ then this is constant in θ. Hence the minimax estimator for

quadratic loss is $\frac{x+\sqrt{n}/2}{n+\sqrt{n}}$.

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 10.16. The ***zero-one loss*** is of the form $L(\theta,\hat{\theta}) = \begin{cases} a & \text{if } |\theta-\hat{\theta}| > b, \\ 0 & \text{otherwise} \end{cases}$ where a, b are positive constants.

The ***absolute error loss*** is of the form $L(\theta,\hat{\theta}) = k|\hat{\theta}-\theta|$ where k is a positive constant.

The ***quadratic loss*** is of the form $L(\hat{\theta},\theta) = k(\hat{\theta}-\theta)^2$ where k is a positive constant.

Remark. See the handwritten notes (lecture 10.5) for diagrams of these loss functions.

Let's see what the Bayes estimate (Bayes rule) is for each of these losses, by minimising the expected posterior loss.

Proposition 10.17. *The Bayes estimate under the:*

1. *zero-one loss with interval radius b tends to the posterior mode as $b \to 0$;*
2. *absolute error loss is the posterior median;*
3. *quadratic loss is the posterior mean.*

Proof.

1. The expected posterior loss is

$$\begin{aligned} \Lambda(x) &= \int \pi(\theta \mid x) L(\theta,\hat{\theta})\,\mathrm{d}\theta \\ &= a\int_{\hat{\theta}+b}^{\infty} \pi(\theta \mid x)\,\mathrm{d}\theta + a\int_{-\infty}^{\hat{\theta}-b} \pi(\theta \mid x)\,\mathrm{d}\theta \\ &\propto 1 - \int_{\hat{\theta}-b}^{\hat{\theta}+b} \pi(\theta \mid x)\,\mathrm{d}\theta. \end{aligned}$$

   So the Bayes rule is to choose $\hat{\theta}(x)$ to maximise $\int_{\hat{\theta}-b}^{\hat{\theta}+b} \pi(\theta \mid x)\,\mathrm{d}\theta$.

   If $\pi(\theta \mid x)$ is unimodal then this $\hat{\theta}$ is the midpoint of the unique interval of length $2b$ on which $\pi(\theta \mid x)$ takes the same value at both ends.

   So as $b \to 0$, $\hat{\theta}$ tends towards the posterior mode.

2. The expected posterior loss is

$$\Lambda(x) = \int_{-\infty}^{\infty} |\hat{\theta}-\theta|\pi(\theta \mid x)\,\mathrm{d}\theta$$

so that

$$\frac{\partial}{\partial\hat{\theta}}\Lambda(x) = \int_{-\infty}^{\infty} \tfrac{\partial}{\partial\hat{\theta}}|\hat{\theta}-\theta|\pi(\theta \mid x)\,\mathrm{d}\theta = \int_{-\infty}^{\infty}(-1)^{\mathbb{1}_{\hat{\theta}>\theta}}\pi(\theta \mid x)\,\mathrm{d}\theta$$
$$= \int_{-\infty}^{\hat{\theta}}\pi(\theta \mid x)\,\mathrm{d}\theta - \int_{\hat{\theta}}^{\infty}\pi(\theta \mid x)\,\mathrm{d}\theta$$

so, setting this to zero, Λ is minimised when

$$\int_{-\infty}^{\hat{\theta}}\pi(\theta \mid x)\,\mathrm{d}\theta = \int_{\hat{\theta}}^{\infty}\pi(\theta \mid x)\,\mathrm{d}\theta,$$

i.e. $\hat{\theta}$ is the median of $\pi(\theta \mid x)$.

3. The expected posterior loss is

$$\begin{aligned}\Lambda(x) &= \mathbb{E}[(\hat{\theta}-\theta)^2 \mid X=x] \\ &= \mathbb{E}[(\hat{\theta}-\mu_x+\mu_x-\theta)^2 \mid X=x] \text{ where } \mu_x \text{ is the posterior mean} \\ &= (\hat{\theta}^2-\mu_x)^2 + 2(\hat{\theta}-\mu_x)\,\mathbb{E}[\theta-\mu_x \mid X=x] + \mathbb{E}[(\theta-\mu_x)^2 \mid X=x] \\ &= (\hat{\theta}-\mu_x)^2 + \mathrm{Var}(\theta \mid X=x).\end{aligned}$$

So Λ is minimised when $\hat{\theta}=\mu_x$, the posterior mean.

□

Chapter 11

The James-Stein Estimator

In this chapter we explore an interesting paradox.

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 $\hat{\mu}_{\text{MLE}} = X$, 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**

$$\hat{\mu}_{\text{JSE}} \coloneqq \left(1 - \frac{p-2}{\sum_{i=1}^{p} X_i^2}\right) X$$

strictly dominates $\hat{\mu}_{\text{MLE}}$ for quadratic loss.

(We will prove this shortly.)

Corollary 11.2. *If $p \geqslant 3$, $\hat{\mu}_{\text{MLE}}$ 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!

Let's see how to prove this.

Lemma 11.3 (Stein's Lemma). *For independent Gaussian random variables $X = (X_1, \ldots, X_p)$ 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,

$$\begin{aligned}\mathbb{E}[(X_i - \mu_i)h(X) \mid \{X_j : j \neq i\}] &= \int_{-\infty}^{\infty} (x_i - \mu_i)h(x)\mathrm{e}^{-(x_i-\mu_i)^2/2}\,\mathrm{d}x_i \\ &= \left[-\mathrm{e}^{-(x_i-\mu_i)^2/2}h(x)\right]_{x_i=-\infty}^{x_i=\infty} + \int_{-\infty}^{\infty} \frac{\partial h(x)}{\partial x_i}\,\mathrm{e}^{-(x_i-\mu_i)^2/2}\,\mathrm{d}x_i \\ &= 0 + \mathbb{E}\left[\frac{\partial h(X)}{\partial X_i} \mid X_j : j \neq i\right]\end{aligned}$$

since h is bounded. Applying the Tower Law again gives the result. □

Proof of Stein's Paradox. Consider the family of estimators $\hat{\mu}_{\text{JSE}} = \left(1 - \frac{a}{\sum X_i^2}\right) X$ indexed by the parameter a. These are called the ***James-Stein estimators***.

Recalling that $\hat{\mu}_{\text{MLE}} = X$, we get

$$R(\mu, \hat{\mu}_{\text{MLE}}) = \sum_{i=1}^{p} \mathbb{E}[(\mu_i - X_i)^2] = p$$

(since $\text{Var}(X_i) = 1$).

On the other hand, writing $\hat{\mu}_i \coloneqq \left(1 - \frac{a}{\sum_j X_j^2}\right) X_i$,

$$\begin{aligned}R(\mu, \hat{\mu}_{\text{JSE}}) &= \sum_{i=1}^{p} \mathbb{E}[(\mu_i - \hat{\mu}_i)^2] \\ &= \sum_{i=1}^{p} \left(\mathbb{E}[(\mu_i - X_i)^2] - 2a\,\mathbb{E}\left[\frac{(X_i - \mu_i)X_i}{\sum_j X_j^2}\right] + a^2\,\mathbb{E}\left[\frac{X_i^2}{\left(\sum_j X_j^2\right)^2}\right] \right).\end{aligned}$$

Now the first term is just 1, since $\text{Var}(X_i) = 1$, 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

$$\begin{aligned}R(\mu, \hat{\mu}_{\text{JSE}}) &= p - (2ap - 4a)\,\mathbb{E}\left[\frac{1}{\sum X_j^2}\right] + a^2\,\mathbb{E}\left[\frac{1}{\sum X_j^2}\right] \\ &= p - (2a(p-2) - a^2)\,\mathbb{E}\left[\frac{1}{\sum X_j^2}\right].\end{aligned}$$

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 $\hat{\mu}_{\text{JSE}}^{(\mu_0)} = \mu + \left(1 - \frac{p-2}{||X-\mu_0||^2}\right)(X - \mu_0)$ can be defined which shrinks X towards an arbitrary point μ_0, and it can easily be shown that this also strictly dominates $\hat{\mu}_{\text{MLE}}$. (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 $\hat{\mu}_{\text{JSE}}$, where $\mathbf{1}_p = (1, \dots, 1)$.

Remark. Observe that when $||X - \mu_0||^2 < p - 2$, the shinkage factor becomes negative. To avoid this problem, we can define

$$\hat{\mu}^{(\mu_0)}_{\text{JSE}+} = \mu_0 + \left(1 - \frac{p-2}{||X-\mu_0||^2}\right)^+ (X - \mu_0)$$

(where x^+ denotes the positive part), which strictly dominates $\hat{\mu}^{(\mu_0)}_{\text{JSE}}$).

It is worth noting that neither $\hat{\mu}^{(\mu_0)}_{\text{JSE}}$ nor $\hat{\mu}^{(\mu_0)}_{\text{JSE}+}$ 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 naïve estimate is the MLE $\hat{p}_i = Y_i/n_i$. These give rise to the estimated number of home runs $\hat{HR}_i = \hat{p}_i \cdot AB_i$ (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 $f_n(y) \coloneqq n^{1/2}\sin^{-1}(2y-1)$. 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

$$JS_i \coloneqq \bar{X} + \left(1 - \frac{p-3}{V}\right)(X_i - \bar{X})$$

for each i, where $\bar{X} = \sum X_i/p$ and $V = \sum (X_i - \bar{X})^2$ (here $p = 17$).

These estimates of the μ_i are shown in the figure, and transforming back will give us estimates $\hat{HR}_s$ 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 naïve estimator, but allows increased risk in estimation of individual components.

	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.*

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 heirarchical models met in chapter 9.

This motivates the empirical Bayes approach.

12.1 Basic setup

Recall that a heirarchical 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 heirarchical Bayesian model by replacing the hyperparameter vector ψ with a point-estimate $\hat{\psi}$ derived from the data.

So we now just have the likelihood $X \sim f(x, \theta)$ and the prior $\theta \sim \hat{\psi}(\theta) = \pi(\theta, \hat{\psi})$.

Remark. Empirical Bayes methods can be viewed as an approximation of a full heirarchical Bayes model that allows us to avoid doing ψ-integrals. One layer of the heirarchy has been 'chopped off'.

Recall that we met this idea briefly in chapter 9 before heirarchical models were introduced.

The reduced model has posterior

$$\hat{\pi}(\theta \mid x) \propto L(\theta, x)\pi(\theta, \hat{\psi})$$

and a ***Bayes estimator*** $\hat{\theta}_{\text{EB}}$ can be calculated using $\hat{\pi}(\theta \mid x)$. So for quadratic loss, we have $\hat{\theta}_{\text{EB}} = \int \theta\hat{\pi}(\theta \mid x)\,\mathrm{d}\theta$, 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 $\hat{\psi}$ of the hyperparameter? We have all the classical frequentist techniques at our disposal. The two most obvious ways are:

- Use the MLE $\hat{\psi} = \text{argmax}_{\psi} p(x \mid \psi)$ where

$$p(x \mid \psi) = \int L(\theta, x)\pi(\theta, \psi)\,\mathrm{d}\theta$$

  is the marginal likelihood.

- Use the method of moments: choose $\hat{\psi}$ such that $\pi(\theta, \hat{\psi})$ 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 $\hat{\alpha}, \hat{\beta}$ such that $\text{Beta}(\hat{\alpha}, \hat{\beta})$ has 'matched moments', i.e.

$$\frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} = m, \quad \frac{\hat{\alpha}\hat{\beta}}{(\hat{\alpha} + \hat{\beta})^2(\hat{\alpha} + \hat{\beta} + 1)} = v.$$

   This solves to $\hat{\alpha} = 1.4, \hat{\beta} = 8.6$.

4. Calculate the Bayes estimate, which for the quadratic loss is the posterior mean. In this case the posterior is $\hat{\pi}(\theta \mid y) = \text{Beta}(5.4, 18.6)$ so the posterior mean is 0.225

This estimate is less than the maximum-likelihood estimate of $\hat{\theta}_{\text{MLE}} = 4/14$ we'd get based solely on the current experiment, not taking into account past experiments.

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).*

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 $\hat{\tau}^2 = \frac{1}{p}\sum_{j=1}^{p}(X_j^2 - 1)$. (This is from the standard result for the MLE for the variance of a Gaussian distribution).

So the estimated posterior distribution is $\theta_i \mid x_i \sim \mathcal{N}\left(x_i \frac{\hat{\tau}^2}{1+\hat{\tau}^2}, \frac{\hat{\tau}^2}{1+\hat{\tau}^2}\right)$. Thus the Bayes estimator for quadratic loss, i.e. the posterior mean, is

$$\hat{\theta}_{\mathrm{EB},i} = X_i \frac{\hat{\tau}^2}{1+\hat{\tau^2}} = X_i \frac{\left(\frac{1}{p}\sum_{j=1}^{p} X_j^2\right) - 1}{\frac{1}{p}\sum_{j=1}^{p} X_j^2} = X_i \left(1 - \frac{p}{\sum X_j^2}\right).$$

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,

$$p(x_i \mid \lambda) = \int_0^\infty \frac{\mathrm{e}^{-\theta_i}\theta_i^{x_i}}{x_i!} \lambda \mathrm{e}^{-\lambda\theta_i} \,\mathrm{d}\theta_i = \left(\frac{1}{1+\lambda}\right)^{x_i} \frac{\lambda}{1+\lambda} \sim \mathrm{Geom}\left(\frac{\lambda}{1+\lambda}\right).$$

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 $\hat{\lambda} = \frac{1}{\bar{x}} = \frac{n}{\sum x_i}$.

Hence our empirical Bayes approximation gives marginal posterior

$$\hat{\pi}(\theta \mid x) \propto L(\theta, x)\pi(\theta, \hat{\lambda}) = \prod_{i=1}^{p} \mathrm{e}^{-\theta_i}\theta_i^{x_i}\hat{\lambda}\mathrm{e}^{-\hat{\lambda}\theta_i}.$$

We recognise from this expression that $\theta_i \mid x_i \sim \Gamma(x_i + 1, \hat{\lambda} + 1)$ for each i. So the EB estimator is the approximated posterior mean,

$$\hat{\theta}_{\mathrm{EB},i} = \frac{\alpha}{\beta} = \frac{x_i + 1}{\hat{\lambda} + 1} = \bar{x}\frac{x_i + 1}{\bar{x} + 1} = x_i \frac{1}{\bar{x} + 1} + \bar{x}\frac{\bar{x}}{\bar{x} + 1}.$$

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 $Y_i \sim \text{Po}(\theta_i)$ independently. Assume that the parameters θ_i are drawn independently from some distribution π whose form we do not know.

The posterior mean is

$$\begin{aligned}\hat{\theta}_i = \mathbb{E}[\theta_i \mid Y_i] &= \int \theta \pi(\theta \mid Y_i)\,\mathrm{d}\theta \\ &= \frac{\int \left(\frac{\theta^{Y_i+1}\mathrm{e}^{-\theta}}{Y_i!}\right)\pi(\theta)\,\mathrm{d}\theta}{\int \left(\frac{\theta^{Y_i}\mathrm{e}^{-\theta}}{Y_i!}\right)\pi(\theta)\,\mathrm{d}\theta} \text{ by Bayes' Theorem} \\ &= \frac{(Y_i+1)p(Y_i+1)}{p(Y_i)}\end{aligned}$$

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

$$\hat{\theta}_i = \frac{(y_i+1)\hat{p}(y_i+1)}{\hat{p}(y_i)} = \frac{(y_i+1)\cdot|\{j : y_j = y_i+1\}|}{|\{j : y_j = y_i\}|}.$$

Chapter 13

Bayesian Hypothesis Tests

We close by applying our Bayesian theory to hypothesis testing.

Throughout this chapter we assume $X = (X_1, \dots, X_n)$ are i.i.d. random variables with $X_i \sim f(x;\theta)$ for each i.

13.1 Simple hypotheses

Suppose we wish to test the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$.

We'll use the decision rule δ_C, where C is some ***critical region***, defined by

$$\delta_C(x) = \begin{cases} H_1 & \text{if } x \in C, \\ H_0 & \text{otherwise.} \end{cases}$$

We write $\alpha = \mathbb{P}(\text{reject } H_0 \mid H_0)$ and $\beta = \mathbb{P}(\text{accept } H_0 \mid H_1)$ for the ***Type I*** and ***Type II error probabilities*** respectively.

Our choice of loss function will be the obvious one:

$$L(\theta, \delta_C(x)) = \begin{cases} a\mathbb{1}_{x\in C} & \text{if } \theta = \theta_0 \\ b\mathbb{1}_{x\notin C} & \text{if } \theta = \theta_1. \end{cases}$$

Lemma 13.1. *The rule δ_C has risk $R(\theta_0, \delta_C) = a\alpha$ for θ_0 and $R(\theta_1, \delta_C) = b\beta$ for θ_1.*

Proof. We have

$$\begin{aligned} R(\theta_0, \delta_C) &= \int L(\theta_0, \delta_C(x)) f(x, \theta_0)\, \mathrm{d}x \\ &= \int a\mathbb{1}_{x\in C} f(x, \theta)\, \mathrm{d}x \\ &= a\int_{x\in C} f(x, \theta_0)\, \mathrm{d}x \\ &= a\alpha \end{aligned}$$

by definition of α, and the proof for θ_1 is indentical. □

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 13.2. *The Bayes risk for δ_C 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 13.3. 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).

How can we find this optimal critical region?

Recall first the following result from frequentist hypothesis testing (we will not use this result but it helps to clarify how the Bayesian approach is different):

Theorem 13.4 (Neyman-Pearson Lemma). *The best test of size α for H_0 against H_1 is a likelihood ratio test with critical region*

$$C = \left\{ x : \frac{f(x, \theta_1)}{f(x, \theta_0)} \geqslant A \right\}$$

for some constant $A > 0$ chosen such that $\mathbb{P}(X \in C \mid H_0) = \alpha$.

Proof. Part A statistics. □

Remark. By 'best test' we mean the test with the highest power. Recall that the power is defined as $1 - \beta$ and the size as α.

Thus in frequentist statistics we fix the Type I error, α, and this determines the value of A.

It turns out that the critical region that minimises the Bayes risk is of the same form:

Theorem 13.5 (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

$$\begin{aligned} p_0 a\alpha + p_1 b\beta &= p_0 a\, \mathbb{P}(X \in C \mid H_0) + p_1 b\, \mathbb{P}(X \in C' \mid H_1) \\ &= p_0 a \int_C f(x, \theta_0)\, \mathrm{d}x + p_1 b \int_{C'} f(x, \theta_1)\, \mathrm{d}x \\ &= p_0 a \int_C f(x, \theta_0)\, \mathrm{d}x + p_1 b \left[1 - \int_C f(x, \theta_1)\, \mathrm{d}x \right] \\ &= p_1 b + \int_C \left[p_0 a f(x, \theta_0) - p_1 b f(x, \theta_1) \right] \mathrm{d}x. \end{aligned}$$

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 13.6. *The Bayes test is a likelihood ratio test with* $A = \frac{p_0 a}{p_1 b}$.

Corollary 13.7. *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

$$\begin{aligned} C &= \left\{ x \in \mathbb{R}^n : \frac{f(x, \mu_0)}{f(x, \mu_1)} \geqslant A \right\} \\ &= \left\{ x \in \mathbb{R}^n : \bar{x} \geqslant \frac{\sigma^2 \log(A)}{n(\mu_1 - \mu_0)} + \frac{1}{2}(\mu_0 + \mu_1) \right\}. \end{aligned}$$

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.

13.1.1 The case of the 0–1 loss function

In the case that L is the 0–1 loss, so $a = b = 1$ and

$$L(\theta, \delta_C(x)) = \begin{cases} 1 & \text{if } \theta = \theta_1 \text{ and } x \in C, \\ 1 & \text{if } \theta = \theta_0 \text{ and } x \notin C, \\ 0 & \text{otherwise,} \end{cases}$$

the Bayes test takes a particularly intuitive form.

Definition 13.8. The ***maximum a posteriori (MAP) test*** chooses the hypothesis with the highest posterior probability $\mathbb{P}(H_i \mid X = x)$.

Theorem 13.9. *The MAP test is the Bayes test under the 0–1 loss.*

Proof. Exercise. □

13.2 Composite hypotheses

Now that we've developed the basic theory of Bayes tests, we're interested in generalising to the case that our hypotheses involve sets of values.

A general ***testing problem*** involves hypothesis

$$H_0 : \theta \in \Theta_0, H_1 : \theta \in \Theta_1$$

where $\Theta_0 \cap \Theta_1 = \emptyset$.

Definition 13.10. A hypothesis $H_j : \theta \in \Theta_j$ is called ***simple*** if Θ_j is a singleton, and is called ***composite*** if Θ_j is *not* a singleton.

13.2.1 The case of a simple null hypothesis

Suppose H_0 is simple and H_1 is composite; write $\Theta_0 = \{\theta_0\}$. If our prior π for θ is a continuous distribution (i.e. it has a density) then the prior probability of H_0 will always be zero. This is not desirable!

Instead, we construct a prior as a weighted mixture of a point mass on $\Theta_0 = \{\theta_0\}$ and a prior distribution π_1 on Θ_1:

$$\pi(\theta) = \begin{cases} p_0 & \text{if } \theta = \theta_0, \\ (1-p_0)\pi_1(\theta) & \text{otherwise.} \end{cases}$$

Using differential notation,

$$\pi(\mathrm{d}\theta) = p_0 \delta_{\theta_0}(\mathrm{d}\theta) + (1-p_0)\pi_1(\mathrm{d}\theta).$$

Proposition 13.11. *Under this mixed prior, the Bayes test for the 0–1 loss (i.e. the MAP test) rejects H_0 iff*

$$\frac{f(x,\theta_0)}{\int_{\Theta_1} f(x,\theta)\pi_1(\theta)\,\mathrm{d}\theta} < \frac{1-p_0}{p_0}.$$

Proof. The marginal distribution for X under this prior is

$$m(x) = \int_\Theta f(x,\theta)\pi(\mathrm{d}\theta) = p_0 f(x,\theta_0) + (1-p_0)\int_{\Theta_1} f(x,\theta_1)\pi_1(\theta)\,\mathrm{d}\theta.$$

Thus the posterior probability of H_0 is

$$\pi(H_0 \mid x) = \pi(\{\theta_0\} \mid x) = \frac{p_0 f(x,\theta_0)}{p_0 f(x,\theta_0) + (1-p_0)\int_{\Theta_1} f(x,\theta)\,\mathrm{d}\theta}.$$

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

$$\begin{aligned} & 2p_0 f(x,\theta_0) < p_0 f(x,\theta_0) + (1-p_0)\int_{\Theta_1} f(x,\theta)\,\mathrm{d}\theta \\ \iff\ & p_0 f(x,\theta_0) < (1-p_0)\int_{\Theta_1} f(x,\theta)\,\mathrm{d}\theta \\ \iff\ & \frac{f(x,\theta_0)}{\int_{\Theta_1} f(x,\theta)\pi_1(\theta)\,\mathrm{d}\theta} < \frac{1-p_0}{p_0}, \end{aligned}$$

giving the result. □

Remark. The expression $\frac{f(x,\theta_0)}{\int_{\Theta_1} f(x,\theta)\pi_1(\theta)\,\mathrm{d}\theta}$ here is called the ***Bayes factor***. We'll meet it soon in more generality.

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 $p_0 = \pi(H_0) = 1/2$ and $\pi_1 = \mathcal{U}[0,1]$. Under this prior, the posterior probability of H_0 is

$$\begin{aligned}\pi(H_0 \mid x) = \pi(\{1/2\} \mid x) &= \frac{p_0 f(x,1/2)}{p_0 f(x,1/2) + (1-p_0)\int_0^1 f(x,\theta)\,\mathrm{d}\theta} \\ &= \frac{\binom{n}{x}2^{-n}}{\binom{n}{x}2^{-n} + \frac{1}{n+1}} \\ &\approx 0.92\end{aligned}$$

in our case. 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.

13.2.2 The case of a point composite null hypothesis

Another common scenario is that Θ_0 is a proper linear subspace of Θ but is not a singleton. This is called a ***point composite hypothesis***. This often arises when we have multiple unknown parameters and our hypotheses involve only some of them.

The following example illustrates how to handle this situation.

Example. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)$ is unknown. We may wish to test the hypotheses

$$H_0 : \mu = 0, \quad H_1 : \mu \neq 0.$$

In this case $\Theta_0 = \{0\} \times \mathbb{R}^+$. This is an example of a point composite hypothesis.

To construct a prior for this test we follow the same approach as for a simple hypothesis, creating a weighted mixture $\pi = p_0\pi_0 + (1-p_0)\pi_1$ of priors π_0 on Θ_0 and π_1 on Θ_1. Now, however, we instead define π_0 as $\delta_0 \otimes \pi_\sigma$ where π_σ is a prior on R^+ for σ. (Previously π_0 was just a Dirac distribution.) As before, π_1 has a density, in this case on $\mathbb{R} \times \mathbb{R}^+$.

The posterior for H_0 will then be

$$\pi(H_0 \mid x) = \pi(\Theta_0 \mid x) = \frac{p_0 m_0(x)}{p_0 m_0(x) + (1-p_0)m_1(x)}$$

where, writing $f(x, \mu, \sigma^2)$ for the likelihood of μ, σ^2 having observed x,

$$m_0(x) = \int_0^\infty f(x, 0, \sigma^2)\pi_\sigma(\sigma)\,\mathrm{d}\sigma$$

is the marginal likelihood under H_0 and

$$m_1(x) = \int_{\mathbb{R}}\int_0^\infty f(x, \mu, \sigma^2)\pi_1(\mu, \sigma^2)\,\mathrm{d}\sigma\,\mathrm{d}\mu$$

is the marginal likelihood under H_1.

13.2.3 The general case

In general, for any hypotheses H_0, H_1, we can construct a mixed prior $\pi = p_0\pi_0 + p_1\pi_1$ (where π_0 is a prior on Θ_0 and π_1 is a prior on Θ_1) and we can write the posterior probability of Θ_0 as

$$\pi(\Theta_0 \mid x) = \frac{p_0 m_0(x)}{p_0 m_0(x) + (1-p_0)m_1(x)}$$

where $m_0(x)$ is the marginal likelihood under H_0 and $m_1(x)$ is the marginal likelihood under H_1.

Remark. As we've seen, the marginal likelihoods in common cases take the forms:

- $m_j(x) = \int_{\Theta_j} f(x, \theta)\pi(\theta \mid H_j)\,\mathrm{d}\theta$ **in the continuous case**,
- $m_j(x) = \sum_{\theta\in\Theta_j} f(x, \theta)\pi(\theta \mid H_j)$ **in the discrete case**,
- $m_j(x) = f(x, \theta_0)$ **in the case of a simple hypothesis**.

In this language we can find a general form for Bayesian hypothesis tests.

Definition 13.12. The Bayes factor of H_0 over H_1 is given by

$$B_{0/1}(X) = \frac{m_0(X)}{m_1(X)}.$$

Theorem 13.13. *The Bayes test under the 0–1 loss (the MAP test) rejects H_0 iff*

$$B_{0/1}(X) < \frac{1-p_0}{p_0}.$$

Proof. Exercise. Follow the reasoning from the case of a simple null hypothesis. □

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.

In the psychokinesis example, the Bayes factor is $B_{0/1} = 12$, corresponding to positive-to-strong evidence in favour of H_0 (no paranormal ability).

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

$$H_0 : \theta \leqslant 0.2, \quad H_1 : \theta > 0.2.$$

Using the prior $\pi(\theta) = 30\theta(1-\theta)^4$ for $0 < \theta < 1$, the hypotheses have prior probabilities

$$p_0 = \pi(H_0) = \pi(\theta \in \Theta_0) = \int_0^{0.2} 30\theta(1-\theta)^4 \,\mathrm{d}\theta \approx 0.345$$

and $p_1 = \pi(H_1) \approx 1 - 0.345$. The priors for θ under each hypothesis are then

$$\pi(\theta \mid H_0) = \frac{30\theta(1-\theta)^4}{p_0}, \quad 0 < \theta \leqslant 0.2$$

and

$$\pi(\theta \mid H_1) = \frac{30\theta(1-\theta)^4}{p_1}, \quad 0.2 < \theta < 1.$$

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

$$\begin{aligned} m_0(x) &= \int_{\Theta_0} f(x,\theta)\pi(\theta \mid H_0)\,\mathrm{d}\theta \\ &= \binom{n}{x} \int_0^{0.2} \theta^x (1-\theta)^{n-x} \frac{30\theta(1-\theta)^4}{p_0}\,\mathrm{d}\theta. \end{aligned}$$

For one batch of size $n = 5$, the value $X = x = 0$ is observed. So

$$m_0(x) = \binom{5}{0} \int_0^{0.2} \frac{30\theta(1-\theta)^9}{p_0}\,\mathrm{d}\theta \approx \frac{0.185}{0.345} = 0.536.$$

Similarly $m_1(x) = \binom{5}{0} \int_{0.2}^1 \frac{30\theta(1-\theta)^9}{p_1}\,\mathrm{d}\theta \approx 0.134$.

So the Bayes factor is $B_{0/1} = \frac{m_0(x)}{m_1(x)} = \frac{0.536}{0.134} = 4 > \frac{1-p_0}{p_0} = 1.89$ so the Bayes test does not reject H_0.

Indeed, the overall marginal likelihood is $m(x) = m_0(x)p_0 + m_1(x)(1-p_0) \approx 0.273$, so the posterior probabilities for the hypotheses are $\pi(H_0 \mid x) = \frac{\pi(x|H_0)p_0}{m(x)} \approx \frac{0.185}{0.273} = 0.678$ and $\pi(H_1 \mid x) \approx 0.322$; we see that H_0 indeed maximises the posterior.

The following example shows that Bayes tests, and the Bayes factor, are not defined when the prior is improper:

Example (Lindley's Paradox). Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\theta, \sigma^2)$ random variables, where σ^2 is known. We wish to test the following hypotheses (one composite, one simple):

$$H_0 : \theta = 0, \quad H_1 : \theta \neq 0.$$

Suppose the prior distribution under H_1 is $\theta \mid H_1 \sim \mathcal{N}(\mu\tau^2)$. The marginal likelihoods in this case

are

$$m_0(x) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum x_i^2\right)$$

and

$$m_1(x) = (2\pi\sigma^2)^{-n/2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}\sum (x_i-\theta)^2\right) \cdot (2\pi\tau^2)^{-1/2} \exp\left(-\frac{(\theta-\mu)^2}{2\tau^2}\right) \mathrm{d}\theta.$$

Completing the square and integrating gives

$$m_1 = (2\pi\sigma^2)^{-n/2} \left(\frac{\sigma^2}{n\tau^2+\sigma^2}\right)^{1/2} \cdot \exp\left[-\frac{1}{2}\left\{\frac{n}{n\tau^2+\sigma^2}(\bar{x}-\mu)^2 + \frac{1}{\sigma^2}\sum (x_i-\bar{x})^2\right\}\right].$$

So the Bayes factor is

$$B_{0/1} = \left(1+\frac{n\tau^2}{\sigma^2}\right)^{1/2} \exp\left[-\frac{1}{2}\left\{\frac{n\bar{x}^2}{\sigma^2} - \frac{n}{n\tau^2+\sigma^2}(\bar{x}-\mu)^2\right\}\right].$$

We see that $B_{0/1} \to \infty$ as $\tau \to \infty$ for all x. So 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 shows why we cannot allow improper priors for Bayes tests!

The more general phenomenon hinted at here — that the frequentist and Bayesian hypothesis tests can disagree in certain situations under certain diffuse priors — is called ***Lindley's paradox***.

13.3 Model selection

We conclude by briefly touching on a Bayesian framework for model selection. Suppose we have k candidate models $\mathfrak{M}_1, \ldots, \mathfrak{M}_k$ for our data x. Each model $\mathfrak{M}_i$ consists of a parametric family $f_i(x, \theta_i)$ for X and a prior $\pi_i(\theta)$ for the unknown parameter θ_i.

We want identify which model is most likely given the data. Suppose we assign to each model a prior probability $\Pi(\mathfrak{M}_i)$; for example $\frac{1}{k}$ in the case of a uniform prior. Write $m_i(x) = m(x \mid \mathfrak{M}_i)$ for the marginal distribution of X under $\mathfrak{M}_i$.

In this framework the Bayes factor of $\mathfrak{M}_j$ over $\mathfrak{M}_i$ is $B_{j/i} = \frac{m(x|\mathfrak{M}_j)}{m(x|\mathfrak{M}_i)}$, and the posterior probability of model $\mathfrak{M}_i$ is

$$\Pi(\mathfrak{M}_i \mid x) = \frac{\Pi(\mathfrak{M}_i) m(x \mid \mathfrak{M}_i)}{\sum_j \Pi(\mathfrak{M}_j) m(x \mid \mathfrak{M}_j)} = \left[\sum_j \frac{\Pi(\mathfrak{M}_j)}{\Pi(\mathfrak{M}_j)} B_{j/i}\right]^{-1}.$$

So in the case of a uniform prior Π, this is just the inverse of the sum of the Bayes factors.

We can then pick the model that maximises this posterior probability; this is the Bayes test/MAP test.