Statistical Inference

Lecture notes for the course wi4455

Version 1.42

Frank van der Meulen
Delft University of Technology, The Netherlands

Preface

These lecture notes support the course "statistical inference" at Delft University of Technology. Just like Young and Smith [2005], this course aims to provide a concise account of the essential elements of statistical inference and theory with particular emphasis on the contrasts among frequentist and Bayesian approaches. I believe it is very important to learn that there are different thoughts on how to do proper statistics. Many people collect data, often with the goal of information based decision making. Next, they seem to hope that statistics will give a clear cut answer to the various questions they may have. This is usually not possible; I will try to teach you why. Clearly, there are always choices to be made in modelling the data. But even if two well-educated statisticians agree on the statistical model, this does not imply they agree on the conclusions to be drawn. This does not invalidate the use of statistics, but you should be aware of the underlying causes for disagreement.

This syllabus is based on various sources. At many places I present the material as in either Young and Smith [2005] or Schervish [1995]. The chapter on decision theory is based on chapters 7 and 8 from Parmigiani and Inoue [2009]. At some places I have used the material from Shao [2003], Keener [2010] and Ghosh et al. [2006].

Starred sections or theorems in these lecture notes are not part of the exam. A couple of articles from that appeared in the statistical literature have been added. Though I strongly advise you to read these, they are not part of exam material. Exercises are scattered throughout the text. Many are from Keener [2010], Schervish [1995] and Young and Smith [2005].

The content of Chapter 2 should be mostly familiar and for that reason I will go through this part rather quickly.

Scripts corresponding to numerical examples can be found at `https://github.com/fmeulen/WI4455`. These scripts are written in either R or Julia.

Thanks to Eni Musta, Marc Corstanje and Laura Middeldorp for providing many exercise solutions and checking parts of these notes. I thank Geurt Jongbloed and students that took the course wi4455 for pointing out typo's, mistakes and suggestions for improving the text. In particular, Joost Pluim (2017-2018).

Delft, August 2022
Frank van der Meulen

Contents

Chapter 1

Preliminaries

In this chapter we set notation and introduce some well known statistical models, including exponential family models. Some examples are presented which cause disagreement about the right type of inference among statisticians. The final section is on stochastic convergence and includes definitions that will be used in later chapters.

1.1 What is statistics about?

Statistics is about extracting information from data. In the recent report Council [2013] (page 3), statistical inference is defined as follows:

> Inference is the problem of turning data into knowledge, where knowledge often is expressed in terms of entities that are not present in the data per se but are present in models that one uses to interpret the data.

It is important to know that for statistics as a field of study, one usually discerns

- **descriptive statistics**, which is about numerically and graphically summarising data;
- **inferential statistics**, which is about planning experiments, drawing conclusions from data and quantifying the uncertainty of statements;
- **statistical decision theory**, which is about using the available information to choose a among a number of alternative actions.

A clear informal introduction is given in the opening section of the worth reading book by Winkler [1972]

> Applications of statistics occur in virtually all fields of endeavour –business, the social sciences, education, and so on, almost without end. Although the specific details differ somewhat in the different fields, the problems can all be treated with the general theory of statistics. To begin, it is convenient to identify three major branches of statistics: descriptive statistics, inferential statistics, and the statistical decision theory. *Descriptive statistics* is a body of techniques for the effective organization, summarization, and communication of data. When the "man on the street" speaks of "statistics", he usually means data organized by the methods of descriptive statistics. *Inferential statistics* is a body of methods for arriving at conclusions extending beyond the immediate data. For example, given some information regarding a small subset of a given population, what can be said about the entire population? Inferential statistics, then, refers

to the process of drawing conclusions or making predictions on the basis of limited information. Finally, *statistical decision theory* goes one step further; instead of just making inferential statements, the decision maker uses the available information to choose among a number of alternative actions.

For most practitioners, descriptive statistics is probably most relevant. Tidying data and preprocessing data for further analysis is typically time-consuming but its importance cannot be understated. This course is mostly concerned with inference and to a lesser extend with statistical decision theory (Chapter 6). A key component of inference is the concept of a **statistical model**.

1.2 Statistical models

We will denote the data (also referred to as measurements or observations) by X. For making probability statements, we are interested in the distribution of X. Loosely speaking, this is a specification of the possible values that X can take and the corresponding probabilities. In introductory courses on probability the distinction between continuous and discrete random variables is made. As an example for a discrete random variable, if we measure counts we may specify that $X \sim Pois(\theta)$ (with $\theta > 0$). This is shorthand notation for assuming that X is a random variable that takes values in the set $\{0, 1, \dots\}$ and that

$$\mathbb{P}(X = x) = e^{-\theta}\frac{\theta^x}{x!}, \qquad x = 0, 1, 2, \dots .$$

The function $x \mapsto \mathbb{P}(X = x)$ is the probability mass function. If we assume $Y \sim Exp(\lambda)$ (with $\lambda > 0$), then this translates to assuming Y takes values in $[0, \infty)$ and

$$\mathbb{P}(a \leq Y \leq b) = \int_a^b \lambda e^{-\lambda y}\mathbf{1}_{[0,\infty)}(y)\,\mathrm{d}y, \qquad a \leq b.$$

Here, $f_Y(y) = \lambda e^{-\lambda y}\mathbf{1}_{[0,\infty)}(y)$ is the probability density function. The distinction between "continuous" and "discrete" random variables is restrictive and unnecessary. As a concrete example, suppose the random variable Z is defined by

$$Z = \begin{cases} X & \text{with probability } w \\ Y & \text{with probability } 1 - w \end{cases}, \tag{1.1}$$

where $w \in (0, 1)$. Clearly, Z is neither discrete nor continuous. Is it possible to define a density (similar to probability mass functions or density functions)? This is certainly not an exceptional example, consider for example a $YExp(1)$, $c > 0$ and define $X = \min(Y, c)$. Verify yourself that X is neither discrete nor continuous.

Using the language of measure theory, we now make the notions of sample space and random variable more precise. Actually, as we are also interested in describing probabilities of random vectors or even random functions, we will talk about random quantities to refer to any of these.

We assume that there is a probability space $(S, \mathcal{F}, \mathbb{P})$ underlying all calculations. A random quantity X is a measurable mapping

$$X : (S, \mathcal{F}) \to (\mathcal{X}, \mathcal{B}),$$

where $\mathcal{X}$ (with σ-field $\mathcal{B}$) is called the **sample space**. Recall the definition of the **pre-image** of a set under a map: if $\phi : A \to B$ is a map and $V \subseteq B$, then the pre-image of V under ϕ is defined by

$$\mathrm{preim}_\phi(V) = \{a \in A \mid \phi(a) \in V\}.$$

Often, this set is denoted by $\phi^{-1}(V)$, but note that the definition for the pre-image does not require ϕ to be invertible. The **distribution** of X, denoted by P, is defined by

$$\mathrm{P}(B) = \mathbb{P}(\mathrm{preim}_X(B)), \qquad B \in \mathcal{B}.$$

Other terminology for the same thing is "probability distribution of X" or "law of X".[1] If X depends on the parameter θ one often writes P_θ for the distribution of X. The expectation of X is written either as $\mathbb{E}X = \int X(s)\,\mathrm{d}\mathbb{P}(s)$ or $\mathrm{E}_\theta\, X = \int x\,\mathrm{dP}_\theta(x)$, the latter showing explicitly the dependence on the parameter θ. Sometimes we write P^X to denote the distribution of the random quantity X.

Example 1.1. Take $(S, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}([0,1]), \mu)$ with μ denoting Lebesgue measure on $[0,1]$. Define $X : [0,1] \to (0,\infty)$ by $X(s) = -\theta \log(s)$, then $\mathrm{P}_\theta([0,x]) = \mu([e^{-x/\theta}, 1]) = 1 - e^{-x/\theta}$. Hence P_θ is in fact the distribution of a random variable with Exponential distribution with mean θ.

While introductory texts in probability usually start from the underlying probability space, rather quickly this space is no longer mentioned as genuine interest lies in the distribution of X.

> **Exercise 1.1** Consider the outcome of a throw with a fair coin. Give two distinct pairs of underlying probability space and random variable that yield the same distribution P, assigning probability $1/2$ to tails and probability $1/2$ to heads.

All computer simulation are built from random numbers, implying that the underlying probability space is $[0,1]^N$ with its Borel-σ-algebra and Lebesgue measure.

Definition 1.2. A **statistical model** (statistical experiment) is a family of probability measures $\{\mathrm{P}_\theta,\ \theta \in \Omega\}$ on a measurable space $(\mathcal{X}, \mathcal{B})$. The set Ω is the **parameter space**. We denote the model by $\mathcal{E} = (\mathcal{X}, \mathcal{B}, \{\mathrm{P}_\theta,\ \theta \in \Omega\})$.

The measures P_θ are often called **sampling probabilities**.

Important note on notation: In many texts the letter Θ is used for the parameter space. In these notes, the parameter space will always be denoted by Ω, and we reserve the symbol Θ for a random quantity (this is particularly useful when considering Bayesian statistics, where all unknowns are treated as random quantities).

In this course we will mainly be dealing with **parametric models**, where the dimension of Ω is finite. Nonparametric models are models which are not parametric. Hence, in such a model at least one of the parameters is infinite-dimensional. The infinite-dimensional parameter is usually a function or measure. Examples of nonparametric models are

1. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{P}$ and it is assumed that $x \mapsto \mathrm{P}((-\infty, x])$ is concave;
2. $(X_1, Y_1), \ldots, (X_n, Y_n)$ and independent and it is assumed that

$$Y_i \mid X_i = x \sim N\left(r(x), \sigma^2\right)$$

for a twice-continuously differentiable regression function r.

Nonparametric models are definitely worthwhile studying, the additional flexibility offered reduces the risk of model misspecification (when compared to a parametric model). By the latter we mean specifying a statistical model which does not include the data-generating distribution. Both computationally and theoretically nonparametric models tend to be more challenging.

[1] Yet stated differently, P is the pushforward of the measure $\mathbb{P}$ under the map X, which is then also denoted $\mathrm{P} = X_\star\mathbb{P}$.

1.2.1 Densities

Often, it is convenient to specify statistical models by densities. In basic probability courses the density refers to either the probability mass function or probability density function. We aim for more generality and for that reason recap the following definitions from measure theory.

Definition 1.3. Suppose μ and ν are measures on the measurable space $(\mathcal{X}, \mathcal{B})$.

- The measure μ is said to be **dominated** by ν, denoted by $\mu \ll \nu$ if

$$\nu(A) = 0 \quad \Rightarrow \quad \mu(A) = 0 \quad \text{for all } A \in \mathcal{B}.$$

- A σ-finite measure μ is **absolutely continuous** with respect to the σ-finite measure ν if there exists a measurable function $f : \mathcal{X} \to [0, \infty]$ such that

$$\mu(A) = \int_A f(x)\, \mathrm{d}\nu(x) \quad \text{for all } A \in \mathcal{B}.$$

  In that case we write

$$f(x) = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)$$

  and call f the **Radon-Nikodym derivative** or simply **density** of μ with respect to ν.

Theorem 1.4 (Radon-Nikodym)**.** A σ-finite measure μ is absolutely continuous with respect to the σ-finite measure ν if and only if $\mu \ll \nu$. The density is ν-almost surely unique.

The relevant definition for statistical models is the following:

Definition 1.5. A statistical model $\{\mathrm{P}_\theta, \theta \in \Omega\}$ is **dominated** if there exists a σ-finite measure $\nu : \mathcal{B} \to [0, \infty]$ such that for all $\theta \in \Omega$ we have $\mathrm{P}_\theta \ll \nu$.

By the Radon-Nikodym theorem, we may represent a dominated model in terms of probability densities $f(\cdot; \theta) = (\mathrm{dP}_\theta / \mathrm{d}\nu)(\cdot)$. Note that the dominating measure is not unique and henceforth densities are not unique. In these notes, the symbol ν is reserved for the dominating measure. Discrete and continuous random variables are now seen to be random variables with density with respect to counting measure and Lebesgue measure respectively. The following example shows that mixed forms also exist. These do pop up in realistic applications of statistics!

Example 1.6. Suppose $X \sim Pois(\theta)$ and $Y \sim Exp(\lambda)$. If we let ν_c denote counting measure on $\{0, 1, 2, \dots, \}$, then

$$\mathrm{P}^X(B) = \int_B e^{-\theta} \frac{\theta^x}{x!}\, \mathrm{d}\nu_c(x).$$

If we let ν_ℓ denote Lebesgue measure on $\mathbb{R}$, then

$$\mathrm{P}^Y(B) = \int_B \lambda e^{-\lambda y} \mathbf{1}_{[0,\infty)}(y)\, \mathrm{d}\nu_\ell(y).$$

Now if Z is defined as in (1.1), then P^Z has density

$$w \frac{\mathrm{dP}^X}{\mathrm{d}\nu_c}(x) \mathbf{1}_{\mathbb{N}}(x) + (1 - w) \frac{\mathrm{dP}^Y}{\mathrm{d}\nu_\ell}(x) \mathbf{1}_{\mathbb{R}\setminus\mathbb{N}}(x)$$

with respect to $\nu_c + \nu_\ell$.

To see this, first note that $\mathrm{P}^X \ll \nu_c + \nu_\ell$ and $\mathrm{P}^Y \ll \nu_c + \nu_\ell$. Hence $\mathrm{P}^Z = w\mathrm{P}^X + (1-w)\mathrm{P}^Y \ll \nu_c + \nu_\ell$. Note that the measures ν_c and ν_ℓ are (mutually) singular: there exists a set A with $\nu_\ell(A) = \nu_c(A^C) = 0$. Simply take $A = \mathbb{N}$ so that $A^C = \mathbb{R} \setminus \mathbb{N}$. Now

$$\begin{aligned}\mathrm{P}^X(B) &= \int_B \frac{\mathrm{d}P^X}{\mathrm{d}\nu_c}(x)\,\mathrm{d}\nu_c(x) = \int_B \frac{\mathrm{dP}^X}{\mathrm{d}\nu_c}(x)\mathbf{1}_{\mathbb{N}}(x)\,\mathrm{d}\nu_c(x)\\ &= \int_B \frac{\mathrm{dP}^X}{\mathrm{d}\nu_c}(x)\mathbf{1}_{\mathbb{N}}(x)\,\mathrm{d}(\nu_c+\nu_\ell)(x).\end{aligned}$$

Similarly, we get

$$\begin{aligned}\mathrm{P}^Y(B) &= \int_B \frac{\mathrm{dP}^Y}{\mathrm{d}\nu_\ell}(x)\,\mathrm{d}\nu_\ell(x) = \int_B \frac{\mathrm{dP}^Y}{\mathrm{d}\nu_\ell}(x)\mathbf{1}_{\mathbb{R}\setminus\mathbb{N}}(x)\,\mathrm{d}\nu_\ell(x)\\ &= \int_B \frac{\mathrm{dP}^Y}{\mathrm{d}\nu_\ell}(x)\mathbf{1}_{\mathbb{R}\setminus\mathbb{N}}(x)\,\mathrm{d}(\nu_c+\nu_\ell)(x).\end{aligned}$$

The result now follows by combining the results in the previous two displays when computing $w\mathrm{P}^X(B) + (1-w)\mathrm{P}^Y(B)$. Please do note that it is important to include the indicators $\mathbf{1}_{\mathbb{N}}$ and $\mathbf{1}_{\mathbb{R}\setminus\mathbb{N}}$ in the density!

In probability and statistics, product spaces are of particular importance. If $(\mathcal{X}, \mathcal{A}, \mu)$ and $(\mathcal{Y}, \mathcal{B}, \nu)$ are measure spaces, then there exists a unique measure $\mu \times \nu$, called the **product measure**, on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \vee \mathcal{B})$ such that

$$(\mu \times \nu)(A \times B) = \mu(A)\nu(B)$$

for all $A \in \mathcal{A}$ and $B \in \mathcal{B}$. Here $\mathcal{A} \vee \mathcal{B}$ is the smallest σ-field containing all sets $A \times B$ with $A \in \mathcal{A}$ and $B \in \mathcal{B}$. Suppose random variables $X_1, \ldots, X_n$ are independent:

$$\mathbb{P}(X_1 \in B_1, \cdots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \times \cdots \times \mathbb{P}(X_n \in B_n)$$

Hence, the distribution of $(X_1, \ldots, X_n)$, say μ, is a product measure

$$\mu(B_1 \times \cdots B_n) = \mu_1(B_1) \times \cdots \times \mu_n(B_n).$$

For future reference, we introduce the following notation.

Notation 1.7. The density of a random quantity X will always be denoted by f_X, or simply f, if there is no risk of confusion with densities of other random quantities. If the density depends on a parameter θ, Then we write $f_X(\cdot;\theta)$.

Exercise 1.2 Check that in example 1.6

$$\mathrm{P}^Z([1,2]) = w\left(e^{-\theta}\theta + e^{-\theta}\frac{\theta^2}{2}\right) + (1-w)(e^{-\lambda} - e^{-2\lambda}).$$

Exercise 1.3 Suppose $X_1, X_2, X_3 \overset{\text{iid}}{\sim} Exp(\theta)$. In introductory courses on statistics you learn about maximum likelihood estimation.

1. Derive the maximum likelihood estimator for θ based on $X = (X_1, X_2, X_3)$.
2. Now suppose we do not fully observe X_3. Instead, we only observe where $X_3 \in [3, 4)$ or not. This means that the data are given by the random vector $X = (X_1, X_2, \mathbf{1}_{[3,4)}(X_3))$. Can you derive the maximum likelihood estimator in this case as well? Note that the likelihood is in fact the density of the distribution of X. In the fully observed case one can take Lebesgue measure on $\mathbb{R}^3$ as dominating measure. What dominating measure can be taken for P^X in the partially observed case?

1.3 Some well known distributions

For future reference we give some examples of well known distributions.

Example 1.8. If $X = 1$ with probability θ and 0 else, we write $X \sim Ber(\theta)$ and say that X has the Bernoulli-distribution with parameter θ. If $X_1, \ldots, X_n$ are independent $Ber(\theta)$-random variables, then $Y = \sum_{i=1}^n X_i \sim Bin(n, \theta)$ and Y is said to have the Binomial distribution. The density of Y is given by

$$f_Y(y; \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, \qquad y = 0, 1, \ldots, n.$$

Example 1.9. The Negative Binomial distribution arises by counting the number of successes in a sequence of independent and identically distributed Bernoulli trials (each with succes probability θ) before a specified (non-random) number of r failures occurs. We write $X \sim NegBin(r, \theta)$, for which

$$f_X(x; \theta) = \binom{x+r-1}{x} \theta^x (1-\theta)^r, \qquad x = 0, 1, \ldots.$$

Exercise 1.4 Verify the form of the density in the preceding example.

Example 1.10. The Geometric distribution arises by counting the number of independent and identically distributed Bernoulli trials (each with success probability θ) necessary for obtaining a first succes. If X denotes the number of trials, then $X \sim Geom(\theta)$ and $f_X(x; \theta) = (1-\theta)^{x-1}\theta$, $x = 1, 2, \ldots$.

Example 1.11. The Poisson distribution arises as a limiting case of the $Bin(n, \theta/n)$-distribution when $n \to \infty$. We write $X \sim Pois(\theta)$ so that $f_X(x; \theta) = e^{-\theta}\theta^x/(x!)$, $x = 0, 1, \ldots$.

Example 1.12. If X is normally distributed with mean μ and variance σ^2 we write $X \sim N\left(\mu, \sigma^2\right)$. The parameter $\rho = 1/\sigma^2$ is called the precision parameter. If $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$, X and Y are independent, then

$$cX + Y \sim N(c\mu_X + \mu_Y, c^2\sigma_X^2 + \sigma_Y^2)$$

for $c \in \mathbb{R}$. The density of the $N\left(\mu, \sigma^2\right)$-distribution is denoted by $\phi(\cdot; \mu, \sigma^2)$. The cumulative distribution function of the $N(0, 1)$-distribution is denoted by Φ. The upper α-quantile of the $N(0, 1)$-distribution is denoted by ξ_α so that $\mathbb{P}(Z \geq \xi_\alpha) = \alpha$.

Definition 1.13. Suppose $Z_1, \ldots, Z_k$ are independent $N(0,1)$-distributed random variables. Define $Z = \begin{bmatrix} Z_1 & \cdots & Z_k \end{bmatrix}'$. A k-dimensional random vector X has the **multivariate normal distribution** with mean vector μ and covariance matrix Σ if X has the same probability distribution as the vector $\mu + LZ$, for a $k \times k$ matrix L with $\Sigma = LL'$ and k-dimensional vector μ. The density of X is then given by

$$f_X(x) = (2\pi)^{-k/2}(\det\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right).$$

More information can be found for instance on the Wikipedia page https://en.wikipedia.org/wiki/Multivariate_normal_distribution. For future reference we include the following lemma.

Lemma 1.14. If the random vectors $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ satisfy

$$\begin{aligned} X &\sim N(m, P) \\ Y \mid X &\sim N(HX + u, R) \end{aligned}$$

then

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} m \\ Hm + u \end{bmatrix}, \begin{bmatrix} P & PH' \\ HP & HPH' + R \end{bmatrix}\right).$$

The covariance for Y can be easily remembered from the law of total variance

$$\begin{aligned} \operatorname{Cov} Y &= \operatorname{Cov} \mathbb{E}[Y \mid X] + \mathbb{E}\operatorname{Cov}(Y \mid X) \\ &= \operatorname{Cov}(HX + u) + \mathbb{E}R = H\operatorname{Cov}(X)H' + R. \end{aligned}$$

Example 1.15. We say $X \sim Exp(\theta)$ if X has density $f_X(x;\theta) = \theta e^{-\theta x}\mathbf{1}_{[0,\infty)}(x)$. If $X_1, \ldots, X_n$ are independent $Exp(\theta)$-random variables, then $Y = \sum_{i=1}^n X_i$ has the $Ga(n,\theta)$-distribution. We have

$$f_Y(y;n,\theta) = \frac{\theta^n}{\Gamma(n)} y^{n-1} e^{-\theta y}\mathbf{1}_{[0,\infty)}(y).$$

This definition applies also to noninteger values of n (provided that $n > 0$).

Example 1.16. We say X has the Beta density, denoted by $X \sim Be(\alpha,\beta)$, if X has density

$$f_X(x;\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbf{1}_{(0,1)}(x)$$

with respect to Lebesgue measure. Here B denotes the Beta-function and $\alpha, \beta > 0$. We have $EX = \alpha/(\alpha+\beta)$ and $\operatorname{Var} X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Example 1.17. Suppose Y has density f_Y. Let $a > 0$ and $b \in \mathbb{R}$. If $X = aY + b$, then $f_X(x;a,b) = a^{-1}f_Y\left(\frac{x-b}{a}\right)$. The family of densities $\{f_X(\cdot;a,b),\ a > 0, b \in \mathbb{R}\}$ is called a **location-scale family** of densities. The family of densities $\{\phi(\cdot;\mu,\sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 \in (0,\infty)\}$ constitutes an important example.

1.3.1 Exponential family models

Many well-known distributions are part of a general class known as the Exponential family. As many results can be directly derived for the Exponential family, this unifies proofs.

Definition 1.18. A parametric family with parameter space Ω and density $f_X(x;\theta)$ with respect to a measure ν on $(\mathcal{X}, \mathcal{B})$ is called an **exponential family** if

$$f_X(x;\theta) = c(\theta)h(x)\exp\left(\sum_{i=1}^{k}\xi_i(\theta)\tau_i(x)\right)$$

for some measurable functions $\xi_1, \dots, \xi_k$, $\tau_1, \dots, \tau_k$ and some integer k.

As the function $c(\theta)$ satisfies

$$c(\theta)^{-1} = \int h(x)\exp\left(\sum_{i=1}^{k}\xi_i(\theta)\tau_i(x)\right)\mathrm{d}\nu(x)$$

we see that we might as well parametrise the family by

$$\xi = (\xi_1(\theta), \dots, \xi_k(\theta)) \in \mathbb{R}^k.$$

Definition 1.19. In an exponential family the **natural parameter** is the vector $\xi = (\xi_1(\theta), \dots, \xi_k(\theta))$ and

$$\Gamma = \left\{\xi \in \mathbb{R}^k : \int h(x)\exp\left(\sum_{i=1}^{k}\xi_i\tau_i(x)\right)\mathrm{d}\nu(x) < \infty\right\}$$

is called the natural parameter space.

One advantage of this parametrisation is that the set Γ is convex. A proof is given in Theorem 2.62 in Schervish [1995]. The quantities $\tau_i(x)$ are sometimes referred to as natural statistics.

Example 1.20. If $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ then $X = (X_1, \dots, X_n)$ forms an exponential family with respect to Lebesgue measure on $\mathbb{R}^n$ with

$$h(x) = (2\pi)^{-n/2} \qquad \tau_1(x) = \sum_{i=1}^{n} x_i \qquad \tau_2(x) = \sum_{i=1}^{n} x_i^2$$

$$c(\theta) = \sigma^{-n}\exp\left(-\frac{n\mu^2}{2\sigma^2}\right) \qquad \xi_1(\theta) = \frac{\mu}{\sigma^2} \qquad \xi_2(\theta) = -\frac{1}{2\sigma^2}.$$

where $\theta = (\mu, \sigma)$. The natural parameter space is given by $\Gamma = \mathbb{R} \times (-\infty, 0)$.

Example 1.21 (Continuation of example 1.16.)**.** If $X \sim Be(\alpha, \beta)$, then the parameter is given by $\theta = (\alpha, \beta)$ and X is within the exponential family with

$$h(x) = \frac{\mathbf{1}_{(0,1)}(x)}{x(1-x)} \qquad c(\theta) = 1/\mathrm{B}(\alpha, \beta)$$

$$(\xi_1, \xi_2) = (\alpha, \beta) \qquad (\tau_1(x), \tau_2(x)) = (\log x, \log(1-x)).$$

The natural parameter space is given by $\Gamma = (0, \infty)^2$.

If the parameter space contains an open set, the family is said to be of **full rank**, else it is called **curved**.

Exercise 1.5

1. Prove that random samples from the following distributions form exponential families: Poisson, geometric, Gamma. What about the negative Binomial distribution?
2. Identify the natural statistics and the natural parameters in each case. What are the distributions of the natural statistics?

Exercise 1.6 Let $Y_1, \ldots, Y_n$ be independent and identically distributed $N\left(\mu, \mu^2\right)$. Show that this model is an example of a curved exponential family.

Exercise 1.7 Suppose $X_1, \ldots, X_n$ are independent Bernoulli random variables with θ_i the success probability for X_i. Suppose these success probabilities are related to a sequence of variables $t_1, \ldots, t_n$, viewed as known constants, through

$$\theta_i = \frac{1}{1 + \exp\left(-\alpha - \beta t_i\right)}, \qquad i = 1, \ldots, n.$$

Show that the joint densities of $X_1, \ldots, X_n$ form a two-parameter exponential family, and identify the statistics τ_1 and τ_2.
Hint: In order to derive the statistics first write down the joint density of the $\{X_i\}$. After you have done this, try to derive the relationship between the logarithm θ_i and $1 - \theta_i$ and substitute this into the joint density.

1.4 Examples of statistical models

Example 1.22. Suppose realisations of tuples (X_i, Y_i) are observed, where the *response* Y_i is assumed to depend on the vector of features X_i. Rather than modelling the joint distribution of (X_i, Y_i) one typically models the conditional distribution of Y_i given X_i. In the classical **linear model** it is assumed that

$$Y_i \mid X_i = x \sim \mu_\theta(x) + \varepsilon_i,$$

where $\{\varepsilon_i\}$ is assumed to be a sequence of mean-zero independent and identically distributed random variables and $\mu_\theta(x) = \theta' x$. It is common to assume that ε_i has a Normal distribution, but more heavily tailed distributions are also possible.

Generalised linear models generalise linear models by allowing for non-Normal response distributions , such as the Gamma, Poisson and Bernoulli-distribution. In Poisson-regression, one assumes

$$Y_i \mid X_i = x \sim Pois\left(e^{\mu_\theta(x)}\right),$$

whereas in logistic regression one assumes

$$Y_i \mid X_i = x \sim Ber\left(\psi(\mu_\theta(x))\right),$$

with $\psi(x) = (1 + e^{-x})^{-1}$. In both cases $\mu_\theta(x)$ is transformed such that the parameter only takes "valid" values (i.e. positive in case of the Poisson-distribution; in $(0, 1)$ for the Bernoulli-distribution).

Example 1.23. Let k be a positive integer and suppose $\{\pi_i,\, i = 1, \dots, k\}$ are numbers in $[0,1]$ that add to unity. A finite **mixture experiment** consists of the composite experiment in which the i-th experiment is chosen with probability π_i. As a simple example, suppose $Y = 1$ with probability π_1 and $Y = 2$ with probability $\pi_2 = 1 - \pi_1$. Suppose $X \mid Y = y \sim N\left(\mu_y, 1\right)$. We have

$$\begin{aligned}\mathbb{P}(X \in B) &= \mathbb{P}(X \in B \mid Y = 1)\mathbb{P}(Y = 1) + \mathbb{P}(X \in B \mid Y = 2)\mathbb{P}(Y = 2) \\ &= \pi_1 \int_B \phi(x; \mu_1, 1)\,\mathrm{d}x + \pi_2 \int_B \phi(x; \mu_2, 1)\,\mathrm{d}x\end{aligned}$$

and hence the density of X with respect to Lebesgue measure is given by

$$f_X(x) = \pi_1 \phi(x; \mu_1, 1) + \pi_2 \phi(x; \mu_2, 1).$$

The joint distribution of (X, Y) can be found upon noting that

$$\mathbb{P}(X \in B, Y = y) = \mathbf{1}_{\{y=1\}} \int_B \phi(x; \mu_1, 1)\,\mathrm{d}x + \mathbf{1}_{\{y=2\}} \int_B \phi(x; \mu_2, 1)\,\mathrm{d}x.$$

Hence the density of (X, Y) is

$$f_{X,Y}(x, y) = \mathbf{1}_{\{y=1\}} \phi(x; \mu_1, 1) + \mathbf{1}_{\{y=2\}} \phi(x; \mu_2, 1).$$

with respect to the product measure of Lebesgue measure and counting measure on $\{0, 1\}$.

Example 1.24. Time series analysis is concerned with the analysis of data gathered discretely over time. Think e.g. of temperature over time, the value of a stock over time, etc. It is assumed that the data are realisations of a sequence of random variables $\{X_{t_i},\, i = 0, \dots, n\}$, obtained at times $\{t_i,\, i = 0, \dots, n\}$. Denote $X_i \equiv X_{t_i}$. Clearly, independence of X_i and X_j (for $i \neq j$) is unreasonable to be assumed in many cases and therefore many models have been thought of for explicitly modelling the dependence structure (this is what many econometricians spend lots of time on). A simple example is given by the auto-regressive model of order 1. In this model we start with a sequence of independent and identically distributed random variables $\{Z_i\}$. Suppose X_0 has distribution μ_0. Then the model postulates that

$$X_{i+1} = \alpha X_i + Z_i.$$

If for example $Z_i \sim N\left(0, \sigma^2\right)$, then the parameter vector in this model is given by $\theta = (\alpha, \sigma^2)$. The distribution of $(X_0, \dots, X_n)$ is denoted by P_θ and the statistical model is given by the set of measures $\{\mathrm{P}_\theta,\, \theta \in \Omega\}$, where Ω is the parameter space.

Example 1.25. Continuous time Markov processes are often defined as solutions of a stochastic differential equation

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \sigma(t, X_t)\,\mathrm{d}W_t, \qquad X_0 = x_0.$$

Here W is a **Wiener process**. If you are unfamiliar with this, think of the above equation of a stochastically perturbed version of the ordinary differential equation $\mathrm{d}x(t) = b(t, x(t))\,\mathrm{d}t$. A realisation of $(X_t,\, t \in [0, T])$ can be simulated on a computer by taking $h > 0$ small and recursively computing

$$X_{t+h} = X_t + hb(t, X_t) + \sigma(t, X_t)\sqrt{h}Z,$$

where $Z \sim N(0, 1)$ is drawn independently of all previously simulated random quantities. A simulated trajectory of a one-dimensional diffusion is in figure 1.4. Suppose now that either b or σ depends on an unknown quantity θ and we have data $X = (X_t,\, t \in [0, T])$ or $X = (X_{t_1}, X_{t_2}, \dots, X_{t_n})$. It is not immediate how to define the distribution of X, its density, or the maximum likelihood estimator. While this example is slightly out of scope for this course, it clearly indicates that defining a (or the?) maximum likelihood estimator is not obvious in some settings.

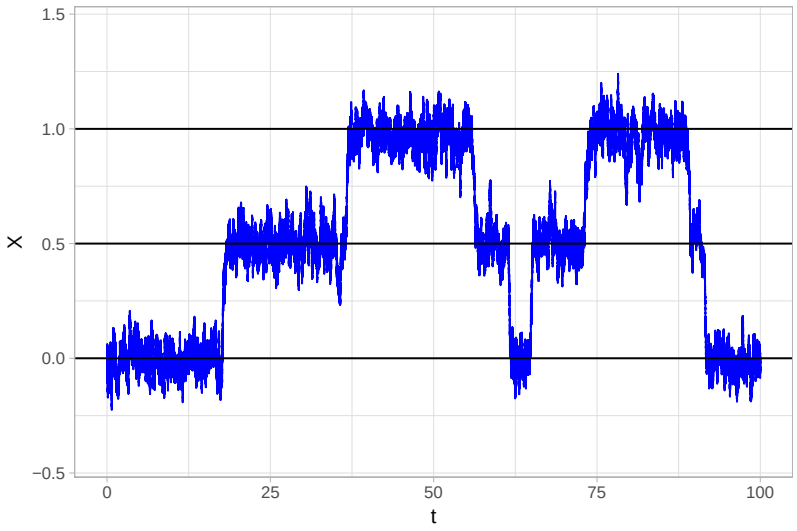

Figure 1.1: Simulation of a trajectory of the diffusion with $b(x) = (-0.2x - \sin(4\pi x))$ and $\sigma = 0.3$, starting at 0. The sin-term in the drift causes multimodal behaviour, whereas the linear part of the drift ensures mean-reversion within the modes.

Example 1.26. Consider a rod of length L with insulated sides that is given an initial temperature $g(x)$ at $x \in [0, L]$. Suppose that $u(x, t)$ is the temperature at $x \in [0, L]$ at time $t \geq 0$. Then u satisfies the heat equation

$$\frac{\partial u}{\partial t} = \theta \frac{\partial^2 u}{\partial x^2}.$$

If the temperature at the end of the rod is kept at 0 degrees, then the boundary conditions to this partial differential equation are given by

$$u(0, t) = u(L, t) = 0.$$

Furthermore, from the initial temperature we obtain the initial condition $u(0, x) = g(x)$. Suppose $(g(x),\ x \in [0, L])$ is known and at time $t = T$ the temperature is measured with noise at points $x_1, \ldots, x_K$. The measurements can be modelled as independent realisations of $Y_1, \ldots, Y_K$, where

$$Y_i = u(x_i, T) + \varepsilon_i.$$

Assume that $\{\varepsilon_i\} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The statistical problem is to estimate (θ, σ^2). Here, the likelihood is easily written down, but computationally hard to evaluate.

1.5 Different views on inferential statistics

There is no universal agreement on how statistical inference should be done. There are two main schools of thought: **classical** (also referred to as **frequentist** or **orthodox**) statistics and **Bayesian** statistics. The results obtained by either of these schools can lead to different conclusions for the same data, even when the same statistical model is employed. Within the Bayesian community, one can distinguish between subjective and objective Bayesian statistics (which causes further disagreement). The present situation is unsatisfying, when one takes into account that probably about 99% of all statistical analyses is conducted by non statisticians anyway.

A very short summary of the historical development in statistics is as follows:

- Objective Bayesian inference, using constant prior densities for unknowns, was prominent from 1775–1925, under the name inverse probability. The idea dates back to Laplace.
- By 1940, the prevailing statistical philosophies were either Fisherian (associated with Ronald Fisher) or frequentist (associated with Jerzy Neyman and Egon Pearson).
- Subjective Bayesian analysis became prominent by 1960 and was strongly advocated by Savage and Lindley.
- Harold Jeffreys revived the objective Bayesian school from 1940-1970.

It is fair to say that most people nowadays approach statistics from a frequentist point of view. Methods such as maximum likelihood and Neyman-Pearson testing (using type I and type II errors) are parts of introductory statistics courses, whereas Bayesian estimation quite often is not. Over the past 20 years Bayesian methods have flourished and to a large extent this is due to a tremendous increase in computational algorithms. Another development is often referred to as "big data", which means that either the amount of experimental units or the number of parameters in the statistical model is very large. It turns out that many ideas from Bayesian analysis can be used in these settings to derive estimators that enjoy favourable frequentist properties (as an example, the posterior mode can sometimes be interpreted as a regularised maximum likelihood estimator).

Because of the impact of choosing either the frequentist or Bayesian paradigm, it is no surprise that there has been a lively debate on the correct way to perform statistical inference. To get an impression of the debate you may wish to read chapter 16 of Jaynes [2003] (an online version is available on http://www-biba.inrialpes.fr/Jaynes/prob.html). Here is one quote from this book.

> During the aforementioned period, the average worker in physics, chemistry, biology, medicine, or economics with a need to analyze data could hardly be expected to understand theoretical principles that did not exist, and so the approved methods of data analysis were conveyed to him in many different, unrelated *ad hoc* recipes in "cookbooks" which, in effect, told one to "Do this... then do that ... and don't ask why.
>
> R.A. Fisher's "Statistical Methods for Research Workers (1925)" was the most influential of these cookbooks.

The book by Jaynes [2003] is very opinionated (but a must read in my opinion), and a somewhat milder discussion on the use of frequentist and Bayesian statistics can be found in two recent articles by the well known statistician Efron (Cf. Efron [1986] and Efron [2005]). The recent, very accessible, book Clayton [2021] strongly advocates Jaynes' approach (which is Bayesian, with probability being interpreted as information). To fight the reproducibility crisis, the inability to replicate experiments (especially in psychology and sociology), he proposes to stop teaching frequentist hypothesis testing, confidence intervals, sufficient statistics, and all those topics that have been taught traditionally in about any course on statistics. We'll take a bit milder point of view here, and shortly recap those topics in Chapter 2.

1.5.1 Example: classical and Bayesian estimation for Bernoulli trials

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Ber(\theta)$ and we are interested in estimating θ. In this section we consider various approaches to this problem, the details will become clearer (hopefully!) as we proceed along our way through this course.

From a frequentist perspective, there are at least three methods for obtaining an estimator:

(A) Maximum likelihood estimation: define the likelihood by

$$L(\theta) = \theta^S(1-\theta)^{n-S}, \qquad \theta \in [0,1] \quad S = \sum_{i=1}^{n} X_i.$$

Its maximiser is given by $\hat{\Theta}_{MLE} = \bar{X}_n$.

(B) Uniformly Minimum Variance Unbiased (UMVU) estimation: among all estimators $T = d(X_1, \ldots, X_n)$ that satisfy $\mathrm{E}_\theta T = \theta$ (if any exist), choose the estimator with minimum variance. It turns out that $\hat{\Theta}_{UMVU} = \bar{X}_n$ which is just the maximum likelihood estimator.

(C) Minimax estimation: choose the estimator that minimises the maximum value of the Mean Squared error. Define the Mean Square Error of T for estimating θ by $\mathrm{MSE}(\theta, T) = \mathrm{E}_\theta(T-\theta)^2$. The minimax estimator is defined as the estimator that minimises the worst case value of the Mean Square Error:

$$\hat{\Theta}_{MM} = \operatorname*{argmin}_T \max_{\theta\in[0,1]} \mathrm{MSE}(\theta, T).$$

It turns out that

$$\hat{\Theta}_{MM} = \frac{n\bar{X}_n + \sqrt{n}/2}{n+\sqrt{n}}.$$

While the third estimator is different from the other two, it is easily seen that asymptotically, when n is large, the difference diminishes.

Within Bayesian statistics all unknowns are treated as random quantities, including in particular the parameter. In this sense, the only distinction between data and parameters hinges on what is observed. The Bayesian statistician then proceeds by defining the joint distribution of (X, Θ), where $X = (X_1, \ldots, X_n)$, in the following way:

$$\begin{aligned} X_1, \ldots, X_n \mid \Theta = \theta &\overset{\text{iid}}{\sim} Ber(\theta) \\ \Theta &\sim Be(\alpha, \beta). \end{aligned}$$

The first line in this hierarchy includes the assumption that $X_1, \ldots, X_n$ are conditionally independent, instead of independent. This implies that $X_1, \ldots, X_n$ are exchangeable: an assumption that is considerably weaker than independence. The second line in this hierarchy specifies the **prior distribution** for Θ. The need for providing this distribution is often seen as a weak aspect of Bayesian statistics by frequentists (the idea being that one can subjectively choose a prior and "anything" can result from that). On the opposite, Bayesians consider it to be a key advantage of the Bayesian approach, enabling to incorporate available prior knowledge. Regarding choice of prior, here we have chosen the $Be(\alpha, \beta)$-distribution as prior as this turns out to simplify calculations. Once the joint distribution of (X, Θ) has been specified, Bayesian statistics is conceptually straightforward ("conceptually", as there may be computationally demanding problems remaining): all inference should be based on the **posterior distribution**, which is the distribution of Θ conditional on X. We have

$$f_{\Theta|X}(\theta \mid x) = \frac{f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)}{\int f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)\, \mathrm{d}\theta}.$$

It turns out that $\Theta \mid X \sim Be(\alpha + S, \beta + n - S)$. A point estimator can then for example be defined by the posterior mean

$$\mathbb{E}[\Theta \mid X] = \frac{\alpha + S}{\alpha + \beta + n}.$$

It is interesting to see that $\hat{\Theta}_{MLE} = \hat{\Theta}_{UMVU} = \bar{X}_n$ is obtained upon letting both α and β tend to 0. Moreover, the minimax estimator $\hat{\Theta}_{MM}$ is obtained by taking $\alpha = \beta = \sqrt{n}/2$. Obviously, if we would have taken another prior distribution, the posterior would change accordingly.

Exercise 1.8 Verify that $\Theta \mid X \sim Be(\alpha + S, \beta + n - S)$.

Whereas the various approaches to estimation are fundamentally different, asymptotically they all agree that $\bar{X}_n$ is a good estimator for θ. The Bernstein-Von Mises theorem essentially states that this is even the case for all prior distributions with density that is strictly positive on $(0, 1)$.

If it comes to hypothesis testing, differences between the frequentist and Bayesian approach become more pronounced. An entertaining and accessible introduction to Bayesian statistics is the article by Lindley and Phillips [1976].

1.5.2 A few intriguing examples

Example 1.27. Suppose we wish to estimate θ based on $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N\left(\theta, \sigma^2\right)$. The statistical experiment is as follows: first a coin is tossed, if it lands heads sample size $n = 2$ is taken, else $n = 1000$ is taken. This is an example of a mixture experiment, as discussed in example 1.23. It is decided to estimate θ by $\hat{\Theta} := \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. As a measure of accuracy, we compute the variance of this estimator. We have

$$\operatorname{Var} \hat{\Theta} = \frac{1}{2}\left(\frac{\sigma^2}{2} + \frac{\sigma^2}{1000}\right) \approx \frac{\sigma^2}{4}.$$

Now would you believe $\sigma^2/4$ is a measure of accuracy if the coin toss resulted in tails? Few statisticians would feel so, but strictly speaking the derivation is correct from a frequentist point of view.

Example 1.28. This is a well known example (see for instance Berger [1985]). We follow the exposition in chapter 1 of Parmigiani and Inoue [2009]. Suppose you have to guess a secret number. You know it is an integer. You can perform an experiment that would yield either the number before it or the number after it, with equal probability. You perform this experiment twice and get the numbers 41 and 43. What is the secret number? Of course you know it is 42, and no statistician would argue about this. If it comes to quantifying uncertainty about this estimate, there is a disagreement however.

Mathematically, we can model the outcomes $x_1 = 41$ and $x_2 = 43$ as realisations of independent random variables X_1 and X_2, where $(i = 1, 2)$

$$X_i = \begin{cases} \theta + 1 & \text{with probability } 1/2 \\ \theta - 1 & \text{with probability } 1/2 \end{cases}.$$

The estimator that led to the result that the secret number ought to be 42 is given by

$$\hat{\Theta} = \frac{X_1 + X_2}{2}.$$

Consider the confidence set for θ given by

$$C(X_1, X_2) = \begin{cases} (X_1 + X_2)/2 & \text{if } X_1 \neq X_2 \\ X_1 - 1 & \text{if } X_1 = X_2 \end{cases}.$$

It is easily verified that

$$\mathbb{P}(C(X_1, X_2) \text{ contains } \theta \mid X_1 \neq X_2) = 1$$
$$\mathbb{P}(C(X_1, X_2) \text{ contains } \theta \mid X_1 = X_2) = 0.5$$

Hence, the set C has coverage 0.75:

$$\mathbb{P}(C(X_1, X_2) \text{ contains } \theta) = 0.75.$$

Within classical statistics, this is the confidence set that is to be reported. However, some statisticians find it rather silly to report this confidence set: you are absolutely sure about the secret number when $x_1 \neq x_2$ is observed.

The following example is known as Stein's paradox and dates back to 1956 (Stein [1956]).

Example 1.29. Suppose $X_1, X_2, X_3 \overset{\text{ind}}{\sim} N(\theta, 1)$. As the likelihood equals

$$L(\theta; X) = (2\pi)^{-3/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{3}(X_i - \theta)^2\right)$$

we find that the maximum likelihood estimator (MLE) equals $\hat{\Theta}_{MLE} = \bar{X}_3$. In the chapter on statistical decision theory we will see that this estimator has many favourable properties, including that it is

- admissible;
- minimax: it is the estimator that minimises $\sup_\theta \mathrm{E}_\theta \|\hat{\Theta} - \theta\|^2$ over all estimators $\hat{\Theta}$.
- asymptotically efficient.

The meaning of these concepts is explained in later chapters.

Now consider the slightly different problem where $X_i \overset{\text{ind}}{\sim} N(\theta_i, 1)$, $1 \leq i \leq 3$. Note that each X_i is now assumed to have its own mean θ_i. In this case the likelihood equals

$$L(\theta; X) = (2\pi)^{-3/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{3}(X_i - \theta_i)^2\right),$$

where $\theta = (\theta_1, \theta_2, \theta_3)$. Hence the maximum likelihood estimator equals $\hat{\Theta}_{MLE} = X = (X_1, X_2, X_3)$. The astonishing fact is that the "James-Stein estimator"

$$\hat{\Theta}_{JS} = \left(1 - \frac{1}{\|X\|}\right) X$$

satisfies

$$\mathrm{E}_\theta \left\|\hat{\Theta}_{JS} - \theta\right\|^2 < \mathrm{E}_\theta \left\|\hat{\Theta}_{MLE} - \theta\right\|^2$$

for all θ. That is, the estimator $\hat{\Theta}_{JS}$ improves upon the MLE. Put differently, the MLE is inadmissible! Convince yourself that this is somewhat counter intuitive: for estimating the mean of X_j, we use all observations $\{X_i\}$, while at the same time all $\{X_i\}$ are assumed independent.

Example 1.30. Consider the statistical model where $Y_i \sim Bin(n_i, p_i)$. Efron and Norris [1975] considered the example where i indexes baseball players and Y_i denotes the number of homeruns out of n_i times at bat by the i-th player. If you are not into sports, you may also think of i indexing hospitals and Y_i being the number of successful operations out of n_i operations in the i-th hospital. A straightforward estimator for p_i is given by $\hat{P}_i = Y_i/n_i$. Now just as in the previous example, it turns out that better estimators can be found if we are interested in the *aggregate* performance of the estimators $\{\hat{P}_i\}$.

We can transform the problem to the setting of the preceding example by applying a **variance stabilising transformation**. Define

$$X_i = \sqrt{n_i} \arcsin\left(2\frac{Y_i}{n_i} - 1\right),$$

then, for n_i large, we have that X_i has approximately the $N\left(\mu_i, 1\right)$ distribution with

$$\mu_i = \sqrt{n_i} \arcsin(2p_i - 1).$$

Cf. Exercise 1.11 ahead. We will return to this example when discussing hierarchical Bayesian models and Markov Chain Monte Carlo methods.

1.6 Stochastic convergence

Though this is not a course on asymptotic statistics, occasionally we will deal with limits of sequences of random quantities $\{X_n\}_{n=1}^{\infty}$. There exist various notions for defining convergence of the sequence to a limiting random quantity X. We only give the most important definitions.

Definition 1.31. A sequence of random vectors $\{X_n\}$ is said to **converge in distribution** to a random vector X if

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x), \qquad n \to \infty$$

for every x at which the limit distribution function $x \mapsto \mathbb{P}(X \leq x)$ is continuous. This is denoted by $X_n \rightsquigarrow X$.

Alternative names are **weak convergence** or **convergence in law**. The "Portmanteau theorem" (lemma 2.2 in van der Vaart [1998] includes equivalence of convergence in distribution to

$$\mathbb{E}f(X_n) \to \mathbb{E}f(X) \quad \text{for all bounded continuous functions } f.$$

Let d be a distance function on $\mathbb{R}^k$ that generates the usual topology, for instance Euclidean distance.

Definition 1.32. A sequence of random vectors X_n is said to **converge in probability** to X if for all $\varepsilon > 0$

$$\mathbb{P}(d(X_n, X) > \varepsilon) \to 0, \qquad n \to \infty.$$

This is denoted $X_n \xrightarrow{\text{p}} X$.

Hence, $X_n \xrightarrow{\text{p}} X$ is equivalent to $d(X_n, X) \xrightarrow{\text{p}} 0$.

Definition 1.33. A sequence of random vectors X_n is said to **converge almost surely** to X if

$$\mathbb{P}\left(\lim_{n\to\infty} d(X_n, X) = 0\right) = 1.$$

This is denoted $X_n \xrightarrow{\text{a.s.}} X$.

We have

$$X_n \xrightarrow{\text{a.s.}} X \quad \text{implies} \quad X_n \xrightarrow{\text{p}} X \quad \text{implies} \quad X_n \rightsquigarrow X.$$

Exercise 1.9 Let $U \sim Unif\,(0,1)$ and define

$$X_n = \begin{cases} U & \text{if } n \text{ if even} \\ 1-U & \text{if } n \text{ if odd} \end{cases}.$$

Show that the sequence $\{X_n\}_n$ converges in distribution, but not in probability.

Exercise 1.10 Suppose random variables $\{X_n\}$ are defined on $([0,1], \mathcal{B}([0,1]), \lambda)$ (where λ is Lebesgue measure) as follows

$$\begin{aligned} &X_1(\omega) = \mathbf{1}_{[0,1]}(\omega) \\ &X_2(\omega) = \mathbf{1}_{[0,1/2]}(\omega) \qquad X_3 = \mathbf{1}_{[1/2,1]}(\omega) \\ &X_4(\omega) = \mathbf{1}_{[0,1/3]}(\omega) \qquad X_5 = \mathbf{1}_{[1/3,2/3]}(\omega) \qquad X_6 = \mathbf{1}_{[2/3,1]}(\omega) \\ &\vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots \end{aligned}$$

1. Verify that $X_n \xrightarrow{\text{p}} 0$.
2. Verify that for any $\omega \in [0,1]$, $X_n(\omega) \nrightarrow 0$.

For future reference, we state the following results.

Lemma 1.34. If $X_n \rightsquigarrow X$ and $Y_n \xrightarrow{\text{p}} c$ for a constant c, then $(X_n, Y_n) \rightsquigarrow (X, c)$.

Theorem 1.35 (Continuous mapping theorem)**.** Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $\mathbb{P}(X \in C) = 1$.

1. If $X_n \rightsquigarrow X$, then $g(X_n) \rightsquigarrow g(X)$;
2. If $X_n \xrightarrow{\text{p}} X$, then $g(X_n) \xrightarrow{\text{p}} g(X)$;
3. If $X_n \xrightarrow{\text{a.s.}} X$, then $g(X_n) \xrightarrow{\text{a.s.}} g(X)$.

Suppose the statistician gets to observe random variables $X_1, \ldots, X_n$ defined on $\mathcal{X}$ with distribution P_θ. Statistical estimation is concerned with choosing the correct value of θ. An estimator is a (measurable) function of $X_1, \ldots, X_n$, say

$$\hat{\Theta}_n \equiv \hat{\Theta}_n(X_1, \ldots, X_n).$$

The estimator is called **(weakly) consistent** if $\hat{\Theta}_n$ converges in P_θ-probability to θ. It is called **strongly consistent** if $\hat{\Theta}_n$ converges P_θ-almost surely to θ.

The following theorem is usually referred to as the **Delta-method**. It shows how the limit law of $g(\Theta_n) - g(\theta)$ can be derived from that of $\Theta_n - \theta$.

Theorem 1.36. Let $g : D \subset \mathbb{R}^k \mapsto \mathbb{R}^m$ be a map defined on a subset of $\mathbb{R}^k$ and differentiable at θ. Let Θ_n be random vectors taking their value in the domain of g. If $r_n(\Theta_n - \theta) \rightsquigarrow T$ for numbers $r_n \to \infty$, then $r_n(g(\Theta_n) - g(\theta)) \rightsquigarrow g'_\theta(T)$.

Note that T is a random vector on $\mathbb{R}^k$ and g'_θ maps T to a random vector in $\mathbb{R}^m$. For $h \in \mathbb{R}^k$, the derivative map $h \mapsto g'_\theta(h)$ is just matrix multiplication:

$$g'_\theta(h) = \begin{bmatrix} \partial_1 g_1(\theta) & \ldots & \partial_k g_1(\theta) \\ \vdots & & \vdots \\ \partial_1 g_m(\theta) & \ldots & \partial_k g_m(\theta) \end{bmatrix} h.$$

The proof can for example be found in section 3.1 of van der Vaart [1998]. We end with a couple of applications of the Delta-method.

Example 1.37. Suppose $X_1, \ldots, X_n$ are iid with expectation θ and variance σ^2. By the central limit theorem

$$\sqrt{n}\left(\overline{X}_n - \theta\right) \rightsquigarrow N\left(0, \sigma^2\right).$$

By the Delta-method we immediately obtain the limiting distribution of $\overline{X}_n^2$

$$\sqrt{n}\left(\overline{X}_n^2 - \theta^2\right) \rightsquigarrow 2\theta N\left(0, \sigma^2\right) \sim N\left(0, 4\theta^2\sigma^2\right).$$

Example 1.38. Another application is that of deriving a *variance stabilising transformation*. As an example, suppose each $X_i \sim Pois\,(\theta)$, then $\sigma^2 = \theta$. Applying the Delta-method with $g(\theta) = 2\sqrt{\theta}$ yields

$$2\sqrt{n}\left(\sqrt{\bar{X}_n} - \sqrt{\theta}\right) \rightsquigarrow \frac{1}{\sqrt{\theta}} N\left(0, \theta\right) \sim N\left(0, 1\right).$$

As the variance in the limit is independent of θ, the transformation is called variance stabilising. The result in the preceding display can be used for deriving an asymptotic confidence interval for θ.

Writing $\sigma^2 = \sigma^2(\theta)$, we see that the transformation g is obtained as

$$g(\theta) = \int_0^\theta \frac{1}{\sigma(x)} dx.$$

Example 1.39. In the multivariate case, by the central limit theorem, we obtain that (under weak moment-assumptions)

$$\sqrt{n}\left([\overline{X}_n, \overline{X_n^2}] - \mu\right) \rightsquigarrow N\left(0, \Sigma\right).$$

Here μ and 0 are vectors in $\mathbb{R}^2$ and Σ a 2×2 matrix. From this result the limiting distribution of $S_n^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ can be derived using the Delta-method by noting that $S_n^2 = g(\overline{X}_n, \overline{X_n^2})$ with $g(x_1, x_2) = x_2 - x_1^2$.

Exercise 1.11 Verify the claim from example 1.30: Let $Y \sim Bin\,(n, p)$ and define

$$X_n = \sqrt{n}\arcsin\left(2\frac{Y}{n} - 1\right).$$

Show that for n large, X_n has approximately the $N\left(\mu_n, 1\right)$ distribution with

$$\mu_n = \sqrt{n}\arcsin(2p - 1).$$

Hint: first note that if $Y \sim Bin\,(n, p)$ then by the central limit theorem:

$$\sqrt{n}(Y/n - p) \rightsquigarrow N\left(0, p(1-p)\right), \quad n \to \infty.$$

Next, take g such that $g'(x) = 1/\sqrt{x(1-x)}$.

Chapter 2

Topics from classical statistics

It is assumed that you have attended an undergraduate course on mathematical statistics covering topics such as sufficiency, maximum likelihood estimation, confidence intervals and Neyman-Pearson testing. Here, we review these concepts briefly. Most concepts should already be familiar to you.

2.1 Sufficiency

A statistic T is a measurable function of the data X. With slight abuse of notation, we often denote $T(X)$ by T, so that T is both used to denote the mapping and the random quantity $T(X)$. The value that T assumes when $X = x$ is denoted by $T(x)$. A sufficient statistic is a statistic with a special property.

Definition 2.1. Let X be a sample from an unknown probability measure in the set $\{\mathrm{P}_\theta,\ \theta \in \Omega\}$. A statistic T is said to be **sufficient** for $\theta \in \Omega$ if the conditional distribution of X given T is known (does not depend on θ).

Whence we observe X and compute a sufficient statistic $T(X)$, the original data X do not contain any further information concerning the unknown θ. In case $X \sim \mathrm{P}_\theta$ has a density with respect to counting measure), we have

$$\mathbb{P}(X = x; \theta) = \sum_t \mathbb{P}(X = x \mid T = t)\mathbb{P}(T = t; \theta). \tag{2.1}$$

This illustrates that we can sample X by first using θ to sample T, and next sample X conditional on T. Note that $\{T = t\}$ has probability zero for "continuous" random variables, and in that case measure theory is required to give meaning to Equation (2.1). This is further complicated upon noticing that $(X, T(X))$ "lives on the diagonal $\{(x, T(x)),\ x \in \mathcal{X}\}$ (hence, if one wants to talk about densities the dominating measure will not be a product-measure).

Finding sufficient statistics from the definition tends to be complicated. The following theorem simplifies this task.

Theorem 2.2 (**Factorisation theorem**)**.** Suppose that X is a sample from P_θ, where $\{\mathrm{P}_\theta,\ \theta \in \Omega\}$ is a family of probability measures on $(\mathbb{R}^n, \mathcal{B}^n)$ dominated by a σ-finite measure ν. Then T is sufficient for $\theta \in \Omega$ if and only if there are nonnegative Borel functions h (which does not depend on P_θ) on $(\mathbb{R}^n, \mathcal{B}^n)$ and g_θ on the range of T such that

$$\frac{\mathrm{dP}_\theta}{\mathrm{d}\nu}(x) = g_\theta(T(x))h(x).$$

Proof. We give the proof in case X is discrete. One direction is easy: suppose T is sufficient for θ, then

$$\begin{aligned}\mathbb{P}(X = x; \theta) &= \mathbb{P}(X = x, T = T(x); \theta) \\ &= \mathbb{P}(T = T(x); \theta)\mathbb{P}(X = x \mid T = T(x)).\end{aligned}$$

Hence take $g_\theta(T(x)) = \mathbb{P}(T = T(x); \theta)$ and $h(x) = \mathbb{P}(X = x \mid T = T(x))$.

For the other direction, suppose the functions g_θ and h exist. Then

$$\mathbb{P}(X = x_0 \mid T = t; \theta) = \frac{\mathbb{P}(X = x_0, T = t; \theta)}{\mathbb{P}(T = t; \theta)}.$$

If $T(x_0) \neq t$ this probability equals 0. If $T(x_0) = t$, then it is equal to

$$\frac{\mathbb{P}(X = x_0; \theta)}{\mathbb{P}(T = t; \theta)} = \frac{g_\theta(T(x_0))h(x_0)}{\mathbb{P}(T = t; \theta)}.$$

Furthermore,

$$\begin{aligned}\mathbb{P}(T = t; \theta) &= \sum_{\{x:\, T(x)=t\}} \mathbb{P}(X = x; \theta) \\ &= \sum_{\{x:\, T(x)=t\}} g_\theta(T(x))h(x) = g_\theta(t) \sum_{\{x:\, T(x)=t\}} h(x).\end{aligned}$$

Combining the preceding 3 displays, we obtain that if $T(x_0) = t$

$$\mathbb{P}(X = x_0 \mid T = t; \theta) = \frac{h(x_0)}{\sum_{\{x:\, T(x)=t\}} h(x)}$$

which does not depend on θ. □

In the continuous case, the proof is much harder due to measure-theoretic technicalities.

There are many sufficient statistics for a family $\{\mathrm{P}_\theta,\ \theta \in \Omega\}$. If T is a sufficient statistic and $T = h(S)$, where h is measurable and S is another statistic, then S is sufficient. This follows trivially from the factorisation theorem. This motivates the following definition.

Definition 2.3. Let $T : \mathcal{X} \to \mathcal{T}$ be a sufficient statistic for $\theta \in \Omega$. T is called a **minimal sufficient** statistic if, for any other statistic $S : \mathcal{X} \to \mathcal{S}$ that is sufficient for $\theta \in \Omega$, there is a measurable function $h : \mathcal{S} \to \mathcal{T}$ such that $T = h(S)$, P_θ-a.s. for all θ.

Example 2.4. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$. The following are all sufficient statistics

$$T_1(X) = (X_1, \ldots, X_n)$$

$$T_2(X) = (X_1^2, \ldots, X_n^2)$$

$$T_3(X) = \sum_{i=1}^{n} X_i^2.$$

The amount of reduction increases from T_1 to T_3. In fact, $T_3(X)$ is minimal (and for sure it can be written as a function of either $T_1(X)$ or $T_2(X)$).

Lemma 2.5. Suppose S and T are both minimal sufficient statistics for $\theta \in \Omega$. Then there exists a function h, which is injective on the range of S, such that $T = h(S)$ P_θ-a.s.

Proof. Since T is minimal sufficient, there exists a measurable function h such that $T = h(S)$. Similarly there exists a measurable function $\tilde{h}$ such that $S = \tilde{h}(T)$. Suppose x and x' are such that $h(S(x)) = h(S(x'))$. This means that $T(x) = T(x')$ which implies

$$S(x) = \tilde{h}(T(x)) = \tilde{h}(T(x')) = S(x').$$

This proves that h is injective on the range of S. □

Hence the minimal sufficient statistic is unique in the sense that two statistics that are one-to-one functions of each other can be treated as one statistic. Establishing that a sufficient statistic is minimal sufficient can be hard. One useful criterion is the following

Lemma 2.6. Suppose for each $\theta \in \Omega$ that P_θ has density $f(x;\theta) = g_\theta(T(x))h(x)$ with respect to a dominating measure ν. If $f(x;\theta) = cf(y;\theta)$ for some $c = c(x,y)$ implies $T(x) = T(y)$, then T is minimal sufficient.

Proof.* We give a proof by contradiction. The proof consists of a few steps:

- If T is minimal sufficient, this means that for any sufficient statistic T^* we have a mapping r such that $T = r(T^*)$. In particular, for all x, y, if $T^*(x) = T^*(y)$, then $T(x) = T(y)$.
- Now suppose T is not min. sufficient. Then there exist x, y such that $T^*(x) = T^*(y)$, but $T(x) \neq T(y)$.
- Use the factorisation theorem:

$$f(x;\theta) = g_\theta(T^*(x))h(x) = g_\theta(T^*(y))h(x) = \frac{h(x)}{h(y)} f(y;\theta).$$

- By the assumption in the theorem, this implies $T(x) = T(y)$. Hence we have reached a contradiction.

□

Another way for establishing is to show that a sufficient statistic is complete and apply the Lehmann-Scheffé theorem. Details are in Section 2.8

Exercise 2.1 Let ϕ be a positive (Borel) function $\mathbb{R}$ such that $\int_a^b \phi(x)\,\mathrm{d}x < \infty$ for a pair $\theta = (a,b)$ with $-\infty < a < b < \infty$. Let $\Omega = \{\theta = (a,b) \in \mathbb{R}^2 \ : \ a < b\}$. Define

$$f(x;\theta) = c(\theta)\phi(x)\mathbf{1}_{(a,b)}(x),$$

with $c(\theta)$ such that $\int f(x;\theta)\,\mathrm{d}x = 1$. Then $\{f(\cdot;\theta),\ \theta \in \Omega\}$ is called a **truncation family**. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(\cdot;\theta)$. Let $X = (X_1, \ldots, X_n)$. Show that $T(X) = (X_{(1)}, X_{(n)})$ is sufficient.

2.1.1 Exponential families

Suppose $X_1, \ldots, X_n$ are IID and each is within the exponential family class:

$$f_{X_j}(x_j;\theta) = c(\theta)h(x_j)\exp\left(\sum_{i=1}^{k} \xi_i(\theta)\tau_i(x_j)\right).$$

If $X = (X_1, \ldots, X_n)$ and $x = (x_1, \ldots, x_n)$, then

$$f_X(x;\theta) = c(\theta)^n \left\{\prod_{j=1}^{n} h(x_j)\right\} \exp\left(\sum_{i=1}^{k} \xi_i(\theta)\left[\sum_{j=1}^{n} \tau_i(x_j)\right]\right).$$

Hence, if we define $t_i(X) = \sum_{j=1}^n \tau_i(X_j)$, then $T(X) = (t_1(X), \dots, t_k(X))$ is sufficient for θ (as a consequence of the factorisation theorem). This statistic is sometimes called the **natural sufficient statistic**.

Lemma 2.7. If X has an exponential family distribution, then so does the natural sufficient statistic $T(X)$, and the natural parameter for T is the same as for X.

Proof. We only give the proof in the "discrete setting". Fix a vector $y = (y_1, \dots, y_k)$ and let

$$\mathcal{I}_y = \{x : t_1(x) = y_1, \dots, t_k(x) = y_k\}.$$

Then

$$\begin{aligned}
\mathbb{P}(t_1(X) = y_1, \dots, t_k(X) = y_k; \theta) &= \sum_{x \in \mathcal{I}_y} \mathbb{P}(X = x; \theta) \\
&= c(\theta)^n \sum_{x \in \mathcal{I}_y} \left\{ \prod_{j=1}^n h(x_j) \right\} \exp\left(\sum_{i=1}^k \xi_i(\theta) t_i(x) \right) \\
&= c(\theta)^n h_0(y) \exp\left(\sum_{i=1}^k \xi_i(\theta) y_i \right),
\end{aligned}$$

where $h_0(y) = \sum_{x \in \mathcal{I}_y} \left\{ \prod_{j=1}^n h(x_j) \right\}$. □

Theorem 2.8. If the natural parameter space Ω of an exponential family contains an open set in $\mathbb{R}^k$, then the natural sufficient statistic $T(X)$ is complete and sufficient.

A proof is given in Chapter 2 of Schervish [1995].

As examples, the sufficient statistics from normal, exponential, Poisson and Bernoulli distributions are complete.

Exercise 2.2 [YS exercise 6.2.] Find a minimal sufficient statistic for θ based on an independent sample of size n from each of the following distributions:

1. $Ga(\alpha, \beta)$ (with $\theta = (\alpha, \beta)$);
2. $Unif(\theta - 1, \theta + 1)$;
3. (*, advanced) the Cauchy distribution with density

$$f(x; \alpha, \beta) = \frac{\beta}{\pi[(x - \alpha)^2 + \beta^2]}, \qquad x \in \mathbb{R}$$

with $\alpha \in \mathbb{R}$ $\beta > 0$ and $\theta = (\alpha, \beta)$. *Hint: a first try would be to see if the density is within the Exponential family, however, this is not the case. Argue from the definition of minimal sufficiency why the vector of order-statistics, $(X_{(1)}, \dots, X_{(n)})$ is minimal sufficient.*

2.2 Information measures

In this section we discuss two measures that are known as information measures. The qualification "information" is a bit misleading, as these measures do not depend on the data but rather on their assumed distribution. From their definition it is not immediately clear why these measures are of any use, but we will see their applicability for example when discussing the Crámer-Rao bound and asymptotic properties of maximum likelihood estimators and Bayesian point estimators.

2.2.1 Fisher information

Definition 2.9. Suppose X has density $f_X(\cdot;\theta)$ with respect to ν and $\theta \in \Omega \subset \mathbb{R}^k$. The following conditions will be referred to as the **FI regularity conditions**.

1. There exists B with $\nu(B) = 0$ such that for all θ, $\partial f_X(x;\theta)/\partial\theta_i$ exists for $x \notin B$ and each i.
2. $\int f_X(x;\theta)\,\mathrm{d}\nu(x)$ can be differentiated under the integral sign with respect to each coordinate of θ.
3. The set $C = \{x : f_X(x;\theta) > 0\}$ is the same for all θ.

The third condition rules out the $Unif\,(0,\theta)$-distribution for example, which is an important counterexample to keep in mind.

Definition 2.10. Assume the FI regularity conditions hold. The **Fisher information matrix** $I(\theta;X)$ about θ based on X is defined as the matrix with elements

$$I_{i,j}(\theta) = \mathrm{Cov}_\theta\left(s_i(\theta;X), s_j(\theta;X)\right).$$

The vector $s(\theta;x)$ with coordinates

$$s_i(\theta;x) = \frac{\partial}{\partial\theta_i}\log f_X(x;\theta)$$

is called the **score function**. That is $s(\theta;x) = \nabla_\theta \log f_X(x;\theta)$.

Example 2.11. If $X \sim Pois\,(\theta)$, then $I(\theta) = 1/\theta$.

Example 2.12. If $X \sim N\left(\mu,\sigma^2\right)$. If $\theta = (\mu,\sigma)$, then

$$I(\theta) = \begin{bmatrix}\sigma^{-2} & 0\\ 0 & 2\sigma^{-2}\end{bmatrix}.$$

Example 2.13. For a scale family we have $f_X(x;a) = a^{-1}f\left(\frac{x}{a}\right)$ (where f is a fixed density). This implies

$$I(a) = \frac{1}{a^2}\int_0^\infty \left(1 + \frac{uf'(u)}{f(u)}\right)^2 f(u)\,\mathrm{d}u.$$

> **Exercise 2.3** Verify examples 2.11, 2.12 and 2.13.

Under FI regularity conditions we can find a different representation for the Fisher information.

Lemma 2.14. Assume the FI regularity conditions are satisfied. Then

$$I_{i,j}(\theta) = -\,\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_X(X;\theta)\right].$$

Proof. First note that

$$\begin{aligned}\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta_i}\log f_X(X;\theta)\right] &= \int\left[\frac{\partial}{\partial\theta_i}\log f_X(x;\theta)\right] f_X(x;\theta)\,\mathrm{d}x\\ &= \int\frac{\partial}{\partial\theta_i}f_X(x;\theta)\,\mathrm{d}x = \frac{\partial}{\partial\theta_i}\int f_X(x;\theta)\,\mathrm{d}x = 0\end{aligned}$$

when the order of differentiation and integration can be interchanged.

Suppose we can differentiate twice under the integral sign (this is valid for example in exponential families), then

$$0 = \int \frac{\partial^2}{\partial\theta_i \partial\theta_j} f_X(x;\theta)\, \mathrm{d}\nu(x) = \mathrm{E}_\theta \left[\frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j} f_X(X;\theta)}{f_X(X;\theta)} \right].$$

Upon taking the expectation on both sides of the relation

$$-\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log f_X(X;\theta) = -\frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j} f_X(X;\theta)}{f_X(X;\theta)} + \frac{\left(\frac{\partial}{\partial\theta_i} f_X(X;\theta)\right)\left(\frac{\partial}{\partial\theta_j} f_X(X;\theta)\right)}{f_X^2(X;\theta)}$$

we find that

$$-\mathrm{E}_\theta \left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log f_X(X;\theta) \right] = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i} \log f_X(X;\theta) \right) \left(\frac{\partial}{\partial\theta_j} \log f_X(X;\theta) \right) \right].$$

□

The following lemma states that Fisher information is additive in case of independent random quantities.

Lemma 2.15. Suppose $X_1, \ldots, X_n$ are independent and the Fisher information is $I_i(\theta)$ for each X_i. If $X = (X_1, \ldots, X_n)$ then the Fisher information $I_{1:n}(\theta)$ of X is given by

$$I_{1:n}(\theta) = \sum_{i=1}^{n} I_i(\theta).$$

Exercise 2.4 Prove lemma 2.15.

Exercise 2.5 Suppose

$$f_X(x;\pi) = c(\pi)h(x) \exp\left(\sum_{i=1}^{k} \pi_i \tau_i(x) \right)$$

(an exponential family in natural parametrisation). Show that

$$I_{i,j}(\pi) = -\frac{\partial^2}{\partial\pi_i\partial\pi_j} \log c(\pi).$$

FI under reparametrisation

If we change the parametrisation of a statistical model, then its Fisher information changes accordingly. That is, Fisher information is not invariant under reparametrisation. A simple example illustrates this.

Example 2.16. Suppose that $X \sim Exp(\theta)$. Then it is easy to verify that $I(\theta) = 1/\theta^2$. Now suppose we use a different parametrisation

$$\psi = g(\theta) = 1/\theta^2 \quad \Longleftrightarrow \quad \theta = g^{-1}(\psi) = 1/\sqrt{\psi},$$

and let Y be a random variable with density

$$f_Y(y;\psi) := f_X(y; g^{-1}(\psi)) = \psi^{-1/2} e^{-y/\sqrt{\psi}}, \qquad \psi \in (0,\infty).$$

While maybe less convenient, a parametrisation in terms of ψ yields the same model as under θ. A straightforward calculation shows that $I(\psi) = 1/(4\psi^2)$. Now we see that notation is tricky here. Let's consider $X \sim Exp(\theta)$ as parametrisation 1, and $Y \sim Exp\left(1/\sqrt{\psi}\right)$ as parametrisation 2. Then we write $I_1(\theta) = \theta^{-2}$ and $I_2(\psi) = 1/(4\psi^2)$. The point we want to make in this example is that $I_2(\psi)$ cannot be obtained from $I_1(\theta)$ by simply plugging in the relation between ψ and θ, i.e. $\theta = 1/\sqrt{\psi}$. This is what is meant by "lack of invariance under reparametrisation".

Lemma 2.17. Let $g : \mathbb{R} \to \mathbb{R}$ be a one-to-one, onto differentiable mapping with differentiable inverse. Assume $X \sim f_X(\cdot;\theta)$, where θ is one-dimensional. Let $\psi = g(\theta)$ and $Y \sim f_Y(\cdot;\psi) := f_X(\cdot; g^{-1}(\psi))$. Then

$$I_2(\psi) = I_1(g^{-1}(\psi)) \left| \frac{\mathrm{d}}{\mathrm{d}\psi} g^{-1}(\psi) \right|^2.$$

Proof. As we assume FI regularity conditions, the score function has mean zero and hence

$$\begin{aligned} I_2(\psi) &= \mathrm{Var}_\psi \left(\frac{\mathrm{d}}{\mathrm{d}\psi} \log f_Y(Y;\psi) \right) = \mathrm{E}_\psi \left[\left(\frac{\mathrm{d}}{\mathrm{d}\psi} \log f_Y(Y;\psi) \right)^2 \right] \\ &= \int \left(\frac{\mathrm{d}}{\mathrm{d}\psi} \log f_Y(y;\psi) \right)^2 f_Y(y;\psi)\,\mathrm{d}y \\ &= \int \left(\frac{\mathrm{d}}{\mathrm{d}\psi} \log f_X(y; g^{-1}(\psi)) \right)^2 f_X(y; g^{-1}(\psi))\,\mathrm{d}y \end{aligned}$$

By the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}\psi} \log f_X(y; g^{-1}(\psi)) = \left. \frac{\mathrm{d}}{\mathrm{d}\theta} \log f_X(y;\theta) \right|_{\theta = g^{-1}(\psi)} \left| \frac{\mathrm{d}}{\mathrm{d}\psi} g^{-1}(\psi) \right|.$$

Combining the previous two displays gives

$$I_2(\psi) = \left| \frac{\mathrm{d}}{\mathrm{d}\psi} g^{-1}(\psi) \right|^2 \left. \int \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_X(y;\theta) \right)^2 f_X(y;\theta)\,\mathrm{d}y \right|_{\theta = g^{-1}(\psi)}$$

which is exactly the stated result. □

We can write the statement of this lemma slightly shorter as

$$I_2(\psi) = I_1(\theta) \left| \frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right|^2.$$

Remark 2.18. If we replace $\log f_Y(y;\theta)$ by by an arbitrary smooth function $R(y;\theta)$, the lemma still holds.

2.2.2 Kullback-Leibler information

Suppose P and Q are two probability measures on a common space.

Definition 2.19. The **Kullback-Leibler (KL) information** of Q from P is defined by

$$KL(P,Q) = \int \log \frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}P$$

if P is absolutely continuous with respect to Q and the integral is meaningful, and $KL(P,Q) = \infty$ otherwise. In case P and Q have densities p and q with respect to a common measure ν this implies

$$KL(P,Q) = \int \log \frac{p(x)}{q(x)} p(x)\,\mathrm{d}\nu(x).$$

In case of parametric families, where $X \sim \mathrm{P}_\theta$ and $\mathrm{P}_\theta \ll \nu$ for all θ with densities $f_X(\cdot;\theta)$, we write

$$KL_X(\theta,\psi) := \mathrm{E}_\theta \log \frac{f_X(X;\theta)}{f_X(X;\psi)}, \qquad \theta,\psi \in \Omega.$$

This is the KL information of ψ from θ based on the data X.

To help understanding the interpretation of $KL(P,Q)$, consider the hypothesis testing problem

$$H_0 : X \sim Q \qquad H_1 : X \sim P.$$

The log of the likelihood-ratio-statistic is given by

$$\lambda(X) = \log \frac{\mathrm{d}P}{\mathrm{d}Q}(X)$$

If H_1 is true, we expect $\mathrm{E}_P\,\lambda(X) = KL(P,Q)$ to be large. This gives the interpretation that

$$\boxed{\text{the larger } KL(P,Q), \text{ the easier it is to distinguish } P \text{ from } Q}.$$

KL information is also known as KL distance though in general $KL(P,Q) \neq KL(Q,P)$ so that $KL(\cdot,\cdot)$ is not a true distance. If $KL(P,Q) = 0$ though, we can conclude that $P = Q$.

Lemma 2.20. $KL(P,Q) \geq 0$ with equality if and only if $P = Q$.

Proof. This is based on Jensen's inequality: as $x \mapsto \log(x)$ is strictly concave, we have

$$\mathrm{E}_P \log \frac{q(X)}{p(X)} \leq \log \mathrm{E}_P \frac{q(X)}{p(X)} = \log 1 = 0.$$

Therefore

$$KL(P,Q) = -\mathrm{E}_P \log \frac{q(X)}{p(X)} \geq 0.$$

As $x \mapsto \log x$ is *strictly* concave, we can only have equality if $P = Q$. □

Exercise 2.6 If $X \sim N(\theta,1)$, show that $KL_X(\theta,\psi) = (\theta-\psi)^2/2$. In this case $KL_X(\theta,\psi)$ is a multiple of squared Euclidean distance.

Exercise 2.7 If $X \sim Ber(\theta)$, show that

$$KL_X(\theta, \psi) = \theta \log \frac{\theta}{\psi} + (1-\theta) \log \frac{1-\theta}{1-\psi}.$$

Exercise 2.8 (Schervish [1995], exercise 45 in chapter 2.) Suppose that $X \sim Unif(0, \theta)$. Find the Kullback-Leibler information $KL_X(\theta_1, \theta_2)$ for all pairs (θ_1, θ_2).

Whereas Fisher information in general is a matrix, KL information is always a nonnegative number. It shares the additivity property of Fisher information under independence.

Lemma 2.21. If $X, Y \overset{\text{ind}}{\sim} \mathrm{P}_\theta$ then

$$KL_{X,Y}(\theta, \psi) = KL_X(\theta, \psi) + KL_Y(\theta, \psi).$$

Exercise 2.9 Prove lemma 2.21.

Contrary to Fisher information, KL information is invariant under reparametrisation.

Lemma 2.22. Suppose g is a bijective mapping. Suppose $X \sim f_X(\cdot; \theta)$ and Y is parametrised by $\theta' = g(\theta)$: $f_Y(\cdot; \theta') = f_X(\cdot; g^{-1}(\theta'))$. Then

$$KL_Y(\theta', \psi') = KL_X(\theta, \psi),$$

where $\psi' = g(\psi)$.

Proof. The proof is easy!

$$\begin{aligned} KL_Y(\theta', \psi') &= \int \log \frac{f_X(y; g^{-1}(\theta'))}{f_X(y; g^{-1}(\psi'))} f_X(y; g^{-1}(\theta'))\, \mathrm{d}y \\ &= \int \log \frac{f_X(y; \theta)}{f_X(y; \psi)} f_X(y; \theta)\, \mathrm{d}y = KL_X(\theta, \psi). \end{aligned}$$

□

Note that the lack of invariance to reparametrisation for Fisher-information is caused by the derivative in its definition and a consequence of the chain-rule.

2.3 Parameter estimation

2.3.1 Unbiased and minimum variance estimation

Definition 2.23. An estimator $\Theta \equiv \Theta(X)$ is **unbiased** for θ if $\mathrm{E}_\theta\, \Theta = \theta$.

This is an intuitively appealing property, but unbiased estimators do not exist in many cases or can be silly, as the following example illustrates.

Example 2.24. Suppose $X \sim Geom(\theta)$, with $\theta \in (0, 1)$. We seek for an unbiased estimator for θ. Since

$$\begin{aligned} \mathrm{E}_\theta\, d(X) &= \sum_{i=1}^{\infty} d(i)\theta(1-\theta)^{i-1} \\ &= d(1)\theta + d(2)\theta(1-\theta) + d(3)\theta(1-\theta)^2 + \ldots \end{aligned}$$

we obtain the following unbiased estimator for θ

$$d(X) = \begin{cases} 1 & \text{if } X = 1 \\ 0 & \text{if } X > 1 \end{cases}.$$

Now d is supposed to estimate $\theta \in (0, 1)$, but can only take the values $\{0, 1\}$.

Exercise 2.10 Suppose $X \sim Pois(\theta)$, then X is a complete sufficient statistic.

1. Verify that $\phi(X)$ is an unbiased estimator for $e^{-3\theta}$ if and only if

$$\sum_{k=0}^{\infty} \phi(k) e^{-\theta} \frac{\theta^k}{k!} = e^{-3\theta}.$$

2. Prove that $(-2)^X$ is unbiased for $e^{-3\theta}$. Do you think this is a sensible estimator?

Historically seen, a commonly advocated strategy by classical statisticians is the following. In the search of optimal estimators, choose the estimator with minimal variance among all unbiased estimators.

Definition 2.25. An estimator $\phi(X)$ for θ is called **UMVU (Uniformly Minimum Variance Unbiased)** if it is unbiased, has finite variance, and for every other unbiased estimator $\psi(X)$ of θ we have

$$\mathrm{Var}_\theta\, \phi(X) \leq \mathrm{Var}_\theta\, \psi(X) \quad \text{for all } \theta \in \Omega.$$

There are various problems with this kind of "optimality criterion". First of all, as we have just seen, unbiased estimators need not exist. Moreover, they may be very hard to derive. Second, what we really want is that some dissimilarity measure between $\phi(X)$ and θ is small. In statistical decision theory this is called a loss function. Say we take $L(\phi, \theta) = (\phi(X) - \theta)^2$. Then $\mathbb{E}_\theta L(\phi, \theta) = (\mathbb{E}_\theta - \theta)^2 + \mathrm{Var}_\theta\, \phi(X)$, the usual bias-variance trade-off of the mean square error. Searching for ϕ to minimise $\mathbb{E}_\theta L(\phi, \theta)$ is different than deriving a UMVU-estimator. By insisting on unbiasedness, we may in fact get bad estimators! So take-home message: unbiasedness, though intuitively appealing, is not at all an optimality criterion and shouldn't play a role in deriving estimators.

There are at least two strategies for deriving that a particular estimator is UMVU for θ (or possibly $g(\theta)$):

- Applying the the Lehmann-Scheffé theorem which states that an unbiased estimator that is a function of a complete sufficient statistic is UMVU. Details are in Section 2.8.
- Showing that the Cramér-Rao lower bound is attained. We state and prove this result below.

In certain cases a lower bound on the variance of unbiased estimators can be derived. The best known of these bounds is the **Cramér-Rao lower bound**. For simplicity we assume the parameter θ to be one-dimensional.

Theorem 2.26. Assume $\Omega \subset \mathbb{R}$ and let $\phi(X)$ be a one-dimensional statistic with $\mathrm{E}_\theta\, |\phi(X)| < \infty$ for all θ. Suppose the FI regularity conditions of definition 2.9 are satisfied, $I(\theta) > 0$ and also that $\int \phi(x) f_X(x; \theta) d\nu(x)$ can be differentiated under the integral sign with respect to θ. Then

$$\mathrm{Var}_\theta\, \phi(X) \geq \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \mathrm{E}_\theta\, \phi(X) \right)^2}{I(\theta)}.$$

Proof. Let $D = C \cap B^c$ from definition 2.9, so that for all θ, $\mathrm{P}_\theta(D) = 1$. We have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta}\,\mathrm{E}_\theta\,\phi(X) &= \frac{\mathrm{d}}{\mathrm{d}\theta}\int \phi(x) f_X(x;\theta)\,\mathrm{d}\nu(x)\\ &= \int \phi(x)\frac{\mathrm{d}}{\mathrm{d}\theta} f_X(x;\theta)\,\mathrm{d}\nu(x)\\ &= \int \phi(x) s(\theta;x) f_X(x;\theta)\,\mathrm{d}\nu(x) = \mathrm{E}_\theta\left[\phi(X)s(\theta;X)\right],\end{aligned}$$

where $s(\theta \mid x) = \frac{\mathrm{d}}{\mathrm{d}\theta}\log f_X(x \mid \theta)$. Upon taking $\phi \equiv 1$ we get $\mathrm{E}_\theta\, s(\theta \mid X) = 0$. Hence

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\,\mathrm{E}_\theta\,\phi(X) = \mathrm{E}_\theta\left[(\phi(X) - \mathrm{E}_\theta\,\phi(X))s(\theta;X)\right].$$

The Cauchy-Schwarz inequality [1] then gives

$$\left|\frac{\mathrm{d}}{\mathrm{d}\theta}\,\mathrm{E}_\theta\,\phi(X)\right| \le \left(\mathrm{E}_\theta\left[(\phi(X) - \mathrm{E}_\theta\,\phi(X))^2\right]\mathrm{E}_\theta\left[s(\theta;X)^2\right]\right)^{1/2}.$$

The result now follows, since the square of the right-hand-side equals $\mathrm{Var}_\theta\,\phi(X)\cdot I(\theta)$.

□

Exercise 2.11 If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N\left(\theta, \sigma^2\right)$, then verify that $\bar{X}_n$ is unbiased for θ and that the Cramér-Rao bound is met.

Now suppose $\phi(X)$ is an unbiased estimator for θ, i.e. $\mathrm{E}_\theta\,\phi(X) = \theta$. By the Cramér-Rao bound, the smallest possible variance for any unbiased estimator is $1/I(\theta)$ (as in this case $\frac{\mathrm{d}}{\mathrm{d}\theta}\,\mathrm{E}_\theta\,\phi(X) = 1$). It is a natural question if there exists an estimator that attains the bound.

Inspecting the proof of theorem 2.26, the lower bound is achieved when the $\le$ in the Cauchy-Schwarz' inequality is in fact an equality sign. This is only the case if $\phi(X) - \mathrm{E}_\theta\,\phi(X) = \phi(X) - \theta$ and $\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_X(X;\theta)$ are linearly related, which entails

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_X(x;\theta) = \lambda(\theta)(\phi(x) - \theta) \quad \text{for all } \theta$$

for a certain mapping $\lambda : \Omega \to \mathbb{R}$. This implies that f_X should satisfy

$$\log f_X(x;\theta) = A(\theta)\phi(x) + B(\theta) + C(x),$$

for some functions A, B and C. This is to be compared with a one-paremeter exponential family with a one-dimensional sufficient statistic $\tau(X)$:

$$f_X(x;\theta) = c(\theta)h(x)\exp\left(\xi(\theta)\tau(x)\right).$$

Hence, the Cramér-Rao lower bound can only be sharp if we are within the exponential family, and an UMVU estimator for θ only exists if $\mathrm{E}_\theta\,\tau(X) = \theta$.

[1] For vectors $x, y \in \mathbb{R}^n$, this inequality states that $|\langle x, y\rangle| \le \|x\|\|y\|$ with equality if and only if $y = \lambda x$ for some $\lambda \in \mathbb{R}$. The inequality holds much more generally (in L^2-spaces). A particular formulation is the following: if X and Y are random variables with $\mathrm{E}\,X = \mathrm{E}\,Y = 0$, $\mathrm{E}\,X^2 < \infty$ and $\mathrm{E}\,Y^2 < \infty$, then $|\,\mathrm{E}[XY]| \le \sqrt{\mathrm{E}\,X^2\,\mathrm{E}\,Y^2}$ with equality if and only if $Y = \lambda X$ for some $\lambda \in \mathbb{R}$.

2.3.2 Maximum likelihood estimation

Definition 2.27. Let X be a random quantity with density $f_X(\cdot;\theta)$, $\theta \in \Omega$. If $X = x$ is observed, the function $L(\theta; x) = f_X(x;\theta)$ considered as a function of θ for fixed x is called the **likelihood function**. Any random quantity $\hat{\Theta}$ such that

$$\max_{\theta\in\Omega} L(\theta; X) = L(\hat{\Theta}; X)$$

is called a **maximum likelihood estimator (MLE)** of θ.

The idea goes back to Fisher (1922). Maximum likelihood gives a simple principle for obtaining estimators by maximisation. Often we will simply write $L(\theta)$ instead of $L(\theta; X)$. The **loglilkelihood** is defined by $l(\theta; X) = \log L(\theta; X)$. The **score function** is defined by

$$s(\theta; X) = \nabla_\theta\, l(\theta; X).$$

The **score equation** is given by $s(\theta; X) = 0$, which is to be interpreted as an equation in θ. In many cases maximum likelihood estimators are obtained from solving the score equation.

Example 2.28. Suppose data are independent realisations of pairs (X_i, Y_i) and it is assumed that

$$Y_i \mid X_i = x_i \sim N\left(\theta' x_i, \sigma^2\right).$$

Assume the marginal density of X is given by $f_X(\cdot;\eta)$, where η is an unknown parameter. The likelihood is given by

$$\begin{aligned} L(\theta, \sigma^2, \eta; \mathcal{D}) &= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(Y_i - \theta' X_i)^2\right) f_{X_i}(X_i \mid \eta) \right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \theta' X_i)^2\right) \left(\prod_{i=1}^{n} f_{X_i}(X_i \mid \eta)\right), \end{aligned}$$

with $\mathcal{D} = \{(X_i, Y_i),\ 1 \le i \le n\}$. This implies that the maximum likelihood estimator for θ is found by minimising

$$\theta \mapsto \sum_{i=1}^{n}(Y_i - \theta' X_i)^2.$$

For this reason these estimators are also called **least squares estimators**.

> **Exercise 2.12** Verify that the maximum likelihood estimator for θ is unbiased.

If the likelihood is not smooth, the MLE may not be obtained as a zero of the score function.

> **Exercise 2.13** Verify that if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Unif\,(0, \theta)$, then the MLE for θ equals
>
> $$X_{(n)} = \max(X_1, \ldots, X_n).$$
>
> Show that this estimator is not unbiased for θ.

> **Exercise 2.14** Give the likelihood for Example 1.26.

The principle of maximum likelihood estimation does not necessarily result in a unique estimator.

Example 2.29. Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} Unif\,(\theta - 1/2, \theta + 1/2)$. The likelihood function is

$$\begin{aligned} L(\theta; X_1, \dots, X_n) &= \mathbf{1}\{\theta - 1/2 \le X_{(1)} \le X_{(n)} \le \theta + 1/2\} \\ &= \mathbf{1}\{X_{(n)} - 1/2 \le \theta \le X_{(1)} + 1/2\}. \end{aligned}$$

So any $\hat{\Theta}_n \in [X_{(n)} - 1/2, X_{(1)} + 1/2]$ is an MLE. In particular, for $\alpha_n \in [0, 1]$

$$\hat{\Theta}_n := \alpha_n(X_{(n)} - 1/2) + (1 - \alpha_n)(X_{(1)} + 1/2)$$

is an MLE. If $g : \mathbb{R} \to [0, 1]$, then we may take $\alpha_n = g(\bar{X}_n)$. In this case $\hat{\Theta}_n$ is not unique. This shows that an MLE need not be a function of the minimal sufficient statistic, which is $(X_{(1)}, X_{(n)})$ in this example.

If $u : \Omega \to \bar{\Omega}$ is a bijective function, then instead of parametrising the model using the parameter $\theta \in \Omega$, we can also parametrise it using $\bar{\theta} = u(\theta) \in \bar{\Omega}$. From the definition of a maximum likelihood estimator it then follows that if $\hat{\Theta}$ is an MLE for θ, then $u(\hat{\Theta})$ is an MLE for $u(\theta)$. This means that the MLE is **invariant under reparametrisation**. For an arbitrary map u, not necessarily bijective, we *define* the (a) MLE for $u(\theta)$ by $u(\hat{\Theta})$. Now $u(\hat{\Theta})$ maximises

$$\psi \mapsto \sup_{\theta \in \Omega : u(\theta) = \psi} L(\theta; X).$$

which is known as the **profile likelihood**.

Example 2.30. Assume $X_1, \dots, X_n$ are independent and identically distributed random variables with the lognormal distribution. This means that we assume that $Y_i := \log X_i \sim N\left(\mu, \sigma^2\right)$. Suppose that we wish to find maximum likelihood estimators for the mean and variance of the X_i. Some tedious computations (which you not need carry out here) reveal that

$$\begin{aligned} \mathrm{E}\, X_1 &= \exp(\mu + \sigma^2/2) =: \xi(\mu, \sigma^2) \\ \mathrm{Var}\, X_1 &= \xi^2(\exp(\sigma^2)) - 1) =: D(\mu, \sigma^2). \end{aligned}$$

To find the MLE for $\mathrm{E}\, X_1$ and $\mathrm{Var}\, X_1$ we therefore can simply first find the MLE for (μ, σ^2) and then plug these into the expressions of the preceding display. But since $Y_i \sim N\left(\mu, \sigma^2\right)$ we have $\hat{\mu}_{MLE} = \bar{Y}_n$ and $\hat{\sigma}^2_{MLE} = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$. Hence we immediately obtain that the MLE's for $\mathrm{E}\, X_1$ and $\mathrm{Var}\, X_1$ equal $\xi(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE})$ and $D(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE})$ respectively.

Exercise 2.15 [YS exercise 8.1.] Let $X_1, \dots, X_n$ be a random sample of size $n \ge 3$ from the exponential distribution with mean $1/\theta$.

1. Find a sufficient statistic $T(X)$ for θ and write down its density.
2. Obtain the maximum likelihood estimator $\hat{\Theta}_n$ and show that it is biased for θ, but a multiple of it is not.
3. Calculate the Cramér-Rao Lower bound for the variance of an unbiased estimator, and explain why you would not expect the bound to be attained in this example. Confirm this by calculating the variance of your unbiased estimator and comment on its behaviour as $n \to \infty$.

2.3.3 Asymptotics

Consistency of maximum likelihood estimators

Assume $X_1, \dots, X_n$ are independent and identically; distributed with density $f(\cdot;\theta)$. The likelihood is given by $L_n(\theta) = \prod_{i=1}^n f(X_i;\theta)$. Suppose the model is **identifiable**:

$$f(\cdot;\theta) \neq f(\cdot;\theta_0) \qquad \text{for every} \qquad \theta \neq \theta_0.$$

Lemma 2.31. Assume $KL_{X_1}(\theta_0,\theta) < \infty$. Then for each θ_0, and each $\theta \neq \theta_0$

$$\lim_{n\to\infty} \mathrm{P}_{\theta_0}(L_n(\theta_0) > L_n(\theta)) = 1$$

Proof. We have that $L_n(\theta_0) > L_n(\theta)$ if and only if

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log\frac{f(X_i;\theta_0)}{f(X_i;\theta)} > 0.$$

By the weak law of large numbers, $M_n(\theta)$ converges in probability (under P_{θ_0}) to

$$\mathrm{E}_{\theta_0} \log\frac{f(X;\theta_0)}{f(X;\theta)} = KL(\theta_0,\theta).$$

By identifiability, there exists an $\varepsilon > 0$ such that $KL(\theta_0,\theta) > \varepsilon$. Now

$$M_n = M_n - KL(\theta_0,\theta) + KL(\theta_0,\theta) \geq -|M_n - KL(\theta_0,\theta)| + KL(\theta_0,\theta).$$

So if $|M_n - KL(\theta_0,\theta)| < \varepsilon/2$, then $S_n > -\varepsilon/2 + \varepsilon = \varepsilon/2 > 0$. Hence

$$\mathrm{P}_{\theta_0}\left(|M_n - KL(\theta_0,\theta)| < \varepsilon/2\right) \leq \mathrm{P}_{\theta_0}\left(M_n > 0\right).$$

If we take the limit for $n \to \infty$ on both sides we get

$$\lim_{n\to\infty} \mathrm{P}_{\theta_0}\left(M_n > 0\right) \geq \lim_{n\to\infty} \mathrm{P}_{\theta_0}\left(|M_n - KL(\theta_0,\theta)| < \varepsilon/2\right) = 1.$$

□

This lemma suggests that the MLE should be consistent, as the likelihood is maximal at the value θ_0 asymptotically. However, some further conditions are needed for obtaining consistency. Precise conditions require solid knowledge of various stochastic convergence concepts and for this reason we do not go into details. Chapter 7.3 of Schervish [1995] and Chapter 5 of van der Vaart [1998] are good starting points for further reading.

Asymptotic normality

In certain "nice" settings, the MLE turns out to be asymptotically Normal. Finding sufficient conditions for such a result is really part of the subject that is known as "asymptotic statistics". A clear treatment of the topic is van der Vaart [1998], from which we adapt some results of Chapter 5 (section 5).

Definition 2.32. A statistical model $\{P_\theta,\ \theta \in \Omega\}$ is called **differentiable in quadratic mean** if there exists a measurable vector-valued function $x \mapsto \dot{\ell}(x;\theta_0)$ such that, as $\theta \to \theta_0$

$$\int \left[\sqrt{f(x;\theta)} - \sqrt{f(x;\theta_0)} - \frac{1}{2}(\theta-\theta_0)'\dot{\ell}(x;\theta_0)\sqrt{f(x;\theta_0)}\right]^2 \nu(\mathrm{d}x) = o\left(\|\theta-\theta_0\|^2\right).$$

Note that if for every x, the map $\theta \mapsto \sqrt{f(x;\theta)}$ is differentiable, then

$$\frac{\partial}{\partial\theta}\sqrt{f(x;\theta)} = \frac{1}{2\sqrt{f(x;\theta)}}\frac{\partial}{\partial\theta}f(x;\theta) = \frac{1}{2}\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)\sqrt{f(x;\theta)}$$

and then a Taylor-expansion of $\theta \mapsto \sqrt{f(x;\theta)}$ at θ_0 suggests that we can take $\dot{\ell}$ to be the score function, i.e. $\frac{\partial}{\partial\theta}\log f(x;\theta)$. However, differentiability in quadratic mean does not require existence of $\frac{\partial}{\partial\theta}f(x;\theta)$ for *every* x. The following theorem is adapted from Theorem 5.39 in van der Vaart [1998].

Theorem 2.33. Suppose that the model $\{\mathrm{P}_\theta,\ \theta \in \Omega\}$ is differentiable in quadratic mean at an inner point $\theta_0 \in \Omega \subset \mathbb{R}^k$. Furthermore, suppose that there exists a measurable function $x \mapsto \dot{\ell}(x)$ with $\int \left(\dot{\ell}(x)\right)^2 f(x;\theta_0)\nu(\mathrm{d}x) < \infty$ such that, for every θ_1 and θ_2 in a neighbourhood of θ_0

$$|\log f(x;\theta_1) - \log f(x;\theta_2)| \leq \dot{\ell}(x)\|\theta_1 - \theta_2\|.$$

If the Fisher information matrix $I(\theta_0)$ is non-singular and the MLE $\hat{\Theta}_n$ is consistent, then, under P_{θ_0},

$$\sqrt{n}(\hat{\Theta}_n - \theta_0) \rightsquigarrow N\left(0, I(\theta_0)^{-1}\right) \quad \text{as} \quad n \to \infty.$$

The take-home message from this theorem is that *in certain cases* maximum likelihood estimators are asymptotically unbiased and attain the Cramér-Rao variance asymptotically. The latter property is called **asymptotic efficiency**. While such efficiency is often put forward as a selling point of maximum likelihood, it may give the wrong impression that the MLE is unique in this sense. As Basu [1975] (page 35) wrote

> We all know that under certain circumstances the ML method works rather satisfactorily in an asymptotic sense. But the community of practising statisticians are not always informed of the fact that under the same circumstances the Bayesian method: "Begin with a reasonable prior measure q of your belief in the various possible values of θ, match it with the likelihood function generated by the data, and then estimate θ by the mode of the posterior distribution so obtained", will work as well as the ML method, because the two methods are asymptotically equivalent.

The Bayesian approach will be dealt with in Chapter 4.

Exercise 2.16 Suppose $X_1, \ldots, X_n$ are independent with common density $f(x;\theta) = \theta e^{-\theta x}\mathbf{1}_{[0,\infty)}(x)$. Use Theorem 2.33 to show the the MLE is asymptotically normal.

Exercise 2.17 A certain amount of smoothness of the map $\theta \mapsto f(\cdot;\theta)$ is essential for obtaining asymptotic normality. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Unif\,(0,\theta)$. Prove that the MLE is given by $X_{(n)} = \max(X_1, \ldots, X_n)$. Show that $-n(X_{(n)} - \theta) \rightsquigarrow Exp\,(1/\theta)$.

Exercise 2.18 Suppose $X_1, \ldots, X_n \overset{\text{ind}}{\sim} N(\theta, 1)$, with $\theta \geq 0$.

1. Show that the MLE is given by $\hat{\Theta}_n = \bar{X}_n \mathbf{1}\{\bar{X}_n \geq 0\}$.
2. (*, advanced) Show that if $\theta > 0$, $\sqrt{n}(\hat{\Theta}_n - \theta) \rightsquigarrow N(0, 1)$.
   Hint: The idea is that if $\theta > 0$, then $\bar{X}_n > 0$ and then $\hat{\Theta}_n = \bar{X}_n$. The asymptotic distribution then follows from the central limit theorem. It requires a bit of work to make this precise. Define the event $A_n = \{\bar{X}_n \geq 0\}$. By the law of large numbers, if $\theta > 0$, then $\mathrm{P}_\theta(A_n) \to 1$ ($n \to \infty$). Let $\mathbf{1}\{A\}$ denote the indicator function of a set A (Hence $\mathbf{1}\{A\} = 1$ if A holds true, else it is zero). Note that $\mathrm{P}_\theta(A) = \mathrm{E}_\theta \mathbf{1}\{A\}$. Take any $u \in \mathbb{R}$

$$\begin{aligned}\mathbb{P}_\theta(\sqrt{n}(\hat{\Theta}_n - \theta) \leq u) &= \mathbb{E}_\theta \mathbf{1}\{\sqrt{n}(\hat{\Theta}_n - \theta) \leq u\} \\ &= \mathbb{E}_\theta \left[\mathbf{1}\{\sqrt{n}(\hat{\Theta}_n - \theta) \leq u\}\mathbf{1}\{A_n\}\right] + \mathbb{E}_\theta \left[\mathbf{1}\{\sqrt{n}(\hat{\Theta}_n - \theta) \leq u\}\mathbf{1}\{A_n^C\}\right]\end{aligned}$$

   Now try to finish the argument.
3. (*, advanced) Show that in case $\theta = 0$

$$\mathrm{P}_0(\sqrt{n}\hat{\Theta}_n \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/2 & \text{if } x = 0 \\ \Phi(x) & \text{if } x > 0 \end{cases}.$$

   Hence, in this case the limit distribution is not normal, but a mixture of a point mass at zero and the standard normal distribution.

Exercise 2.19 [YS exercise 8.10.] A random sample $X_1, \ldots, X_n$ is taken from the normal distribution $N(\mu, 1)$.

1. Show that the maximum likelihood estimator $\hat{\mu}_n$ of μ is the minimum variance unbiased estimator (show that the Cràmer-Rao lower bound is attained).
2. Find its distribution and variance.
3. (*, advanced) Now define a second estimator T_n by

$$T_n = \begin{cases} \hat{\mu}_n & \text{when } |\hat{\mu}_n| \geq n^{-1/4} \\ \frac{1}{2}\hat{\mu}_n & \text{when } |\hat{\mu}_n| < n^{-1/4} \end{cases}$$

   (a) Show that $\mathrm{P}_\mu(T_n \neq \hat{\mu}_n)$ tends to one when $\mu = 0$ but to zero if $\mu > 0$.

   (b) Derive that for $\mu = 0$, the asymptotic distribution of T_n is $N(0, 1/(4n))$.

   (c) Derive that for $\mu > 0$, the asymptotic distribution of T_n is $N(0, 1/n)$.

   Hence, asymptotically, T_n improves upon $\hat{\mu}_n$.

4. (*, advanced) The Mean Squared Error (MSE) for $\hat{\mu}_n$ is given by

$$\mathrm{E}_\mu[(\hat{\mu}_n - \mu)^2] = 1/n = \mathrm{E}_\mu\left[(\hat{\mu}_n - \mu)^2 1_{\{|\hat{\mu}_n| \geq n^{-1/4}\}}\right] + \mathrm{E}_\mu\left[(\hat{\mu}_n - \mu)^2 1_{\{|\hat{\mu}_n| < n^{-1/4}\}}\right],$$

   while the MSE of T_n is

$$\mathrm{E}_\mu[(T_n - \mu)^2] = \mathrm{E}_\mu\left[(\hat{\mu}_n - \mu)^2 1_{\{|\hat{\mu}_n| \geq n^{-1/4}\}}\right] + \mathrm{E}_\mu\left[(0.5\hat{\mu}_n - \mu)^2 1_{\{|\hat{\mu}_n| < n^{-1/4}\}}\right].$$

   Compare the MSE of $\hat{\mu}_n$ and T_n both in case $\mu = 0$ and $\mu > 0$. Is T_n a sensible estimator in practise?

This exercise gives an example of **asymptotic superefficiency**. Asymptotically T_n outperforms the minimum variance unbiased estimator at the point of superefficiency $\mu = 0$.

2.3.4 Maximum likelihood estimation in Exponential family models*

In exponential family models, the MLE is consistent and efficient. The following result is taken from Schervish [1995], Theorem 7.57.

Theorem 2.34. Suppose that $\{X_n\}_{n=1}^{\infty}$ are independent and identically distributed with density

$$f_{X_i}(x; \theta) = c(\theta) \exp\left(\theta' x\right)$$

with respect to the dominating measure ν. Suppose that the natural parameter space Ω is an open subset of $\mathbb{R}^k$. Let $\hat{\Theta}_n$ be the MLE of θ based on $X_1, \ldots, X_n$ (if it exists). Then $\lim_{n\to\infty} \mathrm{P}_\theta(\hat{\Theta}_n \text{ exists}) = 1$ and under P_θ

$$\sqrt{n}(\hat{\Theta}_n - \theta) \rightsquigarrow N\left(0, I(\theta)^{-1}\right).$$

This theorem implies that the MLE is consistent and that if $g : \Omega \to \mathbb{R}$ has continuous partial derivatives, then $g(\hat{\Theta}_n)$ is an asymptotically efficient estimator of $g(\theta)$ ([Schervish [1995], Corollary 7.58 and Corollary 7.59).

2.3.5 Inconsistency of maximum likelihood estimators

Maximum likelihood estimation is among the most popular estimation methods in statistics. In *certain circumstances* (for example exponential family models, see Theorem 2.34), maximum likelihood estimators are asymptotically efficient, by which we mean that the Cramér-Rao bound is attained asymptotically. However, in general, maximum likelihood estimators need not be optimal. Moreover, even if the MLE (or "a" if it is not unique) is optimal, many other estimators (such as the posterior mean under various priors, see later chapters) may be efficient as well. The use of ML techniques is not always accompanied by a clear appreciation of their limitations. From LeCam [1953]:

> ⋯ although all efforts at a proof of the general existence of [asymptotically] efficient estimates ⋯ as well as a proof of the efficiency of ML estimates were obviously inaccurate and although accurate proofs of similar statements always referred not to the general case but to particular classes of estimates ⋯ a general belief became established that the above statements are true in the most general sense.

Another warning is from Zacks [2014]:

> In contrast to statements in many textbooks, maximum likelihood estimation is not a universally good procedure and should not be used in a dogmatic fashion. As we show in examples, there are various cases in which maximum likelihood estimators are inefficient compared with unbiased or other types of estimators. ⋯ The advantage is that in many cases it is very simple to derive the maximum likelihood estimators.

Indeed, there are examples where the MLE is inconsistent. The following example is classical.

Example 2.35 (Neyman-Scott probleem)**.** Let $X_{ij} \overset{\text{ind}}{\sim} N(\theta_i, \sigma^2)$, $i = 1, \dots, n$ and $j = 1, \dots, r$. The unique MLE is given by

$$\hat{\Theta}_i = \frac{1}{r}\sum_{j=1}^{r} X_{ij} \qquad \hat{\Sigma}^2 = \frac{1}{rn}\sum_{i=1}^{n}\sum_{j=1}^{r}(X_{ij} - \hat{\Theta}_i)^2.$$

These can be derived by taking partial derivatives of the loglikelihood function and equating these to zero.

To see that $\hat{\Sigma}^2$ is not consistent, note that the statistics $S_i^2 = \sum_{j=1}^{r}(X_{ij} - \hat{\Theta}_i)^2$ are independent and identically distributed with expectation $\mathrm{E}\, S_i^2 = (r-1)\sigma^2$. Therefore $n^{-1}\sum_{i=1}^{n} S_i^2 \xrightarrow{\text{p}} (r-1)\sigma^2$ and hence

$$\hat{\Sigma}^2 \xrightarrow{\text{p}} \frac{r-1}{r}\sigma^2.$$

In this example, the number of "nuisance parameters" (parameters which are in the model but not of direct interest), $\{\theta_i\}_{i=1}^n$ grows with n.

In case the MLE is inconsistent, it is common to fix this problem by either adjusting the likelihood function (in ways that can appear to be a bit at-hoc) or the estimator. As an example, in the previous Neyman-Scott problem, it is obvious that the estimator $\frac{r}{r-1}\hat{\Sigma}^2$ is consistent for estimating σ^2.

In nonparametric estimation problems inconsistency of the MLE is not uncommon. See for instance Section 2.2 of Groeneboom and Jongbloed [2014] for an example on nonparametric estimation of a decreasing density.

2.4 Hypothesis testing

In hypothesis testing, we wish to decide whether $\theta \in \Omega_0 \subseteq \Omega$ or $\theta \in \Omega_1 \subseteq \Omega$, where $\Omega_0 \cap \Omega_1 = \emptyset$. Usually Ω_0 and Ω_1 are chosen such that $\Omega_0 \cup \Omega_1 = \Omega$. The **null-hypothesis** is given by

$$H_0 : \theta \in \Omega_0,$$

while the **alternative hypothesis** is given by $H_1 : \theta \in \Omega_1$. We recap frequentist hypothesis testing as proposed by Neyman and Pearson.

Definition 2.36. A **nonrandomised test** (T, C) consists of a **test statistic** $T(X)$ together with a **critical region** $C \subseteq \mathbb{R}$. We reject H_0 and accept H_1 if $T(X) \in C$, else we do not reject H_0.

Definition 2.37. The **power function** $\beta : \Omega \mapsto [0, 1]$ of the statistical test (T, C) is defined by

$$\beta(\theta) = \mathrm{P}_\theta(T(X) \in C).$$

Ideally, we would like to have $\beta(\theta) = 0$ if $\theta \in \Omega_0$ and $\beta(\theta) = 1$ if $\theta \in \Omega_1$, but due to randomness this is of course impossible.

An important questions concerns the choice of T and C for a particular statistical model. For the moment suppose we have chosen some statistic T. The critical region should be such that if $T \in C$ we have an indication that H_1 may be true. In simple settings, this is sort of clear. In general, there is no rule for this, and there may be situations where it is unclear how to choose C (examples are given in Clayton [2021]). For now, let's neglect these issues. The classical Neyman-Pearson theory then treats H_0 and H_1 asymmetrically and prescribes to choose the region C as follows:

1. First choose a **significance level** $\alpha \in [0, 1]$ and take C such that the inequality
$$\sup_{\theta \in \Omega_0} \beta(\theta) \leq \alpha \tag{2.2}$$
is satisfied. This ensures the probability of a type I error (incorrectly rejecting H_0) to be bounded by α.
2. Conditional on (2.2), maximise $\beta(\theta)$ for all $\theta \in \Omega_1$ by choosing the volume of C as large as possible. This ensures minimising the type II error (incorrectly not rejecting H_0).

If C is given, the number $\sup_{\theta \in \Omega_0} \beta(\theta)$ is called the significance level of the test.

For choosing a test statistic we need to define an optimality criterion.

Definition 2.38. Suppose (T, C) is a test with significance level α. It is called **Uniformly Most Powerful (UMP)** if

$$\mathrm{P}_\theta(T \in C) \geq \mathrm{P}_\theta(\tilde{T} \in \tilde{C}) \quad \text{for all} \quad \theta \in \Omega_1$$

for all tests $(\tilde{T}, \tilde{C})$ with significance level α (for the same hypothesis testing problem).

Does such a test exist? To be able to answer this question we have to include **randomised tests**. These are tests where the decision for rejecting the null hypothesis not only depends on the observation X, but on an additional independent random variable U as well. Few statisticians will recommend such a test for practical purposes. However, including randomised tests enables us to establish UMP-tests in certain settings. Before we give an example of a randomised test, we note that for a nonrandomised test the power function can be written as

$$\beta(\theta) = \mathrm{E}_\theta\, \phi(X) \quad \text{where} \quad \phi(X) = \mathbf{1}_{\{T(X) \in C\}}.$$

This function ϕ is called the **critical function** of the test: it gives the probability to reject the null hypothesis when X is observed. Clearly, $\phi(X) \in \{0, 1\}$ in case of a nonrandomised test.

Example 2.39. Suppose $X \sim Bin(10, \theta)$ and we wish to test $H_0 : \theta \leq 1/2$ versus $H_1 : \theta > 1/2$. As we only have one observation it is natural to take $T(X) = X$. We will reject H_0 for large values of X and hence the critical region is of the form $C = \{c, c+1, \ldots, 10\}$. To obtain maximal power of the test under the alternative hypothesis, we wish to attain significance level $\alpha = 0.05$ *exactly* (equality in equation (2.2)). Since

$$\sup_{\theta \leq 1/2} \mathrm{P}_\theta(T \in C) = \mathrm{P}_{1/2}(X \geq c) = \begin{cases} 0.055 & \text{if} \quad c = 8 \\ 0.011 & \text{if} \quad c = 9 \end{cases}, \tag{2.3}$$

we see that this is impossible. As a remedy, one can choose $c = 8$ with probability γ and $c = 9$ with probability $1 - \gamma$, where we choose γ such that

$$0.055\gamma + 0.011(1 - \gamma) = \alpha.$$

Taking $\alpha = 0.05$ gives $\gamma \approx 0.89$ and leads to the notion of a *randomised test*. the probability γ is chosen such that *on average* (if we would repeat the testing procedure infinitely often) significance level α is obtained. To make this mathematically precise, define

$$\phi(X, U) = \mathbf{1}_{\{X \geq 8, U \leq \gamma\}} + \mathbf{1}_{\{X \geq 9, U > \gamma\}},$$

where $U \sim Unif(0, 1)$ is independent of X. The interpretation of this function is that we reject when $\phi(X, U) = 1$. We have

$$\mathrm{E}_\theta \, \phi(X, U) = \mathrm{E}_\theta \, \mathrm{E}[\phi(X, U) \mid X],$$

the inner expectation on the right-hand-side being over U.

The critical function of the randomised test is defined by

$$x \mapsto \phi(x) := \mathrm{E}_\theta[\phi(X, U) \mid X = x].$$

This function gives the probability of rejecting H_0 when $X = x$ is observed. In this example, we have

$$\begin{aligned} \phi(x) &= \int \phi(x, u) f_U(u) \, \mathrm{d}u \\ &= \int_0^\gamma \mathbf{1}_{\{x \geq 8\}} \, \mathrm{d}u + \int_\gamma^1 \mathbf{1}_{\{x \geq 9\}} \, \mathrm{d}u \\ &= \gamma \mathbf{1}_{\{x \geq 8\}} + (1 - \gamma)\mathbf{1}_{\{x \geq 9\}} = \gamma \mathbf{1}_{\{8\}}(x) + \mathbf{1}_{\{9,10\}}(x). \end{aligned}$$

So we always reject when $X \in \{9, 10\}$ and if $X = 8$ we reject with probability γ. As a check on our calculations, we establish that

$$\mathrm{E}_{1/2} \, \phi(X) = \gamma \mathrm{P}_{1/2}(X = 8) + \mathrm{P}_{1/2}(X \in \{9, 10\}) = \alpha = 0.05.$$

Definition 2.40. The **critical function** is defined by

$$x \mapsto \phi(x) = \mathrm{P}_\theta(\text{reject } H_0 \mid X = x).$$

Just as in the example, using the law of total probability we have

$$\beta(\theta) = \mathrm{P}_\theta(\text{reject } H_0) = \mathrm{E}_\theta \, \phi(X).$$

Suppose $\Omega_0 = \{\theta_0\}$ and the test statistic is such that we reject for large values. We claim that we can always find values k and $\gamma \in [0, 1]$ such that the test

$$\phi(x) = \gamma \mathbf{1}_{[k,\infty)}(T(x)) + (1 - \gamma)\mathbf{1}_{(k,\infty)}(T(x))$$

satisfies $\mathrm{E}_{\theta_0}\,\phi(X) = \alpha$. The construction goes as follows: first find the smallest value k such that $\mathrm{P}_{\theta_0}(T \geq k)$ is above α (if this probability happens to be exactly α no randomisation is required and this value of k gives a nonrandomised test of level α with $\gamma = 0$). Using equation (2.3) of the example we have $k = 8$ in that case. If we define

$$\gamma = \frac{\alpha - \mathrm{P}_{\theta_0}(T > \bar{k})}{\mathrm{P}_{\theta_0}(T \geq \bar{k}) - \mathrm{P}_{\theta_0}(T > \bar{k})} \tag{2.4}$$

then we have $\mathrm{E}_{\theta_0}\,\phi(X) = \alpha$, as claimed.

Exercise 2.20 Verify this claim.

We now turn to the question of choosing the test statistic. In case $\Omega_0 = \{\theta_0\}$ and $\Omega_1 = \{\theta_1\}$ there is a clear answer to this question. If the distribution under the null or alternative hypothesis is completely specified we call the hypothesis simple. So here we consider the case of testing a simple versus a simple hypothesis. This case is easy to analyse as the power function for a test ϕ has only two values: $\mathrm{E}_0\,\phi$ and $\mathrm{E}_1\,\phi$ (E_i denotes expectation under P_{θ_i}). We look for a test that

$$\begin{aligned} &\text{maximises} && \mathrm{E}_1\,\phi(X) \\ &\text{subject to} && \mathrm{E}_0\,\phi(X) \leq \alpha \end{aligned} \tag{2.5}$$

Denote the density of P_{θ_i} by f_i. [2] The following lemma is known as the **Neyman-Pearson lemma**.

Lemma 2.41. Define the test

$$\phi^*(x) = \begin{cases} 1 & \text{if } f_1(x) > k f_0(x) \\ \gamma(x) & \text{if } f_1(x) = k f_0(x) \\ 0 & \text{if } f_1(x) < k f_0(x) \end{cases},$$

where $k \geq 0$ is a constant and $\gamma : \mathcal{X} \to [0,1]$ an arbitrary function. Then

1. *Optimality*: Let $\alpha^* = \mathrm{E}_0\,\phi^*$. For any k and $\gamma(x)$, ϕ^* is the test that maximises $\mathrm{E}_1\,\phi$ over all tests ϕ with $\mathrm{E}_0\,\phi \leq \alpha^*$. That is ϕ^* is the most powerful size α^*-test.
2. *Existence*: Given $\alpha \in (0,1)$, there exist constants k and γ such that the test ϕ^* with $\gamma(x) = \gamma$ has size exactly equal to α.
3. *Uniqueness*: If the test ϕ has size α and is of maximum power amongst all possible tests with level at most α, then $\phi \mathbf{1}_B = \phi^* \mathbf{1}_B$, where $B = \{x : f_1(x) \neq k f_0(x)\}$.

Proof. Note that the existence part was already proved in the previous discussion (take $T(x) = f_1(x)/f_0(x)$). For the optimality result, let ϕ be any test with $\mathrm{E}_0\,\phi \leq \alpha^*$. Define

$$U(x) = (\phi^*(x) - \phi(x)(f_1(x) - k f_0(x))$$

and verify that $U(x) \geq 0$. Hence

$$\begin{aligned} 0 \leq \int U(x)dx &= \int \phi^*(x) f_1(x) dx - \int \phi(x) f_1(x)\,\mathrm{d}x + \\ &\qquad k\left(\int \phi(x) f_0(x)\,\mathrm{d}x - \int \phi^*(x) f_0(x)\,\mathrm{d}x\right) \\ &= \mathrm{E}_1\,\phi^*(X) - \mathrm{E}_1\,\phi(X) + k\left(\mathrm{E}_0\,\phi(X) - \mathrm{E}_0\,\phi^*(X)\right) \\ &\leq \mathrm{E}_1\,\phi^*(X) - \mathrm{E}_1\,\phi(X) \end{aligned}$$

[2] Existence of densities is guaranteed, the measure $P_{\theta_0} + P_{\theta_1}$ is a dominating measure for example. Note that usually we would denote this density by $f_X(\cdot;\theta_i)$, but simply writing f_i makes the statements in the following easier to read.

which proves the optimality result.

For the uniqueness result, note that if ϕ^* is not unique, then there is a test ϕ with the same size and power as ϕ^*. Then it follows from the previous display that $\int U(x)dx = 0$. As we also have that $U(x) \geq 0$ we conclude $U(x) = 0$ for all x. This implies that if x is such that $f_1(x) \neq k f_0(x)$, then $\phi^*(x) = \phi(x)$. □

Remark 2.42. The NP-setup has become a default choice for many scientists, but has been criticised a lot (the same holds for p-values). Instead of bounding the type I error by α one could for example also derive a test by choosing the critical region C to minimise a linear combination of the type I and type II errors. Take $k > 0$ and suppose $\Omega = \{\theta_0, \theta_1\}$ (the simplest case). Instead of the NP-setup we could choose the critical region to minimise

$$\begin{aligned} k\mathrm{P}_{\theta_0}(X \in C) + \mathrm{P}_{\theta_1}(X \notin C) &= k \int_C f_X(x;\theta_0)\,\mathrm{d}\nu(x) + 1 - \int_C f_X(x;\theta_1)\,\mathrm{d}\nu(x) \\ &= 1 + \int_C \left(k f_X(x;\theta_0) - f_X(x;\theta_1)\right)\,\mathrm{d}\nu(x). \end{aligned}$$

Hence we should take C such that the integrand is negative. That is

$$C = \left\{ x : \frac{f_X(x;\theta_1)}{f_X(x;\theta_0)} \geq k \right\}.$$

The form of the critical region is just as with the standard NP-test, but the criterion to decide to reject is derived differently.

Example 2.43. Suppose $X \sim Exp(\theta)$ and consider $H_0 : \theta = \theta_0 = 1$ (say) against $H_1 : \theta = \theta_1$ where $\theta_1 > 1$. We reject H_0 when

$$\frac{f_X(x;\theta_1)}{f_X(x;\theta_0)} = \theta_1 e^{-(\theta_1 - 1)x} \geq k.$$

This is the case when $x < (\theta_1 - 1)^{-1} \log\left(\theta_1/k\right) =: k'$. Solving $\mathrm{P}_{\theta_0}(X < k') = \alpha$ gives $k' = -\log(1-\alpha)$. Hence, according to the Neyman-Pearson lemma, the optimal test rejects when $\phi(X) = 1$ where

$$\phi(X) = \mathbf{1}_{(-\infty, -\log(1-\alpha))}(X) = \mathbf{1}_{(0, -\log(1-\alpha))}(X).$$

If $\tilde{\phi}$ is any other test with $\mathrm{E}_{\theta_0}\,\tilde{\phi} \leq \alpha$, then $\mathrm{E}_{\theta_1}\,\phi \geq \mathrm{E}_{\theta_1}\,\tilde{\phi}$.

Existence of uniformly most powerful tests is a great deal to expect. It is asking the Neyman-Pearson test for simple vs. simple hypothesis to be the same for every pair of simple hypotheses contained within H_0 and H_1. In example 2.43 it is easy to see that the derived test is UMP for the hypothesis $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$ since the critical function of the test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ does not depend on θ_1. This argument holds more generally in case of monotone likelihood ratios and a one-sided alternative hypothesis.

Definition 2.44. The family of densities $\{f(\cdot;\theta),\ \theta \in \Omega \subseteq \mathbb{R}\}$ is said to be of increasing **monotone likelihood ratio (MLR)** if there exists a function $t(x)$ such that the likelihood ratio $f(x;\theta_2)/f(x;\theta_1)$ is a non-decreasing function of $t(x)$ whenever $\theta_2 \geq \theta_1$.

Example 2.45. Suppose $X_1, \ldots, X_n$ are independent, each with density

$$f_{X_1}(x;\theta) = c(\theta)h(x)e^{\theta\tau(x)}.$$

This is a one-parameter exponential family parametrised in terms of its natural parameter θ. Set $X = (X_1, \ldots, X_n)$. Then,

$$f_X(x) = c(\theta)^n \prod_{i=1}^{n} h(x_i) e^{\theta \sum_{i=1}^{n} \tau(x_i)}.$$

Define $t(x) = \sum_{i=1}^{n} \tau(x_i)$. For any $\pi_2 \geq \pi_1$ we have

$$\frac{f(x;\theta_2)}{f(x;\theta_1)} = \left(\frac{c(\theta_2)}{c(\theta_1)}\right)^n \exp\left((\theta_2 - \theta_1)t(x)\right)$$

which is non-decreasing in $t(x)$. Hence the family is of increasing MLR.

Theorem 2.46 (Young and Smith [2005], theorem 4.2)**.** Suppose X has a distribution from a family which is of increasing MLR with respect to the statistic $t(X)$ and that we wish to test $H : \theta \leq \theta_0$ against $H_A : \theta > \theta_0$. Suppose the distribution of X is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^k$.

1. The test

$$\phi^*(x) = \begin{cases} 1 & \text{if } t(x) > t_0 \\ 0 & \text{if } t(x) \leq t_0 \end{cases}$$

   is UMP among all tests of size $\leq \mathrm{E}_{\theta_0} \phi^*(X)$.

2. Given some $\alpha \in (0, 1]$, there exists some t_0 such that the test ϕ^* has size exactly α.

Two sided hypothesis testing involves testing $H_0 : \theta \in [\theta_1, \theta_2]$ vs. $H_1 : \theta \in (-\infty, \theta_1) \cup (\theta_2, \infty)$, where $\theta_1 < \theta_2$ or $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. Quoting from Young and Smith [2005] (section 7.1):

> In this situation, we cannot in general expect to find a UMP test, even for nice families, such as the monotone likelihood ratio or exponential family models. The reason is obvious: if we construct a Neyman-Pearson test of say $\theta = \theta_0$ against $\theta = \theta_1$ for some $\theta_1 \neq \theta_0$, the test takes quite a different form when $\theta_1 > \theta_0$ from when $\theta_1 < \theta_0$. We simply cannot expect one test to be most powerful in both cases simultaneously.

Exercise 2.21 [YS exercise 4.2.] Let $X_1, \ldots, X_n$ be independent $N\left(\mu, \sigma^2\right)$ random variables where $\sigma^2 > 0$ is a known constant and $\mu \in \mathbb{R}$ is unknown. Show that $X = (X_1, \ldots, X_n)$ has monotone likelihood ratio. Given $\alpha \in (0, 1)$ and $\mu_0 \in \mathbb{R}$, construct a uniformly most powerful test of size α of $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$, expressing the critical region in terms of the standard normal distribution function Φ.

Exercise 2.22 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Exp(\theta)$. Consider $H_0 : \theta \leq 1$ against $H_1 : \theta > 1$. Derive a uniformly most powerful test of size α.

Exercise 2.23 [YS exercise 4.5.] Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Unif\,(0, \theta)$.

1. Show that there exists a uniformly most powerful size α test of $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ and find its form.
2. Let $T = \max(X_1, \ldots, X_n)$. Show that the test

$$\phi(x) = \begin{cases} 1 & \text{if } t > \theta_0 \text{ or } t \leq b \\ 0 & \text{if } b < t \leq \theta_0 \end{cases},$$

where $b = \theta_0 \alpha^{1/n}$, is a uniformly most powerful test of size α for testing H_0 against $H_1' : \theta \neq \theta_0$. (Note that in a more "regular" situation a UMP test of H_0 against H_1' does not exist.)

Exercise 2.24 Let $\sigma_1^2 > 0$ and $\sigma_2^2 > 0$. Suppose $X_1 \sim N(0, \sigma_1^2)$ and $X_2 \sim N(0, \sigma_2^2)$ are independent. Find the critical region for the best symmetric test $H_0 : X_1 \sim N(0, \sigma_1^2),\ X_2 \sim N(0, \sigma_2^2)$ against $H_1 : X_2 \sim N(0, \sigma_1^2),\ X_1 \sim N(0, \sigma_2^2)$.

A symmetric test here is a test that takes the opposite action if the two data values are switched, so $\phi(x_1, x_2) = 1 - \phi(x_2, x_1)$. For a symmetric test the error probabilities under H_0 and H_1 will be equal.

2.5 Does frequentist hypothesis testing answer the right question?

There are essentially three ingredients: a null hypothesis H_0, an alternative hypothesis H_1, and data D. Frequentist hypothesis testing looks at

$$\mathbb{P}\left(\text{data } D\text{, or data more extreme than } D \text{ pointing towards } H_1 \mid H_0\right).$$

Note that it is not just the data D, we also consider hypothetical data we have not seen (and may never see). The issues with p-values have been reported over decades, yet these dominate statistical reporting in the literature.

Bayesian statistics, to be fully discussed in Chapter 4 focusses on something entirely different

$$\mathbb{P}(H_0 \mid D) = \frac{\mathbb{P}(H_0)}{\mathbb{P}(D)}\mathbb{P}(D \mid H_0) = \frac{\mathbb{P}(D \mid H_0)\mathbb{P}(H_0)}{\mathbb{P}(D \mid H_0)\mathbb{P}(H_0) + \mathbb{P}(D \mid H_1)\mathbb{P}(H_1)}, \tag{2.6}$$

the final equality being true if H_1 is of the form not(H_0). So we answer the question what the probability of the hypothesis is, in light of the data. Doesn't that sound more natural? Equation (2.6) is just Bayes' formula, which can be rewritten to

$$\frac{\mathbb{P}(H_1 \mid D)}{\mathbb{P}(H_0 \mid D)} = \frac{\mathbb{P}(D \mid H_1)}{\mathbb{P}(D \mid H_0)}\frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)}.$$

This reads

$$\boxed{\text{posterior odds} = \text{likelihood ratio} \times \text{prior odds}}.$$

It is this prior odds term that frequentist object to. In the specific setting where $X \sim f(\cdot; \theta)$ and $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, we have

$$\frac{\mathbb{P}(D \mid H_1)}{\mathbb{P}(D \mid H_0)} = \frac{f(x; \theta_1)}{f(x; \theta_0)}.$$

The Neyman-Person Lemma states that this is the optimal frequentist test-statistic. Therefore, at least in this simple setting, both approaches agree that the likelihood-ratio is of central importance. However, as we will see, the way this is used to reach a decision (accept or reject) is fundamentally different.

2.6 P-values

Suppose (T, C_α) is a (nonrandomised) test with significance level α and we observe a realisation t from T. The rule is to reject the null hypothesis when $t \in C_\alpha$. This gives a binary decision rule: either reject or not. A common criticism is that this is not informative enough and one should also provide a measure of strength of evidence in favour (or against) the null hypothesis. A perfect candidate would be the probability that the null hypothesis is true. However, as the parameter is supposed to be fixed in classical statistics this measure is not available (in Bayesian statistics it is).

Suppose we test $H_0 \ : \ \theta \in \Omega_0$. Because of the defining relation for C_α:

$$\sup_{\theta \in \Omega_0} \mathrm{P}_\theta(T \in C_\alpha) \leq \alpha,$$

decreasing α yields a critical region of smaller volume. Now if $\alpha = 1$, then certainly $t \in C_\alpha$. By successively decreasing the value of α, at a certain value we will no longer have that $t \in C_\alpha$. The tipping point is precisely what the p-value is.

Definition 2.47. Suppose (T, C_α) is a (nonrandomised) test with significance level α and we observe a realisation t from T. The **p-value** is defined as

$$p = \inf\{\alpha \in [0,1] \ : \ t \in C_\alpha\}.$$

Hence, it is the smallest value of α for which we reject the null-hypothesis when observing the realisation t. Since $p \leq \alpha_0$ if and only if $t \in C_{\alpha_0}$ we can use the p-value to decide upon rejecting of the hypothesis for any level $\alpha_0 \in [0,1]$.

Use of p-values has been much disputed in the statistical literature. For an interesting discussion, see Schervish [1996] where it is shown that the common informal use of p-values as measures of support or evidence for hypotheses has serious logical flaws.

2.6.1 Criticism and misuse of p-values

Many practitioners mistakenly believe that statistics requires performing a statistical test and reporting its p-value. Even worse, complete diagrams are given to "help" the experimenter to find the "right" test. An example is given in Figure 2.1, which was taken from the book McElreath [2015] (who strongly advocates against this recipe-style towards statistics). Clearly, such an approach completely ignores proper descriptive statistics. Moreover, the assumed statistical model should be reported, and it should be assessed that this model is reasonable for the available dataset.

There is a general misconception among practitioners that the p-value provides a measure of support, where small values favour the alternative hypothesis. The following simple example explains what is wrong with this.

Example 2.48. Suppose $X \sim Bin(n, \theta)$ and we wish to test $H_0 \ : \ \theta = 1/2$ versus $H_1 \ : \ \theta \neq 1/2$. Suppose we observe $x = n/2 + \sqrt{n}$ and n is such that x is integer valued. The p-value for this problem is given by

$$p = 2\mathrm{P}_{1/2}(X \geq x) = 2\mathrm{P}_{1/2}\left(\frac{X - n/2}{\sqrt{n/4}} \geq 2\right) \approx 2\mathbb{P}(Z \geq 2) \approx 0.0455,$$

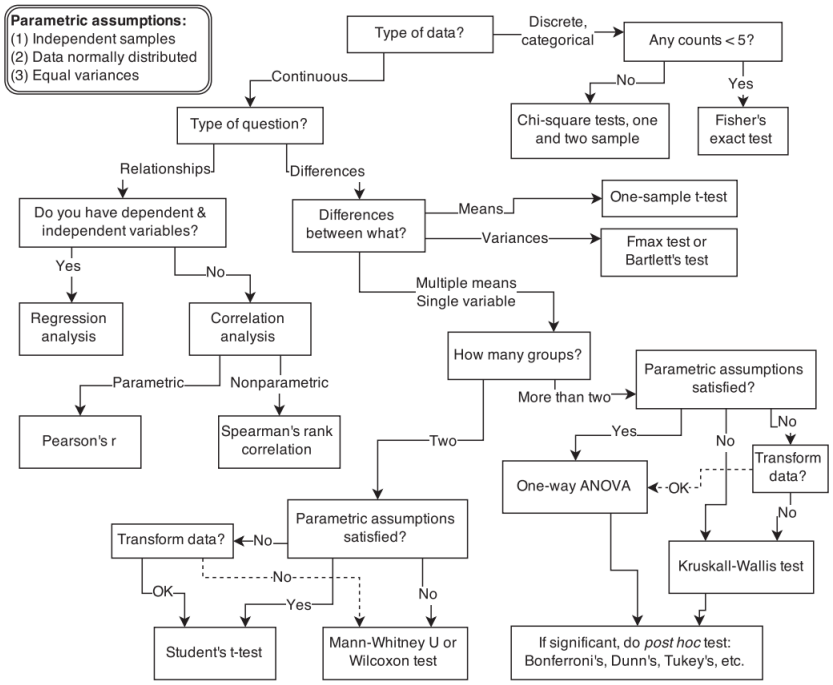

Figure 2.1: Scheme to find the "right" test.

where $Z \sim N(0,1)$. The first $\approx$-sign is obtained by applying the central limit theorem, which requires n to be sufficiently large. Hence it appears we have evidence against $\theta = 1/2$ when using significance-level $\alpha = 0.05$, the evidence being the same for all values of n for which the central-limit-theorem-approximation is valid. However, the observed proportion of successes is given by

$$\frac{x}{n} = \frac{1}{2} + \frac{1}{\sqrt{n}}$$

which becomes arbitrarily close to $1/2$ when n is taken sufficiently large.

In many practical settings of hypothesis testing, H_0 is by definition not true. Some examples:

- Belgians and Germans do equally well on an IQ-test;
- the proportion of males in the population is 0.75;
- the correlation between two quantities in the population equals 0.

Hence, with a sufficiently large sample size, H_0 is always rejected (as it should). But then we only find out what we already knew before collecting data! [3]

Furthermore, there are problems with the interpretation of rejecting a null hypothesis. This can be seen from the following line of reasoning which is often seen in journal articles (taken from an article by Cohen [1994])

[3] An argument in favour of point-null hypothesis testing $H_0 : \theta = \theta_0$ is that it should be viewed as a limiting case of $H_0 : \theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$, with $\varepsilon \downarrow 0$. As we will see later, also for Bayesian statistics point-null hypothesis testing gives certain specific difficulties.

> If H_0 is true, then this result (statistical significance) would probably not occur.
>
> This result has occurred.
>
> Then H_0 is probably not true and therefore formally invalid.

The first line corresponds to finding a critical region C for a test statistic T. The second line corresponds to discovering that the observed value of T, denoted by t, satisfies $t \in C$.

This reasoning is incorrect, as seen from the following

> If a person is an American, then he is probably not a member of congress. *True, right?*
>
> This person is a member of congress.
>
> Therefore, he is probably not an American.

Only if the word "probably" would not appear, and the first line (the major premise) would be correct, then this line of reasoning would be correct. Hence, there is a logical flaw in frequentist hypothesis testing.

2.6.2 Tricks to get a significant result

Why are many research papers that claim a significant results which cannot be replicated? Chapter 14 in Briggs [2008] discusses various ways to "cheat" in statistics. Unfortunately, I believe this cheating does regularly happen in practice. Here, we extract some remarks on "publishable p-values", as discussed in section 14.5 of Briggs [2008] (all quotes in this section are extracted from here). [4] Briggs argues as follows:

> Most journals, say in medicine or those serving fields ending with "ology", are slaves to p-values. Papers have a difficult, if not impossible, time getting published unless authors can demonstrate for their study a p-value that is publishable, that is, that is less than 0.05. Sometimes, the data are not cooperative and the p-value that you get from using a common statistic is too large to see the light of print. This is bad news, because if you are an academic, you must publish papers else you can't get grants, and if you don't get grants, then you do not bring money into your university, and if you don't bring money into your university, the Dean is unhappy, and if they Dean is unhappy you do not get tenure, and if you do not get tenure, then you are out the door and you feel shame.

So what to do? Here are some ways to "improve" your results:

1. Increase your sample size. This will always work with point-null hypothesis testing. This immediately implies that reporting a p-value without sample size is meaningless.
2. Change your test-statistic. As Briggs argues

   > This comes from the useful loophole in classical theory that there is no rule which specifies which statistic you can use in any situation. Thus, though some creativity and willingness to spend time with your statistical software, you can create small p-values where others see only despair. This isn't so easy to do in R because you have to know the names of the alternate statistics, but it's cake in software like SAS, which usually prints out dozens of statistics in standard situations, which is one reason SAS is worth its exorbitant price. Look around at the advertising brochures of statistical software and you will see that the openly boast of the large number of tests on offer.

[4] I believe Briggs [2008] has a point here, but these quotes give a wrong impression that all scientists/practitioners deliberately mess around with p-values just to get a significant result. But yes, I believe unfortunately this does take place. Any statistical method can be misused, but that does not imply the method is invalid. This also applies to Bayesian statistics, to be discussed in a later chapter. Bayesian hypothesis testing requires choosing a prior and cut-off, and these can be changed to influence the result, just like changing the test-statistic in frequentist statistics can be (mis)used.

For testing the difference of two proportions one can think of the z-statistic, the proportion test (with or without continuity correction), χ^2-test, Fisher's exact test, McNemar's test, logistic regression. Now choose the statistic that gives a low p-value, don't tell about other tests. [5]

3. If the test has a two-sided alternative, change the formulated to a one-sided alternative. This will allow you to halve the p-value.

4. Change what is accepted as publishable.

   Our last option, if you cannot lower your p-value any other way, is to change what is accepted as publishable. So, instead of a p-value of 0.05, use 0.10 and just state that this is the level you consider as statistically significant. I haven't seen any other number besides 0.10, however, so if your p-value is larger than this the best you can do is to claim that your results are "suggestive" or "in the expected direction". Don't scoff, because this sometimes works. You can really only get away with this in secondary and tertiary journals (which luckily are increasing in number) or in certain fields where the standard of evidence is low, or when your finding is one which people want to be true. This worked for second-hand smoking studies, for example, and currently works for anything said to be negatively associated with global warming.

This is about the right time to start reading the article "Abandon Statistical Significance" by McShane et al., which is at the very end of these lecture notes. It illustrates that there is much doubt on hypothesis testing using p-values as there are many questionable aspects on the procedure. Another account against the use of p-values for statistical evidence is Meester [2019] (in dutch).

Exercise 2.25 Explain how trick 3 works.

2.7 Confidence sets

A disadvantage of a point estimator is that no quantification about the uncertainty of the estimate is included. Say someone tells you upon repeated coin tossing the fraction of heads equals 0.6. This is a point-estimate, but anyone would have more confidence in asserting that the coin is not fair if 0.6 is the result of seeing 600 times heads out of 1000 throws compared to 6 out of 10 throws.

Confidence sets are random sets that contain the parameter with a predefined probability.

Definition 2.49. Let $g : \Omega \to G$ be a function, let η be the collection of all subsets of G, and let $R : \mathcal{X} \to \eta$ be a function. The function R is a level-γ **confidence set** for $g(\theta)$ if for every $\theta \in \Omega$

$$\mathrm{P}_\theta(g(\theta) \in R(X)) \geq \gamma$$

(an implicit assumption is that $\{x : g(\theta) \in R(x)\}$ is measurable). The confidence set is **exact** if $\mathrm{P}_\theta(g(\theta) \in R(X)) = \gamma$ for all $\theta \in \Omega$. If $\inf_{\theta\in\Omega} \mathrm{P}_\theta(g(\theta) \in R(X)) > \gamma$, the confidence set is **conservative**.

To highlight that it is really the set that is random (and not the parameter θ) one sometimes writes

$$\mathrm{P}_\theta(R(X) \ni g(\theta)) \geq \gamma$$

[5] Choosing just one favourable is cheating, but one can state all results. If these are in conflict, that may indicate that the evidence for the alternative hypothesis is perhaps not so strong.

as it is common to put the random quantity on the left in probability statements (this is of course not necessary). The interpretation of confidence sets is somewhat subtle: on repeated replication of the statistical experiment, the fraction of sets that contain the "true" parameter θ is at least γ; for a given dataset, the confidence set either contains θ, or not (in practice we don't know which of the two cases applies).

The construction of confidence sets is a bit of an art. It is super easy to find a conservative confidence set: just take Ω!. Naturally, we aim for a set that has confidence as close as possible to γ while being of minimal size. There are some tricks/procedures for constructing confidence sets:

1. find a pivotal quantity (sometimes also called pivot);
2. find an asymptotic pivot;
3. "invert" the critical region of a point-null hypothesis test;
4. use the bootstrap (a computer intensive simulation procedure).

In case a pivotal quantity exists, the construction of a confidence set is relatively easy

Definition 2.50. A **pivot** is a function $h : \mathcal{X} \times \Omega \to \mathbb{R}$ whose distribution does not depend on the parameter.

Suppose $h(X, \theta)$ is pivotal. If its distribution is known, then we can find constants c_ℓ and c_r such that

$$\mathrm{P}_\theta(c_\ell \leq h(X, \theta) \leq c_r) = \gamma \quad \text{for all } \theta \in \Omega. \tag{2.7}$$

Suppose for ease of exposition that $\Omega \subset \mathbb{R}$. If moreover $h(X, \theta)$ is of reasonably manageable form, we can rewrite this relation as

$$\mathrm{P}_\theta(L(X) \leq \theta \leq U(X)) = \gamma \quad \text{for all } \theta \in \Omega.$$

This would then deliver the exact level-γ confidence set $R(X) = [L(X), U(X)]$.

Example 2.51. Assume $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N\left(\theta, \sigma^2\right)$, where both θ and σ are unknown. Let

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Then, if $X = (X_1, \ldots, X_n)$,

$$h(X, \theta) = \sqrt{n}\frac{\bar{X}_n - \theta}{S_n} \sim t\,(n-1)$$

is pivotal. Hence, using the quantiles of the $t\,(n-1)$-distribution, we can find constants c_ℓ and c_r such that (2.7) holds under $P'_{(\theta,\sigma^2)}$. In this case we can rewrite this expression as

$$\mathrm{P}_{(\theta,\sigma^2)}\left(\bar{X}_n - c_r\frac{S_n}{\sqrt{n}} \leq \theta \leq \bar{X}_n - c_\ell\frac{S_n}{\sqrt{n}}\right) = \gamma$$

for all (θ, σ^2).

Exercise 2.26 Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Unif\,(0, \theta)$. Show that $X_{(n)}/\theta$ is pivotal and construct a confidence interval for θ of level γ.

Pivots need not exist. In that case one can try to find an asymptotic pivotal to derive an asymptotic confidence set. Suppose the MLE of a one-dimensional parameter is asymptotically normal:

$$\sqrt{nI(\theta_0)}(\hat{\Theta}_n - \theta_0) \rightsquigarrow N(0,1).$$

Suppose $\widehat{I(\theta_0)}$ is a weakly consistent estimator for $I(\theta_0)$. Then

$$\sqrt{n\widehat{I(\theta_0)}}(\hat{\Theta}_n - \theta_0) \rightsquigarrow N(0,1).$$

Hence, the interval

$$\left[\hat{\Theta}_n - \xi_{\alpha/2}\left(n\widehat{I(\theta_0)}\right)^{-1/2}, \hat{\Theta}_n + \xi_{\alpha/2}\left(n\widehat{I(\theta_0)}\right)^{-1/2}\right]$$

has asymptotic coverage probability $1-\alpha$.

One possibility is the plug-in estimator $\widehat{I(\theta_0)} = I(\hat{\Theta}_n)$. Results by Efron and Hinkley [1978], suggest that it is better to use the **observed Fisher information**, $\widehat{I(\theta_0)} := -n^{-1}l_n''(\hat{\Theta}_n)$ (in the sense that the coverage probability of the confidence interval is closer to $1-\alpha$ when using the observed Fisher information instead of the plug-in estimator for the Fisher information).

There is a direct relation between point null hypothesis testing and constructing confidence sets. The idea is that a $1-\alpha$ confidence set is obtained from those values θ_0 for which the hypothesis $H_0 : g(\theta) = \theta_0$ is not rejected at level α. As a consequence, confidence sets can be derived from hypothesis tests and vice versa.

Proposition 2.52 (Schervish [1995], proposition 5.48)**.** Let $g : \Omega \to G$ be a function.

- For each $y \in G$, let ϕ_y be a level-α nonrandomised test of $H_0 : g(\theta) = y$. Let $R(x) = \{y : \phi_y(x) = 0\}$. Then R is a level-$(1-\alpha)$ confidence set for $g(\theta)$.
- Let R be a level-$(1-\alpha)$ confidence set for $g(\theta)$. For each $y \in G$ define

  $$\phi_y(x) = \mathbf{1}_{\{y \notin R(x)\}}.$$

  Then, for each y, ϕ_y has level α as a test of $H_0 : g(\theta) = y$.

Proof.

- We need to show that $\mathrm{P}_\theta(g(\theta) \in R(X)) \geq 1-\alpha$. Now

  $$\mathrm{P}_\theta(g(\theta) \in R(X)) = \mathrm{P}_\theta(\phi_{g(\theta)}(X) = 0\}) = 1 - \mathrm{E}_\theta\, \phi_{g(\theta)}(X).$$

  The final equality holds since the test is nonrandomised and hence takes values in $\{0,1\}$. The result follows since $\sup_{g(\theta)=y} \mathrm{E}_\theta\, \phi_y(X) \leq \alpha$.

- We need to show that $\sup_{g(\theta)=y} \beta(\theta) \leq \alpha$. Now

  $$\beta(\theta) = \mathrm{E}_\theta\, \phi_y(X) = \mathrm{P}_\theta(y \notin R(X)) = 1 - \mathrm{P}_\theta(y \in R(X)).$$

  Hence

  $$\sup_{g(\theta)=y} \beta(\theta) = 1 - \inf_{g(\theta)=y} \mathrm{P}_\theta(y \in R(X)) = 1 - \inf_{g(\theta)=y} \underbrace{\mathrm{P}_\theta(g(\theta) \in R(X))}_{\geq 1-\alpha} \leq \alpha.$$

□

As in many settings no "best" hypothesis test exists, it is not surprising that constructing a "best" confidence interval (set) is far from trivial.

2.8 Some additional results on complete statistics*

Definition 2.53. A statistic T is said to be **complete** for $\theta \in \Omega$ if for any Borel function g, $\mathrm{E}_\theta\, g(T) = 0$ voor all θ, implies $g(T) = 0$, P_θ-a.s for all θ.

If T is complete and $S = h(T)$, then S is complete. The **Lehmann-Scheffé theorem** states that a complete sufficient statistic is minimal sufficient.

Theorem 2.54. Suppose T is sufficient and complete for $\theta \in \Omega$, then T is minimal sufficient.

Proof. Let S be a minimal sufficient statistic. Then there exists a measurable function h such that $S = h(T)$. Define $g(S) = \mathrm{E}[T \mid S]$, which does not depend on θ, as S is sufficient. Taking expectations, we get $\mathrm{E}_\theta[g(S)] = \mathrm{E}_\theta[\mathrm{E}[T \mid S]] = \mathrm{E}_\theta[T]$. Hence $\mathrm{E}_\theta[T - g(S)] = 0$ for all θ. By completeness, this implies $T = g(S)$, P_θ-a.s. As both $S = h(T)$ and $T = g(S)$, we conclude that S and T are one-to-one functions of each other. Hence T is minimal sufficient. □

Theorem 2.55 (Lehmann-Scheffé)**.** If T is a complete statistic, then all unbiased estimators of $g(\theta)$ that are functions of T alone are equal P_θ-a.s. for all θ. If there exists an unbiased estimator that is a function of a complete sufficient statistic, then it is UMVU.

Proof. Suppose both $\phi_1(T)$ and $\phi_2(T)$ are unbiased estimators for $g(\theta)$. Then

$$\mathrm{E}_\theta\left(\phi_1(T) - \phi_2(T)\right) = 0$$

for all θ and by completeness it follows that $\phi_1(T) = \phi_2(T)$, P_θ-a.s. This proves the first statement.

For the second statement, suppose there is an unbiased estimator $\phi(X)$ with finite-variance. Define

$$\phi_3(T) = \mathrm{E}[\phi(X) \mid T].$$

Then $\phi_3(T)$ is unbiased. Let $L(x) = (x - \theta)^2$. Then

$$\begin{aligned}\mathrm{Var}_\theta\, \phi_3(T) &= \mathrm{E}_\theta\, L(\phi_3(T)) = \mathrm{E}_\theta\, L([\mathrm{E}(\phi(X) \mid T]) \\ &\leq \mathrm{E}_\theta\, \mathrm{E}[L(\phi(X)) \mid T] = \mathrm{E}_\theta\, L(\phi(X)) = \mathrm{Var}_\theta\, \phi(X).\end{aligned}$$

The inequality follows from the conditional Jensen inequality and $x \mapsto L(x)$ being convex. □

Exercise 2.27 Suppose $X \sim Exp(\theta)$. Show that no unbiased estimator for θ exists. Proceed as follows:

1. Show that $\phi(X)$ is an unbiased estimator for θ if and only if
$$\int_0^\infty \phi(x) e^{-\theta x}\, \mathrm{d}x = 1.$$
2. Show that $X\phi(X) = 0$, P_θ almost surely.
Hint: use the fact that X is a complete and sufficient statistic for θ.
3. As this implies that $\phi(X) = 0$, conclude that $\phi(X)$ cannot be unbiased for θ.
4. Does an unbiased estimator exist for $1/\theta$ exist?

Chapter 3

The likelihood principle

In this chapter we discuss three principles: the likelihood- sufficiency- and conditionality principle.

3.1 Three principles in statistics

Suppose X has density $f_X(\cdot;\theta)$. If T is sufficient for θ, then by the factorisation theorem there exist measurable functions h and g_θ such that $f_X(x;\theta) = g_\theta(T(x))h(x)$. As will be discussed in section 6.8, "optimal" statistical procedures only depend on sufficient statistics (if the loss function is convex, this is part of the Rao-Blackwell theorem in section 6.8). In particular, when the model allows for a minimal sufficient statistic, we only have to consider procedures depending on this statistic. This motivates the following principle.

Definition 3.1 (**Weak Sufficiency principle (WSP)**)**.** Two observations x and x' factorising through the same value of a sufficient statistic T, that is, such that $T(x) = T(x')$, must lead to the same inference on θ.

The second principle that we discuss here can be attributed to Fisher [1959] and Barnard [1949] and was formalised by Birnbaum [1962]. In its definition, the notion of *information* is to be considered in the general sense of the collection of all possible inferences on θ.

Definition 3.2 (**Likelihood principle (LP)**)**.** The information brought by an observation x about θ is entirely contained in the likelihood function $L(\theta;x)$. Moreover, if x and x' are two observations depending on the same parameter θ (possibly in different experiments), such that there exists a constant c satisfying $L(\theta;x) = cL'(\theta;x')$ for every θ, they bring the same information about θ and must lead to identical inferences.

As Savage [1961] put it

> The likelihood principle says this: the likelihood function, long known to be a minimal sufficient statistic, is much more than merely a sufficient statistic, for given the likelihood function in which an experiment has resulted, *everything* else about the experiment – what its plan was, what different data might have resulted from it, the conditional distributions of statistics under given parameter values, and so on –is irrelevant.

Under this principle, quantities that depend on the sampling distribution of a statistic, which is in general not a function of the likelihood function alone, are irrelevant for statistical inference. This principle is only valid when

1. inference is about the same parameter θ;
2. θ includes *every* unknown factor in the model.

A classical illustration is the following example.

Example 3.3. Suppose 9 Bernoulli trials are observed with success, and 3 Bernoulli trials are observed with failure. In case we assume 12 observations were to be collected, then the number of successes X has the $Bin(12, \theta)$-distribution and the likelihood for $x = 9$ equals

$$L(\theta; x) = \binom{12}{9} \theta^9 (1-\theta)^3.$$

In case we assume we continued experimenting until we had obtained 3 failures, then the number of successes X' has the negative Binomial distribution: $X' \sim NegBin(3, \theta)$. In this case

$$L'(\theta; x') = \binom{11}{9} \theta^9 (1-\theta)^3.$$

The likelihood principle implies that inference on θ should be identical for both models. Hence, the rule which determines to stop gathering any more samples is irrelevant.

Now consider the problem of testing $H_0 : \theta = 1/2$ versus $H_1 : \theta > 1/2$. Assuming the Binomial distribution, the p-value equals

$$\left\{ \binom{12}{9} + \binom{12}{10} + \binom{12}{11} + \binom{12}{12} \right\} \left(\frac{1}{2}\right)^{12} \approx 0.073.$$

Assuming the Negative Binomial distribution, the p-value equals

$$\sum_{k=9}^{\infty} \left(\frac{1}{2}\right)^{k+3} \binom{k+2}{k} \approx 0.03.$$

Using significance level $\alpha = 0.05$, the calculated p-values lead to different conclusions. From this simple example we can clearly see that as the p-value depends on a tail probability, it violates the likelihood principle. This conclusion generalises to all frequentist hypothesis testing.

This example illustrates that classical hypothesis testing does not satisfy the LP.

Related to this example you may wish to read the first section of the article Lindley and Phillips [1976]. In this work the Bernoulli experiment consists of throwing a metal drawing pin (thumb tack) and observing whether it fell with the point uppermost (abbreviated to U) or with the point resting on the table, downwards (abbreviated to D). Regarding the criterion to stop sampling they remark:

> In other words, the significance (in the technical sense) to be associated with the hypothesis of equal chances depends heavily on what other results could have been achieved besides the 9 U's and 3 D's reported. Thus was (10,2) or (10,3) an alternative possibility? And yet it is rare for anyone to ask a scientist for this information.In fact, in the little experiment with the drawing pin I continued tossing until my wife said "Coffee's ready". Exactly how a significance test is to be performed in these circumstances is unclear to me.

Since the Bayesian approach is entirely based on the posterior distribution which only depends on x through the likelihood, the LP is automatically satisfied in a Bayesian setting [1]. However, when it comes to estimation, the LP does not imply Bayesian estimators, as for example maximum likelihood is a frequentist implementation of the LP.

[1] However, in case Jeffrey's prior is used in a Bayesian analysis, the likelihood principle is violated.

Surely not all statisticians accept the LP. Indeed, there is still controversy among researchers about its validity and implications. [2] However, Birnbaum [1962] proved that if you accept the WSP together with the following principle, then you necessarily must accept the LP. We give a precise statement in section 3.2.

Definition 3.4 (**Conditionality Principle (CP)**)**.** If two experiments on the parameter θ, $\mathcal{E}_1$ and $\mathcal{E}_2$, are available and if one of these two experiments is selected with probability p, the resulting inference on θ should only depend on the selected experiment.

Note that it is assumed that the probability p does not depend on θ. Both the WSP and CP appear to be reasonable. As they imply the LP this has far reaching consequences.

To illustrate the point of the conditionality principle, we consider an example. It is copied from the recent article by Gandenberger [2015] and resembles the example given in chapter 7.2 of Young and Smith [2005].

Example 3.5. Suppose you work in a laboratory that contains three thermometers, T1, T2, and T3. All three thermometers produce measurements that are normally distributed about the true temperature being measured. The variance of T1's measurements is equal to that of T2's but much smaller than that of T3's. T1 belongs to your colleague John, so he always gets to use it. T2 and T3 are common lab property, so there are frequent disputes over the use of T2. One day, you and another colleague both want to use T2, so you toss a fair coin to decide who gets it. You win the toss and take T2. That day, you and John happen to be performing identical experiments that involve testing whether the temperature of your respective indistinguishable samples of some substance is greater than 0°C or not. John uses T1 to measure his sample and finds that his result is just statistically significantly different from 0°C. John celebrates and begins making plans to publish his result. You use T2 to measure your sample and happen to measure exactly the same value as John. You celebrate as well and begin to think about how you can beat John to publication. "Not so fast", John says. "Your experiment was different from mine. I was bound to use T1 all along, whereas you had only a 50% chance of using T2. You need to include that fact in your calculations. When you do, you'll find that your result is no longer significant."

Gandenberger [2015] comments on this example

> According to radically 'behaviouristic' forms of frequentism, John may be correct. You performed a mixture experiment by flipping a coin to decide which of two thermometers to use, and thus which of two component experiments to perform. The uniformly most powerful level α test for that mixture experiment does not consist of performing the uniformly most powerful level α test for whichever component experiment is actually performed. Instead, it involves accepting probability of Type I error greater than α when T3 is used in exchange for a probability of Type I error less than α when T2 is used, in such a way that the probability of Type I error for the mixture experiment as a whole remains α (see Cox [1958], p. 360). Most statisticians, including most frequentists, reject this line of reasoning. It seems suspicious for at least three reasons. First, the claim that your measurement warrants different conclusions from John's seems bizarre. They are numerically identical measurements from indistinguishable samples of the same substance made using measuring instruments with the same stochastic properties. The only difference between your procedures is that John was 'bound' to use the thermometer he used, whereas you had a 50% chance of using a less precise thermometer. It seems odd to claim

[2] The topic of this chapter is complicated, but I feel you should at least have taken notice of the principles discussed. Somewhat confusingly, the three principles are not always stated in exactly the same form in the literature. I am not a specialist on this topic and have tried to summarise some of the main findings in an accessible way.

that the fact that you could have used an instrument other than the one you actually used is relevant to the interpretation of the measurement you actually got using the instrument you actually used. Second, the claim that John was "bound" to use T1 warrants scrutiny. Suppose that he had won that thermometer on a bet he made ten years ago that he had a 50% chance of winning, and that if he hadn't won that bet, he would have been using T3 for his measurements. According to his own reasoning, this fact would mean that his result is not statistically significant after all. The implication that one might have to take into account a bet made ten years ago that has nothing to do with the system of interest to analyse John's experiment is hard to swallow. In fact, this problem is much deeper than the fanciful example of John winning the thermometer in a bet would suggest. If John's use of T1 as opposed to some other thermometer with different stochastic properties was a nontrivial result of any random process at any point in the past that was independent of the temperature being measured, then the denial of Weak Conditionality Principle as applied to this example implies that John analysed his data using a procedure that fails to track evidential meaning. Third, at the time of your analysis you know which thermometer you received. How could it be better epistemically to fail to take that knowledge into account?

The following is a numerical case to this example. Suppose $\beta \in [0,1]$ and X has density

$$f(x) = \beta\phi(x;\mu,1) + (1-\beta)\phi(x;\mu,\sigma^2).$$

Assume σ^2 is known and for one experiment $\beta = 1$ and for the other experiment $\beta = 1/2$. Assume we wish to test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ at level $\alpha = 0.05$. In case $\beta = 1$, we reject when $|x| > 1.96$. If $\beta = 1/2$ we reject when $|x| > k$, where k solves

$$\frac{1}{2}\int_k^\infty \phi(x;0,1)\,\mathrm{d}x + \frac{1}{2}\int_k^\infty \phi(x;0,\sigma^2)\,\mathrm{d}x = 0.025.$$

This is equivalent to solving

$$\mathbb{P}(Z > k) + \mathbb{P}(\sigma Z > k) = 0.05,$$

where $Z \sim N(0,1)$. If $\sigma = 5$ then $k = 8.22$; if $\sigma = 0.5$, then $k = 1.65$. Suppose you observe in both experiments the value $x = 3$. Then H_0 is rejected in the non-mixture experiment, whereas it is not rejected in the mixture experiment when $\sigma = 5$.

We conclude this section with Birnbaum's result.

Theorem 3.6. The CP and WSP together imply LP.

This means that if you accept the intuitively reasonable principles CP and WSP, you necessarily accept the LP and its consequences.

Exercise 3.1 [YS exercise 7.4.] Suppose $X \sim N(\theta,1)$ or $X \sim N(\theta,4)$, depending on whether the outcome, Y, of tossing a fair coin is heads ($y = 1$) or tails ($y = 0$). It is desired to test $H_0 : \theta = -1$ against $H_1 : \theta = 1$. Show that the most powerful (unconditional) size $\alpha = 0.05$ test is the test with rejection region given by $x \geq 0.598$ if $y = 1$ and $x \geq 2.392$ if $y = 0$.

Suppose instead that we condition on the outcome of the coin toss in construction of the tests. Verify that, given $y = 1$, the resulting most powerful size $\alpha = 0.05$ test would reject if $x \geq 0.645$, while, given $y = 0$, the rejection region would be $x \geq 2.290$.

Exercise 3.2 ⟨Berger and Wolpert (1988), exercise 1.17 in Robert [2007].⟩ Consider an experiment with outcomes in $\{1, 2, 3\}$ and probability mass functions $f(\cdot \mid \theta)$, $\theta \in \{0, 1\}$ given by

x	1	2	3
$f(x; 0)$	0.9	0.05	0.05
$f(x; 1)$	0.1	0.05	0.85

1. Show that the procedure that rejects the hypothesis $H_0 : \theta = 0$ vs $H_1 : \theta = 1$ when $X \in \{2, 3\}$ has a probability of 0.9 to be correct (both under H_0 and H_1).
2. Now suppose you get the realisation $x = 2$. The frequentist test then rejects. Based on the likelihood ratio, is there strong evidence to reject?

Exercise 3.3 ⟨Cox (1958), exercise 1.20 in Robert [2007].⟩ In a research laboratory, a physical quantity θ can be measured by a precise but often busy machine, which provides a measurement $X_1 \sim N(\theta, 0.1)$, with probability 0.5, or through a less precise but always available machine, which gives $X_2 \sim N(\theta, 10)$. Suppose you obtain an observation and wish to make a confidence interval for θ, say with confidence 95%. If you take into account that both machines could have been selected, show that the half-width of the confidence interval equals 5.19, while the half-width of the confidence interval obtained from the precise machine equals 0.62.
Hint: In order to derive the half-width for the case where both machines could have been selected, first think about what the density will be in that case. See the previous page for an example of such a combined density. Given this density, try to derive the half-width.

Exercise 3.4 Show by means of an example that the principle of unbiased estimation does not respect the likelihood principle.
Hint: construct unbiased estimators for θ in example 3.3

3.2 Proof of Birnbaum's result*

This section is based on appendix B in Ghosh et al. [2006].

Before giving the proof, we recall a few definitions. A statistical experiment $\mathcal{E}$ is formally defined by the triplet

$$\mathcal{E} = (\mathcal{X}, \mathcal{B}, \{P_\theta, \theta \in \Omega\}).$$

A finite mixture of experiments $\mathcal{E}_1, \ldots, \mathcal{E}_k$ with mixture probabilities $\pi_1, \ldots, \pi_k$ (not depending on θ), is defined as a two-stage experiment where one first selects experiment $\mathcal{E}_i$ with probability π_i and then observes $x_i \in \mathcal{X}_i$ (see also example 1.23).

Let $\mathcal{E}_1 = (\mathcal{X}_1, \mathcal{A}_1, \{\mathrm{P}_{\theta,1}, \theta \in \Omega\})$ and $\mathcal{E}_2 = (\mathcal{X}_2, \mathcal{A}_2, \{\mathrm{P}_{\theta,2}, \theta \in \Omega\})$ be two experiments. Suppose $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$.

Definition 3.7. If one makes the same inference on θ if one performs $\mathcal{E}_1$ and observes x_1 or performs $\mathcal{E}_2$ and observes x_2, then we write

$$(\mathcal{E}_1, x_1) \sim (\mathcal{E}_2, x_2)$$

and say $(\mathcal{E}_1, x_1)$ and $(\mathcal{E}_2, x_2)$ are **equivalent**.

If we assume the measures $P_{\theta,1}$ and $P_{\theta,2}$ admit densities $f^{(1)}(\cdot;\theta)$ and $f^{(2)}(\cdot;\theta)$ respectively, then we can reformulate

LP $f^{(1)}(x_1;\theta) = c f^{(2)}(x_2;\theta)$ for all $\theta \in \Omega$ and some $c > 0$
$\Rightarrow (\mathcal{E}_1, x_1) \sim (\mathcal{E}_2, x_2)$.

WSP $S(x_i) = S(x_i')$ for some sufficient statistic S for θ
$\Rightarrow (\mathcal{E}_i, x_i) \sim (\mathcal{E}_i, x_i')$.

CP for the mixture experiment $\mathcal{E} = \sum_{i=1}^k \pi_i \mathcal{E}_i$, we have $(\mathcal{E}, (i, x_i)) \sim (\mathcal{E}_i, x_i)$ for any $i = 1, \ldots, k$ and $x_i \in \mathcal{E}_i$.

Note [LP] gives equivalence of different experiments, while [WSP] is concerned with a single experiment.

Proof of theorem 3.6. Assume $f^{(1)}(x_1;\theta) = c f^{(2)}(x_2;\theta)$ for all $\theta \in \Omega$ and some $c > 0$. Consider the mixture experiment

$$\mathcal{E} = \frac{1}{1+c}\mathcal{E}_1 + \frac{c}{1+c}\mathcal{E}_2.$$

Within the experiment $\mathcal{E}$, the points $(1, x_1)$ and $(2, x_2)$ have densities

$$\frac{1}{1+c} f^{(1)}(x_1;\theta) \quad \text{and} \quad \frac{c}{1+c} f^{(2)}(x_2;\theta)$$

respectively. These are the same as we assumed that $f^{(1)}(x_1;\theta) = c f^{(2)}(x_2;\theta)$.

This equality, together with the existence of a sufficient statistic S in the mixture experiment $\mathcal{E}$, imply that (use lemma 2.6) $S((1, x_1)) = S((2, x_2))$. By the sufficiency principle this gives

$$(\mathcal{E}, (1, x_1)) \sim (\mathcal{E}, (2, x_2)).$$

By the conditionality principle

$$(\mathcal{E}, (1, x_1)) \sim (\mathcal{E}_1, x_1) \quad \text{and} \quad (\mathcal{E}, (2, x_2)) \sim (\mathcal{E}_2, x_2)$$

Combining the preceding two displays gives $(\mathcal{E}_1, x_1) \sim (\mathcal{E}_2, x_2)$.

□

Chapter 4

Bayesian statistics

This chapter gives an introduction to the Bayesian approach to statistics. We discuss some justifications, such as exchangeability. We discuss prior specification, an essential requirement for Bayesian inference. Subsequently graphical and Bayesian models are connected by means of hierarchical models. In the final sections we discuss empirical Bayes methods and Bayesian asymptotics.

4.1 Setup

Bayesian statistics is based on the "axiom" that uncertainties can only be described with probability. As the Bayesian statistician D.V. Lindley put it:

> Whatever way uncertainty is approached, probability is the only sound way to think about it.

This means that, for anything we don't know, the only logical way to describe our beliefs is probability. In particular, the belief about the parameter θ in a statistical model should be described by a probability distribution. One way to view this distribution is that it reflects our state of knowledge (information) about the parameter before seeing any data. On the contrary, in classical statistics, θ is considered a fixed unknown parameter. As within Bayesian statistics the parameter is considered to be a random quantity, we denote it with a capital letter Θ. Its distribution μ_Θ is called the **prior distribution**. The density of the data X, is then in fact to be interpreted as the conditional density of X given Θ (with respect to some σ-finite measure ν). Hence, whereas we would write $f_X(x;\theta)$ in classical statistics, we now write $f_{X|\Theta}(x \mid \theta)$. Note that the "|"-sign really denotes "conditional on" and for that reason within classical statistics we wrote ";" to merely say "depends on θ as well". Now densities need not always exist, as illustrated in some of the examples that follow. For that reason, we first give a general definition of a Bayesian statistical model and afterwards specialise to the case where densities exist. The following definition turns out to be important.

Definition 4.1. A **Markov kernel** with source $(X, \mathcal{A})$ and target $(Y, \mathcal{B})$ is a map $\kappa : X \times \mathcal{B} \to [0,1]$ such that

- for each $B \in \mathcal{B}$, $x \mapsto \kappa(x, B)$ is $\mathcal{A}$-measurable;
- for each $x \in X$, $B \mapsto \kappa(x, B)$ is a probability measure on $(Y, \mathcal{B})$.

As a first example, suppose $X \sim \mathrm{P}_\theta$ with $\mathrm{P}_\theta \ll \nu$. Write $f(x;\theta) = \frac{\mathrm{dP}_\theta}{\mathrm{d}\nu}(x)$. Then $\mathrm{P}_\theta(B) = \int_B f(x;\theta)\,\mathrm{d}\nu(x)$. If $\theta \mapsto f(x;\theta)$ is measurable, then $\kappa(x, B) := \mathrm{P}_\theta(B)$ defines a Markov kernel. As a second example, suppose $\{Z_n\}_n$ is a discrete time Markov chain taking values in a countable set. Then (under a mild measurability condition) $\kappa(z, B) := \mathbb{P}(Z_{n+1} \in B \mid Z_n = z)$ is a Markov kernel.

4.1.1 Definition of a Bayesian statistical experiment

Suppose we are given the statistical experiment $\mathcal{E} = (\mathcal{X}, \mathcal{B}, \{\mathrm{P}_\theta,\ \theta \in \Omega\})$. The probabilities P_θ are sometimes called *sampling probabilities*, where θ is the *parameter*. Suppose Ω is endowed with a σ-field $\mathcal{B}_\Omega$ and probability measure μ_Θ, such that $\theta \mapsto \mathrm{P}_\theta(B)$ is measurable for each $B \in \mathcal{B}$. Then we can define the measure

$$\mu_{(\Theta,X)}(A \times B) = \int_A \mathrm{P}_\theta(B)\, \mathrm{d}\mu_\Theta(\theta) \tag{4.1}$$

on $\Omega \times \mathcal{X}$. A *Bayesian statistical experiment* is defined by

$$\mathcal{E} = \left(\Omega \times \mathcal{X}, \mathcal{B}_\Omega \otimes \mathcal{B}, \mu_{(\Theta,X)}\right).$$

Hence, compared to the definition of a statistical experiment there is additionally the prior that is part of the definition. The decomposition in (4.1) says informally that to sample from (Θ, X) one first samples θ from μ_Θ and next x from $\mathrm{P}_\theta(\cdot)$. In Bayesian statistics we are interested in the "reverse" way of sampling: first sample x from its marginal, followed by sampling θ conditional on x. By marginalisation, we can define the **predictive measure** $\mu_X(B) = \mu_{(\Theta,X)}(\Omega \times B)$. If there exists a Markov kernel Π such that

$$\mu_{(\Theta,X)}(A \times B) = \int_B \Pi_x(A) \mu_X(\mathrm{d}x). \tag{4.2}$$

then the Bayesian statistical experiment is called **regular**. In that case we have

$$\int_A \mathrm{P}_\theta(B) \mu_\Theta(\mathrm{d}\theta) = \int_B \Pi_x(A) \mu_X(\mathrm{d}x) \tag{4.3}$$

and Π is called a **version of the posterior distribution**. Equation (4.3) defines the Markov kernel Π by **disintegration**. We sometimes write $\Pi_x(A) = \mu_{\Theta|X}(A \mid x)$ to highlight that the posterior gives the conditional distribution of the parameter, conditional on the data.

Summarising: in a Bayesian statistical experiment we two ingredients: *(i)* the sampling probabilities $\mathrm{P}_\theta(B)$ (which define an "ordinary", non-Bayesian statistical experiment) and *(ii)* the prior measure μ_Θ. These define the measure (4.2). Then μ_X is defined by marginalisation and the posterior Π_x by disintegration (Equation (4.3)).

Important note on notation: the notation P_θ, μ_Θ, $\mu_{(\Theta,X)}$, μ_X and Π_x will be used at many instances. In particular, the prior and posterior measure are denoted by μ_Θ and Π_x respectively.

4.1.2 Dominated Bayesian statistical models

If we assume existence of densities, deriving the posterior becomes easier. Hence, suppose $\mathrm{P}_\theta \ll \nu$ for all $\theta \in \Omega$ and denote the density, also known as the likelihood, by $L(\theta; x)$. Then

$$\begin{aligned}\mu_{(\Theta,X)}(A \times B) = \int_A \mathrm{P}_\theta(B) \mu_\Theta(\mathrm{d}\theta) &= \int_A \int_B L(\theta; x) \nu(\mathrm{d}x) \mu_\Theta(\mathrm{d}\theta) \\ &= \int_B \int_A L(\theta; x) \mu_\Theta(\mathrm{d}\theta) \nu(\mathrm{d}x)\end{aligned}$$

Since

$$\mu_X(B) = \int_B f_X(x) \nu(\mathrm{d}x) \quad \text{with} \quad f_X(x) = \int_\Omega L(\theta; x) \mu_\Theta(\mathrm{d}\theta)$$

we get

$$\mu_{(\Theta,X)}(A \times B) = \int_B \int_A \frac{L(\theta; x)}{f_X(x)} \mu_\Theta(\mathrm{d}\theta) \mu_X(\mathrm{d}x) \tag{4.4}$$

(provided that $f_X(x) > 0$; see Remark 4.2). First, this shows that the measure $\mu_\Theta \otimes \mu_X$ acts in a natural way as a dominating measure:

$$\frac{\mathrm{d}\mu_{(\Theta,X)}}{\mathrm{d}(\mu_\Theta \otimes \mu_X)}(\theta, x) = \frac{L(\theta; x)}{f_X(x)}. \tag{4.5}$$

Secondly, from (4.4) it follows that the posterior is given by

$$\Pi_x(A) = \int_A \frac{L(\theta; x)}{f_X(x)} \mu_\Theta(\mathrm{d}\theta).$$

If we additionally assume $\mu_\Theta \ll \xi$ and denote $\frac{\mathrm{d}\mu_\Theta}{\mathrm{d}\xi}(\theta) = f_\Theta(\theta)$ then

$$\Pi_x(A) = \int_A \frac{L(\theta; x) f_\Theta(\theta)}{\int_\Omega L(\theta; x) f_\Theta(\theta) \xi(\mathrm{d}\theta)} \xi(\mathrm{d}\theta).$$

Clearly, the posterior depends on the data only via the likelihood: the likelihood principle is satisfied.

Considering x fixed, we see from this expression that the posterior measure is dominated by the measure ξ and has density

$$f_{\Theta|X}(x \mid \theta) := \frac{\mathrm{d}\Pi_x}{\mathrm{d}\xi}(\theta) = \frac{L(\theta; x) f_\Theta(\theta)}{\int_\Omega L(\theta; x) f_\Theta(\theta) \xi(\mathrm{d}\theta)}.$$

Note that the term on the denominator does not depend on θ. Therefore, one often sees this equation written as

$$f_{\Theta|X}(x \mid \theta) \propto L(\theta; x) f_\Theta(\theta)$$

which reads

$$\boxed{\text{posterior density} \propto \text{likelihood} \times \text{prior density}}.$$

This is the formula often seen in introductory books on Bayesian statistics.

Remark 4.2. This is a bit of a technical remark. One may wonder what happens if $f_X(x) = 0$ in (4.4). Formally, rather than (4.5), we can define

$$\frac{\mathrm{d}\mu_{(\Theta,X)}}{\mathrm{d}(\mu_\Theta \otimes \mu_X)}(\theta, x) = k(\theta, x)$$

with

$$k(\theta, x) = \begin{cases} \frac{L(\theta;x)}{f_X(x)} & \text{if} \quad f_X(x) \neq 0 \\ c & \text{if} \quad f_X(x) = 0 \end{cases}$$

where $c \in \mathbb{R}$ is arbitrary. Let $C_0 = \{x : \ f_X(x) = 0\}$. This is valid, since

$$\mu_X(C_0) = \int_{C_0} \mu_X(\mathrm{d}x) = \int_{C_0} f_X(x) \nu(\mathrm{d}x) = 0.$$

4.1.3 Examples: dominated case

Example 4.3. Assume $X_1, \ldots, X_n \mid \Theta = \theta \overset{\text{ind}}{\sim} Ber(\theta)$ and $\Theta \sim Be(\alpha, \beta)$. Then

$$f_{\Theta|X}(\theta \mid x) \propto \theta^S (1-\theta)^{n-S} \theta^{\alpha-1} (1-\theta)^{\beta-1} \mathbf{1}_{[0,1]}(\theta).$$

where $X = (X_1, \ldots, X_n)$ and $S = \sum_{i=1}^n X_i$. It is easily seen that the posterior has the $Be(S + \alpha, n - S + \beta)$-distribution. Note that the **posterior mean** is given by

$$\mathbb{E}[\Theta \mid X] = \frac{S + \alpha}{n + \alpha + \beta}$$

and that it behaves like $\bar{X}_n$ when n is large.

The following example is very important. It shows that a standard measurement model with the Normal distribution is fully tractable, if the variance is assumed to be known and a Gaussian prior on the mean is used.

Example 4.4. Suppose

$$\begin{aligned} X_1, \dots, X_n \mid \Theta = \theta &\overset{\text{ind}}{\sim} N\left(\theta, \sigma^2\right) \\ \Theta &\sim N\left(\mu_0, \sigma_0^2\right) \end{aligned}$$

where we assume σ^2 to be known. The parameter μ_0 and σ_0^2 are also assumed to be known and are part of the prior specification. If we set $X = (X_1, \dots, X_n)$, then

$$\Theta \mid X \sim N(\mu_1, \sigma_1^2)$$

with

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}. \tag{4.6}$$

and

$$\mu_1 = \frac{1/\sigma_0^2}{1/\sigma_1^2}\mu_0 + \frac{n/\sigma^2}{1/\sigma_1^2}\bar{X}_n = w_n\mu_0 + (1 - w_n)\bar{X}_n,$$

where $w_n = \sigma^2/(\sigma_0^2 + n\sigma^2)$ is the ratio fo the prior over the posterior precision. Equation (4.6) reads: posterior precision equals prior precision + data precision (precision being defined a the inverse of variance). Note that the posterior mean is a convex combination of the prior mean μ_0 and the sample average $\bar{X}_n$ and that $w_n \to 0$ if $n \to \infty$. Note that for large values of n, $\sigma_1^2 \approx \sigma^2/n$ and hence $\mu_1 \approx \bar{X}_n$. One says that as the sample size gets large the prior gets "washed away".

Exercise 4.1 Verify the calculations of the preceding example.

Exercise 4.2 Suppose $X_1, \dots, X_n \mid \Theta = \theta \overset{\text{iid}}{\sim} Pois(\theta)$ and $\Theta \sim Ga(\alpha, \beta)$. Derive the posterior distribution and compute the posterior mean.

4.1.4 An example where the posterior is not dominated by the prior*

In dominated models, the posterior is absolutely continuous with respect to the prior. The following example shows how non-dominated models can show up. This is typically the case when the statistical observation carries a piece of deterministic information about the parameter θ.

Assume

$$X = \begin{cases} \theta & \text{with probability} \quad p \\ \theta + Z & \text{with probability} \quad 1 - p \end{cases}$$

Here, Z has density f_Z with respect to Lebesgue measure. So in this model an observation is either exactly equal to the parameter θ or additive noise is superimposed. Now

$$\mathrm{P}_\theta(B) = \mathbb{P}(X \in B) = p\mathbf{1}_B(\theta) + (1 - p)\int_B f_Z(x - \theta)\,\mathrm{d}x. \tag{4.7}$$

Let μ_Θ denote the prior measure and assume it has density f_Θ. We have

$$\begin{aligned}\mu_{(\Theta,X)}(A\times B) &= \int_A \left(p\mathbf{1}_B(\theta) + (1-p)\int_B f_Z(x-\theta)\,\mathrm{d}x \right) f_\Theta(\theta)\,\mathrm{d}\theta \\ &= \int_{A\cap B} p f_\Theta(\theta)\,\mathrm{d}\theta + \int_B \int_A (1-p) f_Z(x-\theta) f_\Theta(\theta)\,\mathrm{d}\theta\,\mathrm{d}x \\ &= \int_B \left(p f_\Theta(x)\mathbf{1}_A(x) + (1-p)\int_A f_Z(x-\theta) f_\Theta(\theta)\,\mathrm{d}\theta \right) \mathrm{d}x\end{aligned}$$

The second equality follows from Fubini's theorem and the final equality follows from

$$\int_{A\cap B} p f_\Theta(\theta)\,\mathrm{d}\theta = \int_B p f_\Theta(\theta)\mathbf{1}_A(\theta)\,\mathrm{d}x = \int_B p f_\Theta(x)\mathbf{1}_A(x)\,\mathrm{d}x.$$

Since $\mu_X(B) = \mu_{(\Theta,X)}(\Omega\times B)$, this implies that

$$\mu_X(B) = \int_B f_X(x)\,\mathrm{d}x \quad \text{with} \quad f_X(x) = \left(p f_\Theta(x)\mathbf{1}_\Omega(x) + (1-p)\int_\Omega f_Z(x-\theta) f_\Theta(\theta)\,\mathrm{d}\theta \right).$$

Thus we have

$$\mu_{(\Theta,X)}(A\times B) = \int_B \frac{p f_\Theta(x)\mathbf{1}_A(x) + (1-p)\int_A f_Z(x-\theta) f_\Theta(\theta)\,\mathrm{d}\theta}{f_X(x)} \mu_X(\,\mathrm{d}x).$$

Therefore, the posterior is given by

$$\Pi_x(A) = \frac{p f_\Theta(x)\mathbf{1}_A(x) + (1-p)\int_A f_Z(x-\theta) f_\Theta(\theta)\,\mathrm{d}\theta}{p f_\Theta(x)\mathbf{1}_\Omega(x) + (1-p)\int_\Omega f_Z(x-\theta) f_\Theta(\theta)\,\mathrm{d}\theta}.$$

In particular, the posterior measure assigns probability $pm(x)/f_X(x)$ to $\{x\}$. Now whereas the prior is absolutely continuous with respect to Lebesgue measure, the posterior is not.

4.1.5 Basu's example

The example below is taken from Basu [1975]. The main learning goal of this example is that in "non-standard" examples the maximum likelihood estimator and estimators derived from the Bayesian point of view can behave very differently.

Suppose a ticket is drawn from an urn with 1000 tickets of which 980 are labelled 10θ and 20 labelled θ. Based on the random draw from the urn we wish to estimate θ. Just to imagine, suppose the ticket drawn reads $x = 5$. Then we know that either $\theta = 5$ or $10\theta = 5$, the latter being equivalent to $\theta = 1/2$. We know that the estimate $\hat\theta = x/10$ is correct in 98% of the cases, while $\hat\theta = x$ only in 2% of the cases. Hence, we have a "good" estimator using $\hat\Theta = X/10$. This is in fact the MLE.

Now we adjust the problem slightly, and change the label 10θ on each of the 980 tickets by $a_i\theta$ for the i-th ticket ($1 \le i \le 980$), where $a_i \in (9.9, 10.1)$, all distinct. Reasoning similarly as before we now have $\hat\Theta = X$ which is correct only with probability 0.02 (2%). It seems strange that a slight adjustment of the problem leads to completely different performance of the estimator, while using the same principle to choose the estimator (maximum likelihood). The example can easily be made more extreme by taking $a_i \in (9.999, 10.0001)$ for example.

Let's consider a Bayesian approach where we assume Θ gets assigned a prior with density $m(\theta)$ (with respect to Lebesgue-measure) that is supported on $[0,\infty)$, say. The problem can generally be stated as drawing

a ticket from an urn with tickets labeled $a_i\theta$, where the i-th ticket is chosen with probability p_i, $1 \leq i \leq N$ (the preceding simply corresponds to $p_i = 1/N$). We have

$$\mathrm{P}_\theta(B) = \sum_{i=1}^{N} p_i \mathbf{1}_B(a_i\theta).$$

Then

$$\begin{aligned}
\mu_{(\Theta,X)}(A \times B) &= \int \mathbf{1}_A(\theta)\mathrm{P}_\theta(B)m(\theta)\,\mathrm{d}\theta \\
&= \sum_{i=1}^{N} p_i \int \mathbf{1}_A(\theta)\mathbf{1}_B(a_i\theta)m(\theta)\,\mathrm{d}\theta \\
&= \sum_{i=1}^{N} p_i \int_B \mathbf{1}_A(u_i/a_i)m(u_i/a_i)a_i^{-1}\,\mathrm{d}u_i \\
&= \sum_{i=1}^{N} p_i \int_B \mathbf{1}_A(x/a_i)m(x/a_i)a_i^{-1}\,\mathrm{d}x \\
&= \int_B \sum_{i=1}^{N} p_i \mathbf{1}_A(x/a_i)m(x/a_i)a_i^{-1}\,\mathrm{d}x = \int_B \sum_{i=1}^{N} w_i(x)\mathbf{1}_A(x/a_i)\,\mathrm{d}x
\end{aligned}$$

using the substitution $u_i = a_i\theta$ at the third equality sign and where $w_i(x) = p_i m(x/a_i)a_i^{-1}$. Hence (take $A = (0, \infty)$)

$$\mu_X(B) = \int_B f_X(x)\,\mathrm{d}x \quad \text{with} \quad f_X(x) = \sum_{i=1}^{N} w_i(x).$$

Verify yourself that $\int f_X(x)\,\mathrm{d}x = 1$, as should. This gives

$$\mu_{(\Theta,X)}(A \times B) = \int_B \frac{\sum_{i=1}^{N} w_i(x)\mathbf{1}_A(x/a_i)}{\sum_{i=1}^{N} w_i(x)}\mu_X(\,\mathrm{d}x).$$

If we define $\bar{w}_j(x) = w_j(x)/\sum_{i=1}^{N} w_i(x)$, then we can conclude from this we conclude that that the posterior measure Π is given by

$$\Pi_x(A) = \sum_{i=1}^{N} \bar{w}_i(x)\mathbf{1}_A(x/a_i).$$

That is, the posterior is concentrated on the set set $\{x/a_1, \ldots, x/a_N\}$ and the mass on $\{x/a_j\}$ equals $\bar{w}_j(x)$. As the observation brings deterministic information about the parameter, the posterior is not absolutely continuous with respect to the prior. On the contrary, the prior and posterior "live" on different sets: the measures are *singular*.

Note that if $\theta \mapsto m(\theta)$ is continuous, $(a_1, \ldots, a_N) \mapsto \Pi_x\left(\{x/a_j\}\right)$ is continuous. In other words, the posterior will be effected in a smooth way when adjusting the problem slightly. The posterior mean estimator

$$\tilde{\Theta} := \int \theta\theta\Pi_X(\,\mathrm{d}\theta) = \sum_{j=1}^{N} \bar{w}_j(X)\frac{X}{a_j}$$

will inherit this smooth behaviour. Note that completely different behaviour when compared to the MLE. From this example we clearly see that the MLE and posterior mean can behave radically different: in the former case we look for a maximiser whereas in the latter case we average over the parameter space.

Exercise 4.1. Verify that $w_i(x)$ can be interpreted as the probability of $\{X = a_i\theta\}$.

Exercise 4.2. There is an alternative way to write the model where we set $X = \alpha\Theta$ and (α, Θ) gets assigned the prior distribution

$$\begin{aligned}\Theta \mid A = a \sim a\Theta \quad &\text{with } \Theta \text{ having density } m \\ A \sim \mathrm{Cat}((1,\dots,N),&(p_1,\dots,p_N)),\end{aligned}$$

the second line denoting the categorical distribution that puts mass p_i on i. Show using marginalisation that

$$\mu_\Theta(C) = \sum_{i=1}^{N} \int_C w_i(x)\,\mathrm{d}x.$$

4.1.6 Prediction

The Bayesian paradigm also accommodates inference about future observables in a natural way. If Y denotes a future observation, then in a dominated Bayesian statistical experiment the **posterior predictive density** of Y is given by

$$f_{Y|X}(y \mid x) = \int f_{\Theta,Y|X}(\theta, y \mid x)\xi(\,\mathrm{d}\theta) = \int \frac{f_{X,Y|\Theta}(x, y \mid \theta) f_\Theta(\theta)}{f_X(x)} \xi(\,\mathrm{d}\theta). \tag{4.8}$$

The formula simplifies in case one assumes that X and Y are independent, conditional on Θ. In that case

$$f_{Y|X}(y \mid x) = \int f_{Y|\Theta}(y \mid \theta) f_{\Theta|X}(\theta \mid x)\xi(\,\mathrm{d}\theta).$$

Alternatively, one could make Y part of Θ, obtain the posterior distribution and marginalise out the components of Θ that do not involve Y.

A typical setting here would involve a sequence of conditionally independent random variables $\{X_i\}_{i=1}^{m+n}$ given Θ, where $X = (X_1, \dots, X_n)$ and $Y = (X_{n+1}, \dots, X_{n+m})$. Another important setting is that of state-space models, which we will discuss later on. Note that the preceding display shows that in prediction we average over the posterior. So rather than fixing one value of the parameter (as is usually done in prediction for frequentist statistics) uncertainty on the parameter is properly taken into account.

Example 4.5 (Laplace's Law of Succession)**.** Laplace considered the problem of determining the probability that the sun will rise tomorrow on the assumption that it has risen n times in succession.

He made the following assumptions:

1. The probability of the sun rising on any day is constant and unknown.
2. This unknown probability is a random variable Θ which is uniformly distributed on $[0, 1]$. This reflects the total ignorance of the probability of the sun rising (or not).
3. Successive sunrises are independent events (actually, independent conditional on the parameter Θ.

This lead Laplace to the following model

$$X_1, \dots, X_n \mid \Theta = \theta \overset{\text{ind}}{\sim} Ber(\theta) \qquad \Theta \sim Unif(0, 1).$$

If $X = (X_1, \dots, X_n)$, then

$$f_{\Theta|X}(\theta \mid x) \propto \theta^S (1-\theta)^{n-S} \mathbf{1}_{[0,1]}(\theta),$$

where $S = \sum_{i=1}^{n} X_i$. Hence the posterior follows from $\Theta \mid X \sim Be(S+1, n-S+1)$. Using equation (4.8) we get

$$f_{X_{n+1}|X}(1 \mid x) = \int_0^1 \theta(n+1)\theta^n \, \mathrm{d}\theta = \mathbb{E}[\Theta \mid X] = \frac{n+1}{n+2},$$

where $x \in \mathbb{R}^n$ is the vector with all elements equal to 1. Laplace' intuition for this result was that we essentially have 2 extra observations, as we assume that both "rise" and "not rise" can happen. This explains the factor $n+2$ in the denominator.

A more general point of view is to consider θ and x containing all unobserved and observed variables respectively. Then the Bayesian point of view can be summarised as deriving the distribution of the unobserved variables conditional on the observed variables. In prediction, that amounts to added future values y to θ. Then we would have a Bayesian experiment with unobserved variable $\theta' := (\theta, y)$ and observed variable x. Then the predictive distribution is simply the y-marginal of the distribution of $\theta' \mid x$. From this one sees that it is not always entirely clear what should be called likelihood or prior (is the distribution on y like a prior or likelihood?). *The main point is that the distinction between likelihood and prior is irrelevant: the important distinction is between observed and non-observed variables.* We further illustrate this in the upcoming section.

Important note on Bayesian notation. Often, "Bayesian notation" is used. This amount to writing $p(y \mid x)$ instead of $f_{Y|X}(y \mid x)$. Be aware that this can be a bit tricky: as an example, the expression $p(y^2 \mid x)$ is meant to denote $f_{Y^2|X}(y^2 \mid x)$ and NOT $f_{Y|X}(y^2 \mid x)$. An advantage of this notation is that it enhances readability of formulas.

4.1.7 Do we need to discern between prior, likelihood, data, parameters?

The frequentist distinction between data (that are treated as realisations of random quantities) and parameters (considered fixed) is absent in Bayesian statistics. Rather, one can view the approach as a way to specify the joint distribution of parameters and data. More loosely, both parameters and data are variables that have a distribution. Inference is then prescribed to be based on conditioning on observed variables. As an example, consider for the model

$$\begin{aligned} Y_i \mid X_i = x_i, \mu, \sigma^2 &\overset{\text{ind}}{\sim} N\left(\mu' x_i, \sigma^2\right) \\ X_1, \ldots, X_n \mid \beta &\overset{\text{iid}}{\sim} f_\beta \\ \mu, \beta, \sigma &\overset{\text{ind}}{\sim} g_{\mu,\beta,\sigma} \end{aligned}$$

Here, we do not fully specify the densities f and g. The third line though imposes the restriction that μ, β and σ are apriori independent. This is a Bayesian analogue of the regression model. In frequentist statistics, one usually directly writes down the conditional distribution of $Y_i \mid X_i$, assuming the X_i are fixed. Why? First of all, it is convenient, as it may be difficult to find a distribution for the X_i, especially when this is a high-dimensional vector. Now let's see what the Bayesian approach gives us. Note that the following variables are involved (I switch to Bayesian notation, writing all variables in lower-case): $y_1, \ldots, y_n, x_1, \ldots, x_n, \mu, \beta, \sigma^2$. The variables that are observed are $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$. Hence, the posterior distribution satisfies

$$\begin{aligned} p(\mu, \beta, \sigma^2 \mid y_1, \ldots, y_n, x_1, \ldots, x_n) &\propto p(\mu, \beta, \sigma^2) \prod_{i=1}^{n} p(y_i \mid x_i, \mu, \sigma^2) p(x_i \mid \beta) \\ &= \left(p(\beta) \prod_{i=1}^{n} p(x_i \mid \beta) \right) \left(p(\mu) p(\sigma^2) \prod_{i=1}^{n} p(y_i \mid x_i, \mu, \sigma^2) \right) \end{aligned}$$

Hence, because the model prescribes that β and (μ, σ^2) are statistically independent (which is an assumption made by the statistician), the posterior distribution of (β, σ^2) can be obtained by only specifying the conditional distribution of $y_i \mid x_i$. This sheds light on the underlying assumption that explains why we can think of $x_1, \dots, x_n$ as being fixed.

It gets more interesting when there are "missing data", a somewhat strange name for referring to the case where for some couple (x_j, y_j), either x_j or y_j is not observed. If y_j is not observed, then the posterior density is $p(\mu, \beta, \sigma^2, y_j \mid x_1, \dots x_n, \{y_i,\ i \neq j\})$, which is proportional to the expression in the preceding display. So in this case, a posteriori, (y_j, μ, σ^2) and β are independent. Now would you call y_j a parameter, observation, missing observation,...? The point is, it is not necessary to think about this. [1] We simply write the hierarchical model as above (we come back to hierarchical modelling later in this chapter) and condition on observed variables.

Exercise 4.3 What if some x_j is not observed, where y_j is observed. Do we still have posterior independence?

4.1.8 Bayesian updating

The posterior can be updated as data arrive sequentially in time. The following formulas are sometimes referred to as **Bayesian updating**. Assume that for $X_1, \dots, X_n$ are conditionally independent given Θ. If we define $\underline{x}_n = (x_1, \dots, x_n)$, then

$$p(\theta \mid \underline{x}_{n+1}) = p(\theta \mid \underline{x}_n, x_{n+1}) = \frac{p(\theta, x_{n+1} \mid \underline{x}_n)}{p(x_{n+1} \mid \underline{x}_n)}$$
$$= \frac{p(x_{n+1} \mid \theta, \underline{x}_n) p(\theta \mid \underline{x}_n)}{p(x_{n+1} \mid \underline{x}_n)} = \frac{p(x_{n+1} \mid \theta) p(\theta \mid \underline{x}_n)}{p(x_{n+1} \mid \underline{x}_n)}$$

Hence

$$p(\theta \mid \underline{x}_{n+1}) \propto p(x_{n+1} \mid \theta) p(\theta \mid \underline{x}_n)$$

which reveals that the posterior based on n observations serves as a prior for the next observation. Intuitively this makes much sense and this also explains why Bayesian inference easily incorporates sequential processing of data. Moreover, even if one observes $\underline{x}_n$, it may be computationally advantageous to compute the posterior sequentially using Bayesian updating.

4.1.9 Posterior mean, median, credible sets

From the posterior distribution, various quantities can be derived. Recall the posterior is denoted Π_x.

Definition 4.6. The **posterior mean** is defined by

$$\mathbb{E}[\Theta \mid X] = \int_\Omega \theta \Pi_X(\mathrm{d}\theta).$$

Definition 4.7. Assume θ is one-dimensional. The **posterior γ-quantile** is defined as the γ-quantile of the posterior distribution.

[1] In my opinion, thinking about variables and dividing these into observed and nonobserved variables is a very useful thing. I was much influenced by the talk https://www.youtube.com/watch?v=yakg94HyWdE&t=6s by Richard McElreath who works in anthropology and is the author of a much acclaimed book on Bayesian statistics.

Both the posterior mean and posterior median are point estimators. The Bayesian equivalent of a confidence set is a credible set, which is a set with pre described posterior probability.

Definition 4.8. The set A is a **level γ-credible set** for θ if $\Pi_X(A) \geq \gamma$.

Contrary to confidence sets, interpretation of credible sets is straightforward. Nevertheless, just like confidence sets, credible sets are not unique. It is not obvious how to choose a credible set when for example the posterior density is multimodal.

The previous three summary measures of the posterior distribution appear a bit at hoc. A formal way to derive these uses a loss function (details are in chapter 6).

4.1.10 An example

Example 4.9. This example is taken from Berger [2006]. Suppose we are dealing with a medical problem where within a population the probability that someone has a particular disease is given by θ_0. Hence, if $D = \{\text{patient has the disease}\}$ then $\theta_0 = \mathbb{P}(D)$. A diagnostic reading results in either a positive (P) or negative (N) reading. Let

$$\theta_1 = \mathbb{P}(P \mid D) \qquad \theta_2 = \mathbb{P}(P \mid D^c).$$

By Bayes' theorem:

$$\psi = \mathbb{P}(D \mid P) = \frac{\theta_0\theta_1}{\theta_0\theta_1 + (1-\theta_0)\theta_2} = g(\theta_0, \theta_1, \theta_2) \quad \text{(say)}. \tag{4.9}$$

The statistical problem is as follows: based on data $X_i \sim Bin\left(n_i, \theta_i\right)$, $i = 0, 1, 2$ (arising from medical studies), find a $100(1-\alpha)\%$ confidence or credible set for θ. At first sight this may seem like a very simple problem, but it is not straightforward how to construct a confidence set from a classical perspective.

Within the Bayesian framework, all unknowns quantities get assigned a probability distribution. We choose

$$X_i \mid \Theta_i = \theta_i \overset{\text{ind}}{\sim} Bin\left(n_i, \theta_i\right)$$
$$\Theta_i \overset{\text{ind}}{\sim} Be\left(a, b\right).$$

It is easy to verify that $\Theta_i \mid X_i \sim Be\left(X_i + a, n_i - X_i + b\right)$, $i = 0, 1, 2$ (independently). To construct the desired credible set, we use Monte Carlo simulation:

1. For $i \in \{0, 1, 2\}$, draw $\Theta_i \mid X_i = x_i \sim Be\left(x_i + a, n_i - x_i + b\right)$ to obtain realisations $\theta_0, \theta_1, \theta_2$.
2. Set $\psi = g(\theta_0, \theta_1, \theta_2)$.
3. Repeat steps (1) and (2) a large number of times (say B times) to obtain from step (2) the numbers $\psi^{(1)}, \ldots, \psi^{(B)}$.

Finally, use the $\alpha/2$-th upper and lower quantiles of $\psi^{(1)}, \ldots, \psi^{(B)}$ to form the desired credible set.

The use of **Monte Carlo simulation** in this example features more generally in Bayesian statistics. Only in very specific situations the posterior density can be computed in closed form.

In Figure 4.1 we present histograms of posterior samples when the data are simulated with

$$(n_0, \theta_0) = (17{,}000, 0.01) \qquad (n_1, \theta_1) = (10, 0.9) \qquad (n_2, \theta_2) = (100, 0.05)$$

leading to $x_0 = 163$, $x_1 = 8$ and $x_2 = 4$. The prior parameters were taken $a = b = 1$. The value of ψ corresponding to the θ-values we used for generating the data equals 0.154, which is well within the topleft

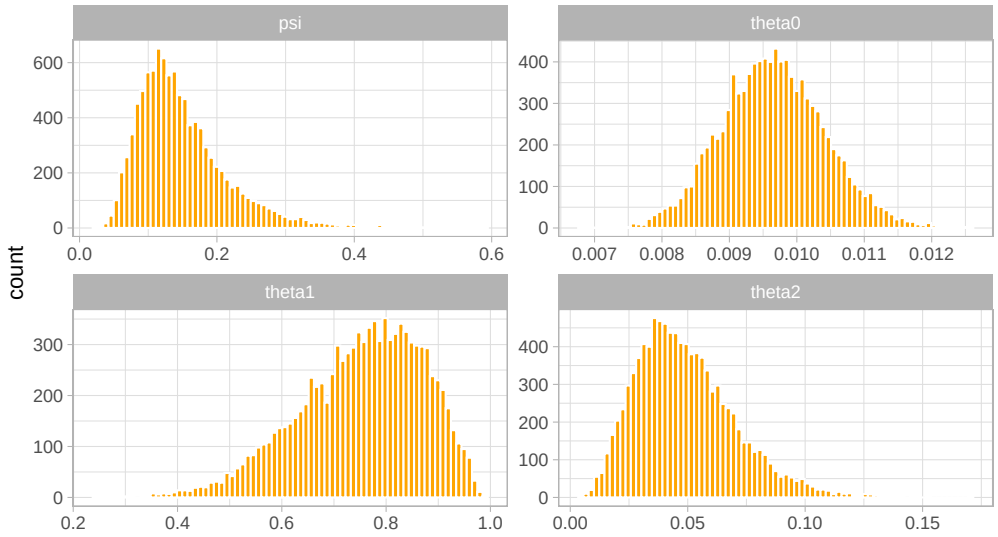

Figure 4.1: Numerical illustration to Example 4.9. Here, for generating the data we took $(n_0, \theta_0) = (17,000, 0.01)$, $(n_1, \theta_1) = (10, 0.9)$ and $(100, 0.05)$ leading to data $x_0 = 163$, $x_1 = 8$ and $x_2 = 4$. The results are based on $10,000$ Monte-Carlo draws from the posterior. Note that the data-generating value of ψ equals 0.154.

histogram. Note that the histogram for θ_0 is very peaked around 0.01, which should not come as a surprise as $n_0 = 17,000$ which is fairly large.

Exercise 4.4 [YS exercise 3.2.] Suppose that Y, the number of heads in n tosses of a coin, is binomially distributed with index n and with parameter θ and that the prior distribution on Θ is $Be(\alpha, \beta)$.

1. Find the posterior of Θ.
2. Suppose that a-priori to any tossing the coin seems to be fair, so that we would take $\alpha = \beta$. Suppose also that tossing yields 1 tail and $n - 1$ heads. How large should n be in order that we would just give odds of 2 to 1 in favour of a head occurring at the next toss? Show that for $\alpha = \beta = 1$ we obtain $n = 4$.

Exercise 4.5 Suppose $X = (X_1, \dots, X_n)$ is a random sample from the $Unif\,(0, \theta)$-distribution.

1. The Pareto family of distributions, with parameters a and b, prescribes density $f_\Theta(\theta) = ab^a/\theta^{1+a}\mathbf{1}_{(b,\infty)}(\theta)$ $(a, b > 0)$. Derive the posterior.
2. Suppose b is small, verify that the maximum likelihood and posterior mode are about the same. Note that the statistical model for computing the MLE is different, namely $X_1, \dots, X_n \overset{\text{iid}}{\sim} Unif\,(0, \theta)$.
3. Compute the predictive distribution of a (new) future observation.

   Hint: Note that the future observation is independent of the previous observations, if we condition on θ.
4. Verify that if b is small then asymptotically (n large) the predictive density is about uniform on $[0, x_{(n)}]$ (which would be the classical approach, where an estimator (in the case the MLE), is plugged in).

Note that the Julia script `pareto_example.jl` gives a Monte-Carlo scheme to approximate the posterior density.

Exercise 4.6 [YS exercise 3.12.] Assume $\theta \in \mathbb{R}$ and that the posterior density $\theta \mapsto f_{\Theta|X}(\theta \mid x)$ is unimodal. Show that if we choose $\theta_1 < \theta_2$ to minimise $\theta_2 - \theta_1$ subject to

$$\int_{\theta_1}^{\theta_2} f_{\Theta|X}(\theta \mid x)\,\mathrm{d}\theta = 1 - \alpha$$

for given $\alpha \in (0, 1)$, then we have $f_{\Theta|X}(\theta_1 \mid x) = f_{\Theta|X}(\theta_2 \mid x)$.
Hint: In order to solve this minimisation problem, apply the method of Lagrange multipliers.

4.2 An application

4.2.1 Bayesian updating for linear regression

In the following we use "Bayesian notation" throughout. Furthermore, we use the multivariate normal distribution, see Section 1.3 for its definition.

Suppose we have observations $y_1, \dots, y_n$ satisfying a linear regression model

$$y_i = \theta_1 + \theta_2 t_i + \varepsilon_i \qquad \varepsilon_i \overset{\text{ind}}{\sim} N\left(0, \sigma^2\right).$$

The times $t_1 < t_2 < \dots$ are the observation times. We assume for simplicity that σ^2 is known. If we define

$$H_i = \begin{bmatrix} 1 & t_i \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix},$$

then we can can write

$$y_i \sim N\left(H_i\theta, \sigma^2\right).$$

Define $y = \begin{bmatrix} y_1 & \dots & y_n \end{bmatrix}'$. The likelihood is given by

$$L(\theta; y) = p(y \mid \theta) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(y_i - H_i\theta)^2\right).$$

That is,

$$p(y \mid \theta) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2}(y - H\theta)'(\sigma^2 I_n)^{-1}(y - H\theta)\right),$$

where

$$H = \begin{bmatrix} H_1 \\ \vdots \\ H_n \end{bmatrix} = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}.$$

Clearly, $y \mid \theta \sim N_n(H\theta, \sigma^2 I_n)$. We take $\theta \sim N_2(m_0, P_0)$ a priori.

Exercise 4.7 Show that $\theta \mid y \sim N_2(\nu, C)$, where

$$C^{-1} = H'\sigma^{-2}H + P_0^{-1}$$

and

$$\nu = C\left((H'\sigma^{-2}y + P_0^{-1}m_0\right).$$

That is, both the prior and posterior distribution are normal. Put differently, the chosen prior is conjugate for the given statistical model.
Hint: you may use the results in Example 4.4 here.

Bayesian updating refers to the following observation: if we let $y_{1:k} = \begin{bmatrix} y_1 & \cdots & y_k \end{bmatrix}'$ then

$$\begin{aligned} p(\theta \mid y_{1:k}) &\propto p(y_{1:k} \mid \theta)p(\theta) \\ &= p(y_{1:k-1} \mid \theta)p(y_k \mid \theta)p(\theta) \\ &\propto p(y_k \mid \theta)p(\theta \mid y_{1:k-1}). \end{aligned}$$

The equality on the second line follows from $y_{1:k-1}$ and y_k being independent, conditional on θ. Therefore, if we wish to find the posterior after k observations, we can obtain it by considering only the k-th observation coming in with prior distribution for θ equal to the posterior of θ based on the first $k-1$ observations.

Exercise 4.8 Suppose that $\theta \mid y_{1:k} \sim N\left(m_k, P_k\right)$. Show that

$$P_k^{-1} = H_k'\sigma^{-2}H_k + P_{k-1}^{-1}$$

and

$$m_k = P_k\left(H_k'\sigma^{-2}y_k + P_{k-1}^{-1}m_{k-1}\right).$$

Exercise 4.9 Use the Woodbury matrix identity (https://en.wikipedia.org/wiki/Woodbury_matrix_identity) to show that

$$P_k = P_{k-1} - P_{k-1}H_k'\left(H_kP_{k-1}H_k' + \sigma^2\right)^{-1}H_kP_{k-1}.$$

4.2.2 State-space models

There is nothing special about the chosen form of H_k, the updating formulas of exercise 4.8 holds generally under the assumption that $y_1, \dots, y_n$ are independent (conditional on θ) with $y_k \mid \theta \sim N\left(H_k\theta, \sigma^2\right)$. Now let's assume the parameter θ is not constant, but in fact a signal that evolves over time. Say we have

$$\theta_k = A\theta_{k-1} + q_{k-1} \qquad q_{k-1} \sim N(0, Q).$$

So in total we have the model

$$\begin{aligned} y_k &= H_k\theta_k + \varepsilon_k && \text{observation model} \\ \theta_k &= A\theta_{k-1} + q_{k-1} && \text{signal} \end{aligned}$$

This is an example of a linear **state-space model**. We could for instance have that $\theta_k \in \mathbb{R}^2$ and $H_k = \begin{bmatrix} 1 & 0 \end{bmatrix}$. This corresponds to only observing the first component of the signal with noise. We aim to sample from $\theta_k \mid y_{1:k}$. This is known as the **filtering problem**. If we can do this, then we are able to reconstruct/estimate not only the first component of the signal, but the second component as well!

Suppose for simplicity that $\{H_k\}_{k=1}^n$, A, Q and σ^2 are known. At time 0, before any observation has been obtained, we assume $\theta_0 \sim N\left(m_0, P_0\right)$ (just as in the previous section; this is the prior). The **Kalman** filter gives the formulas for updating $\theta_{k-1} \mid y_{1:k-1}$ to $\theta_k \mid y_{1:k}$. It consists of two steps:

1. The **prediction step**. We have
$$\theta_k \mid y_{1:k-1} \sim N\left(m_k^-, P_k^-\right)$$
with
$$\begin{aligned} m_k^- &= Am_{k-1} \\ P_k^- &= AP_{k-1}A' + Q \end{aligned} \tag{4.10}$$

2. The **update step**. Here, we use $p(\theta_k \mid y_{1:k-1})$ as a prior for the incoming observation $y_k \sim N\left(H_k\theta_k, \sigma^2 I\right)$. By exercises 4.8 and 4.9 we find that $\theta_k \mid y_{1:k} \sim N\left(m_k, P_k\right)$ with
$$P_k = P_k^- - P_k^- H_k' \left(H_k P_k^- H_k' + \sigma^2\right)^{-1} H_k P_k^-.$$
and
$$m_k = P_k^- \left(H_k'\sigma^{-2} y_k + \left(P_k^-\right)^{-1} m_k^-\right).$$

Exercise 4.10 Verify the formulas in (4.10) for the prediction step of the Kalman filter. *Hints:*

1. Note that the distribution of $\theta_k \mid y_{1:k-1}$ can be obtained as the marginal distribution of $(\theta_k, \theta_{k-1}) \mid y_{1:k-1}$.
2. Explain why
$$p(\theta_k, \theta_{k-1} \mid y_{1:k-1}) = p(\theta_k \mid \theta_{k-1}) p(\theta_{k-1} \mid y_{1:k-1}).$$
3. Apply lemma 1.14 to deduce that the joint distribution of $(\theta_k, \theta_{k-1}) \mid y_{1:k-1}$ is multivariate normal with the given parameters.

The Kalman filter is implemented in all main engineering packages such as Matlab, Python, R and Julia. Normality and linearity in its description lead to the closed-form expressions of the filter. In case of nonlinearity there are many related algorithms such as the extended and unscented Kalman filter.

Remark 4.10. Many familiar time-series models such as autoregressive and moving average processes can be written as a state-space model.

4.3 Justifications for Bayesian inference

People like Bayesian methods for different reasons, or dislike the methods at all. From a philosophical point of view, certain axioms of rational behaviour imply that the Bayesian approach is the only sensible way for doing inference. See for instance Bernardo and Smith [1994] and Jaynes [2003]. In the initial chapters of Jaynes [2003] the author thinks about how to devise a robot that acts *logically under uncertainty* and reaches the conclusion that it necessarily updates probabilities under incoming information as in Bayesian statistics. The idea of making an "inferential machine" (robot) is very popular these days and receives a lot of attention in research in statistics and artificial intelligence (machine learning). Probabilistic programming languages are a key example of this. [2] The Bayesian procedure is very strict in the sense that once prior/likelihood are fixed, all inference depends on the posterior. Exactly this property makes it suitable for automatisation on a computer.

Decision theorists relate to complete class theorems, which roughly assert that only Bayesian procedures are admissible (Cf. chapter 6). To others, the notion of exchangeability is appealing because it asserts the existence of a prior distribution on the parameter (Cf. section 4.3.1).

4.3.1 Exchangeability

Suppose that the random variables $X_1, \dots, X_n$ represent the results of successive tosses of a coin, with values 1 and 0 corresponding to the results "Heads" and "Tails" respectively. Analysing the meaning of the usual frequentist model under which the $\{X_i\}_{i=1}^n$ are independent and identically distributed with $\theta = \mathbb{P}(X_i = 1)$ fixed, the condition of independence would imply, for example, that

$$\mathbb{P}(X_n = x_n \mid X_1 = x_1, \dots, X_{n-1} = x_{n-1}) = \mathbb{P}(X_n = x_n)$$

and, therefore, the results of the first $n-1$ tosses would not change my uncertainty about the result of the n-th toss. The classical statistician would naturally react that this is true, but we do learn about the unknown parameter θ. Yet, the independence assumption seems to be unnatural (author's opinion!) and indeed, Bayesians often motivate their models by the weaker notion of exchangeability. Regarding the example just given, Schervish [1995] (page 7) remarks that

> It seems unfortunate that so much machinery as assumptions of mutual independence and the existence of a mysterious fixed but unknown θ must be introduced to describe what seems, on the surface, to be a relatively simple situation.

Definition 4.11. A finite set $X_1, \dots, X_n$ of random quantities is said to be **exchangeable** if every permutation of $(X_1, \dots, X_n)$ has the same joint distribution as every other permutation. An infinite collection is exchangeable if every finite subcollection is exchangeable.

The motivation for the definition of exchangeability is to express symmetry of beliefs about the random quantities in the weakest possible way. The definition merely says that the labeling of the random quantities is immaterial. If $X_1, \dots, X_n$ are IID (Independent and Identically Distributed), then $X_1, \dots, X_n$ are exchangeable:

$$\begin{aligned} f_{X_1,\dots,X_n}(x_1, \dots, x_n) &= \int f_{X_1,\dots,X_n|\Theta}(x_1, \dots, x_n \mid \theta)\mu_\Theta(\mathrm{d}\theta) \\ &= \int \prod_{i=1}^n f_{X_i|\Theta}(x_i \mid \theta)\mu_\Theta(\mathrm{d}\theta) \end{aligned}$$

The converse is not true and hence assuming exchangeability is a weaker requirement than IID.

[2] I believe in this sense Jaynes was way ahead in his thinking.

Example 4.12. Suppose that, for each n, the joint density of $X_1, \dots, X_n$ is

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = \frac{\Gamma(a+n)b^a}{\Gamma(a)\left(b + \sum_{i=1}^n x_i\right)^{a+n}} \qquad \text{all } x_i > 0.$$

For each n, the random variables $X_1, \dots, X_n$ are exchangeable. The density of $(X_1, \dots, X_n)$ follows upon integrating out Θ from the joint density of $(X_1, \dots, X_n, \Theta)$, when

$$X_1, \dots, X_n \mid \Theta = \theta \overset{\text{iid}}{\sim} Exp(\theta) \qquad \Theta \sim Ga(a, b).$$

The connection between exchangeability and Bayesian statistics is captured by a famous representation theorem by De Finetti. We first give the result in case of Bernoulli random variables.

Theorem 4.13 (DeFinetti's representation theorem for Bernoulli random variables)**.** An *infinite* sequence $\{X_n\}_{n=1}^\infty$ of Bernoulli random variables is exchangeable if and only if there is a random variable Θ taking value in $[0, 1]$ such that, conditional on $\Theta = \theta$, $\{X_n\}_{n=1}^\infty$ are IID $Ber(\theta)$. Furthermore, if the sequence is exchangeable, then the distribution of Θ is unique and $\sum_{i=1}^n X_i/n$ converges to Θ almost surely.

Hence, assuming exchangeability implies existence of a random variable Θ such that the X_i are IID conditional on Θ. This connects well with the Bayesian paradigm, in which all unknowns are equipped with a probability distribution.

The general version is somewhat complicated but conveys the same message.

Theorem 4.14 (DeFinetti's representation theorem*)**.** Let $(S, \mathcal{F}, \mu)$ be a probability space, and let $(\mathcal{X}, \mathcal{B})$ be a Borel space. For each n, let $X_n : S \to \mathcal{X}$ be measurable. The sequence $\{X_n\}_{n=1}^\infty$ is exchangeable if and only if there is a random probability measure $\mathbf{P}$ on $(\mathcal{X}, \mathcal{B})$ such that, conditional on $\mathbf{P} = P$, $\{X_n\}_{n=1}^\infty$ are IID with distribution P. Furthermore, if the sequence is exchangeable, then the distribution of $\mathbf{P}$ is unique, and $\mathbf{P}_n(B)$ converges to $\mathbf{P}$ almost surely for each $B \in \mathcal{B}$.

A lengthy discussion and proofs of these result can be found in sections 1.4 and 1.5 of Schervish [1995]. De Finetti's theorem shows that if we consider an *infinite* collection of random variables then these are exchangeable if and only if the collection is conditionally IID.

Example 4.15. Exchangeability and independence are just 2 different relations between random variables. Independence implies exchangeability; the reverse implication is false. A simple example illustrates this: consider a box with 3 balls, labelled A, B and C. Let X be the label of the first ball drawn from the box, and Y the label of the second ball drawn from the box. We consider drawing without replacement, so in the end one ball remains in the box. The joint distribution of (X, Y) is given by

	A	B	C
A	0	1/6	1/6
B	1/6	0	1/6
C	1/6	1/6	0

Clearly X and Y are exchangeable, but they are not independent. To see this: $\mathbb{P}(Y = A \mid X = B) = 1/2$ while $\mathbb{P}(Y = A) = 1/3$.

In case of drawing *with* replacement, X and Y would be independent, and in that case the table would be filled with 9 times the number $1/9$. This would then give $\mathbb{P}(Y = A \mid X = B) = \mathbb{P}(Y = A) = 1/3$.

Exercise 4.11 If $X_1, \dots, X_n$ are IID conditional on Θ, prove that for $i \neq j$

$$\text{Cov}(X_i, X_j) = \text{Var}\, \mu(\Theta) \geq 0,$$

where $\mu(\Theta) = \mathbb{E}[X_1 \mid \Theta]$. *Hint*: $\mathbb{E}X_1 = \mathbb{E}\mathbb{E}[X_1 \mid \Theta]$.

4.4 Choosing the prior

Besides the choice of the statistical model, the Bayesian approach requires additionally the specification of the prior. This is often seen by classical statisticians as a weak point of the Bayesian formalism, claiming that by appropriate choice of the prior any conclusion can be derived, making the inference subjective. Answers to this critique from the Bayesian side include:

1. Frequentist procedures are neither objective, for example the chosen significance level is often arbitrary. Moreover, the choice of statistical model is usually not objective as well and much more influential (and this influence remains even in the large sample limit). Often, for the same problem, multiple tests exist, none of which can be classified as optimal. For a pre specified significance level, different tests may give rise to conflicting conclusions. Choosing among the tests appears subjective.

2. The influence of the prior distribution usually vanishes as the sample size increases (see section 4.7).

3. In certain cases one can choose a prior that has least influence on the resulting posterior. Deriving such a prior, relative to the statistical model, is key to what is called **objective Bayesian** statistics.

4. Under certain assumptions, any "admissible" statistical procedure is essentially Bayesian (we make this precise in chapter 6).

5. "Optimal" frequentist procedures are Bayesian procedures with respect to a particular prior.

We go more into detail on the final two points of this list in chapter 6. It is fair to say that choice of prior can be very influential in small samples giving the statistician the possibility to take a prior that biases the posterior in a favourable direction. Think about a pharmaceutical company choosing the prior on efficacy of a medicament itself! The point is that there do exist reasonable priors on which consensus can be reached. However, it is very hard to define in general what "reasonable" means here. Hence, in Bayesian statistics, at some stage the statistician has to trust his/her prior and investigate in which sense this choice affects the posterior. If its influence is strong, a common approach is to make the prior more robust by spreading out its mass. Note that in frequentist statistics there is no prior, but test/confidence-intervals are often based on asymptotic derivations and then, rather on trusting a prior, the statistician has to trust that the sample size is sufficiently large to justify asymptotic approximations.

Conjugate priors constitute a convenient class of priors.

Definition 4.16. If $f_\Theta \in \mathcal{G}$ implies $f_{\Theta|X} \in \mathcal{G}$ for some class of densities $\mathcal{G}$ then we call the class $\mathcal{G}$ **conjugate** for the statistical model.

Note that conjugacy is a property of the model together with the family of priors considered. Conjugate priors are very convenient and often lead to a relatively simple formula for the posterior density. We give a number of examples.

Example 4.17. $X_i \mid \Theta = \theta \sim N(\theta, 1)$ with $\Theta \sim N(a, \tau^2)$.

Example 4.18. $X_i \mid \Theta = \theta \sim Ga(\alpha, \theta)$ (α known) with $\Theta \sim Ga(a, b)$.

Example 4.19. $X_i \mid \Theta = \theta \sim Bin(n, \theta)$ with $\Theta \sim Beta(a, b)$.

Exercise 4.12 Verify conjugacy in these examples.

Exercise 4.13 [YS exercise 5.3.] Find the general form of a conjugate prior density for θ in a Bayesian analysis of the one-parameter exponential family density

$$f(x;\theta) = c(\theta)h(x)\exp(\theta t(x)), \quad x \in \mathbb{R}.$$

Another way for obtaining a prior is based on subjective grounds of the experimenter or historical data. In the former case one speaks of **subjective Bayesian procedures**. On the opposite, the **objective Bayes** school tries to choose the prior that affects the resulting posterior in a minimal way. A first thing that comes to mind for objective choice of the prior is to take the uniform prior over the parameter set (assuming it is bounded for simplicity). This won't work, as can be seen from the following example.

Example 4.20. Suppose $X \mid \Theta = \theta \sim Ber(\theta)$ and take the uniform prior over the parameter set: $\Theta \sim Unif(0,1)$. The prior is supposed to express ignorance about θ. However, if we are ignorant about θ, we are ignorant about θ^2 as well (for example). If we define $\Psi = \Theta^2$, then

$$f_\Psi(\psi) = f_\Theta(\sqrt{\psi})\frac{1}{2\sqrt{\psi}} = \frac{1}{2\sqrt{\psi}}\mathbf{1}_{[0,1]}(\psi).$$

This implies that small realisations of Ψ are more probable than large realisations and we are no longer ignorant about Ψ.

We can set this up more formally, following an example presented in Kleijn [2020]. Suppose we have a statistical model that is parametrised in two different ways. Say we have

$$\phi_1 : (0,1) \to \mathcal{P} \quad \text{with} \quad \phi_1(\tau) = N(0,\tau)$$

and

$$\phi_2 : (0,1) \to \mathcal{P} \quad \text{with} \quad \phi_2(\sigma) = N\left(0,\sigma^2\right).$$

From a statistical modelling perspective, these two models are equivalent (in both cases we have a Normal distribution, where the variance is assumed to be in $(0,1)$). More generally, assume

$$\phi_1 : \Omega_1 \to \mathcal{P} \quad \text{and} \quad \phi_2 : \Omega_2 \to \mathcal{P}.$$

For the following derivation, we assume each Ω_i is a measurable space with σ-algebra $\mathcal{B}_i$ and $\mathcal{P}$ is a measurable space with σ-algebra $\mathcal{E}$.[3] Assume in addition that $\phi_1, \phi_1^{-1}, \phi_2, \phi_2^{-1}$ are measurable. Assuming Ω_1 to be bounded, we can define

$$\Pi_1(A) = \mu(A)/\mu(\Omega_1) \qquad A \in \mathcal{B}_1,$$

μ denoting Lebesgue measure. Hence Π_1 is the rescaled Lebesgue measure on Ω_1, representing the uniform distribution on Ω_1. This measure induces a measure on $\mathcal{P}$ by

$$\Pi_1'(B) = (\Pi_1 \circ \phi_1^{-1})(B) \qquad B \in \mathcal{E}.$$

But then, using ϕ_2 we can push-back to Ω_2 and obtain the measure [4]

$$\Pi_2(C) = (\Pi_1' \circ \phi_2)(C) = \left(\Pi_1 \circ (\phi_1^{-1} \circ \phi_2)\right)(C) \qquad C \in \mathcal{B}_2.$$

Hence, starting with a uniform prior on Ω_1, we obtain a prior on Ω_2 that in general will not be uniform. Let's check this for our example:

$$\tau(\sigma) := (\phi_1^{-1} \circ \phi_2)(\sigma) = \phi_1^{-1}(N\left(0,\sigma^2\right)) = \sigma^2.$$

[3] Formally, we should state precisely the topology and $\mathcal{P}$ and $\mathcal{E}$; this is somewhat out of scope for this course.

[4] If you are familiar with differential geometry, you may recognise a resemblance with moving along overlapping charts.

Hence, as $\Pi_2(C) = (\Pi_1 \circ \tau)(C)$ we get that the induced density on Ω_2 is given by

$$\pi_2(\sigma) = \pi_1(\tau(\sigma)) \left| \frac{\mathrm{d}\tau}{\mathrm{d}\sigma} \right| = 1 \cdot 2\sigma = 2\sigma. \tag{4.11}$$

So, taking parametrisation ϕ_2, the prior is non uniform and favours values close to 1 (if for example $\Omega_1 = (0, 1)$).

A natural question that appears is whether it is possible to choose a prior such that the prior is invariant under reparametrisation. So what is meant by this? It means that we have a way of constructing the prior, such that the following two derivations lead to the same result:

1. Construct the prior using parametrisation ϕ_1 (in terms of τ) and transfer it to parametrisation ϕ_2 (in terms of σ) using $\Pi_2(\cdot) = \left(\Pi_1 \circ (\phi_1^{-1} \circ \phi_2)\right)(\cdot)$.
2. Construct the prior using parametrisation ϕ_2 (in terms of σ).

In certain cases this is possible and its construction dates back to 1946 and is due to Jeffreys. Often, it leads to "improper priors", a concept that we will now discuss.

4.4.1 Improper priors

Definition 4.21. A prior distribution with "density" f_Θ satisfying

$$\int f_\Theta(\theta)\, \mathrm{d}\theta = \infty$$

is called an **improper prior distribution**.

Improper priors do not follow the rules of probability theory and should be interpreted as expressing *degree of belief*. They can only be used as long as the posterior is proper. The idea is that since all Bayesian inference depends only on the posterior, the prior actually being a density is not really important. As you can guess, not all statisticians agree on using improper priors! Moreover, it is really not guaranteed that the posterior is proper. As an example, suppose $X \mid \Theta = \theta \sim Bin(n, \theta)$ and we use the improper prior

$$f_\Theta(\theta) \propto \frac{1}{\theta(1-\theta)}$$

(this prior is known as Haldane's prior). It follows that

$$f_{\Theta|X}(\theta \mid x) \propto \frac{1}{\theta(1-\theta)} \theta^x (1-\theta)^{n-x} = \theta^{x-1}(1-\theta)^{n-x-1}.$$

Now suppose $x = 0$ is observed. Then due to the fact that $\theta \mapsto 1/\theta$ is not integrable near zero, the map $\theta \mapsto f_{\Theta|X}(\theta \mid 0)$ fails to be integrable. Hence there is no well-defined posterior in this case.

The following example shows why improper priors can be a natural choice sometimes.

Example 4.22. Let g be a density (with respect to Lebesgue measure) on $\mathbb{R}$ and assume the location model

$$f_{X|\Theta}(x \mid \theta) = g(x - \theta)$$

and $\theta \in \mathbb{R}$ is the location parameter. If no prior information is available, it makes sense to assume the likelihood of an interval $[a, b]$ is proportional to $b - a$. Therefore, the prior is proportional to Lebesgue measure on $\mathbb{R}$. We write $f_\Theta(\theta) \propto k$ ($k \in \mathbb{R}$), or $f_\Theta(\theta) \propto 1$.

Exercise 4.14 Derive the posterior in case g is the density of the standard normal distribution.

The importance of improper priors is due to

1. their appearance as noninformative (objective) priors.
2. their ability to recover usual estimators like maximum likelihood within the Bayesian paradigm.

Often, improper priors arise as limits of a sequence of proper prior distributions. They are generally more acceptable to non-Bayesians, as they can lead to estimators with frequentist validation, such as minimaxity (related to improper least favourable priors, Cf. chapter 6).

4.4.2 Jeffreys' prior

In 1946, Sir Harold Jeffreys described an ingenious method for constructing an "objective" prior distribution. As we will see, the prior is constructed from the likelihood and this is at odds with the Bayesian paradigm (and the likelihood principle). As a consequence (and warning): subjective Bayesians refute this construction. Jeffreys' method is worthwhile studying for at least two reasons:

1. historically, it was the first approach to deal with the problem of Example 4.20 by constructing a prior that is "invariant under reparametrisation";
2. many frequentist estimators are recovered by using Jeffreys' prior.

Definition 4.23. Assume FI-regularity conditions. The Jeffreys prior is defined by

$$\mathrm{d}\Pi(\theta) = \sqrt{\det I(\theta)}\,\mathrm{d}\theta.$$

If $\Pi(\Omega) < \infty$, then Jeffreys prior is usually normalised to a probability measure.

In case the parameter is one-dimensional, it is not hard to see that this prior satisfies the required invariance. From Equation (4.11) it follows that the invariance is satisfied by a prior with density π when

$$\pi(\sigma) = \pi(\tau(\sigma)) \left| \frac{\mathrm{d}\tau}{\mathrm{d}\sigma} \right|.$$

By Lemma 2.17 Jeffreys' prior satisfies this equation.

Example 4.24. We return to the example where

$$\phi_1 : (0,1) \to \mathcal{P} \quad \text{with} \quad \phi_1(\tau) = N(0,\tau)$$

and

$$\phi_2 : (0,1) \to \mathcal{P} \quad \text{with} \quad \phi_2(\sigma) = N\left(0,\sigma^2\right).$$

It is easily verified that indeed Jeffreys' prior transforms correctly. First, it is elementary to verify that

$$I_1(\tau) = \frac{1}{2\tau^2} \qquad \text{and} \qquad I_2(\sigma) = \frac{2}{\sigma^2}.$$

Now, starting from the parametrisation in terms of τ, Jeffreys' prior is given by

$$\pi_1(\tau) = \frac{1}{\tau\sqrt{2}}.$$

This prior can be transported to σ which gives density

$$\pi_1''(\sigma) = \pi_1(\tau(\sigma))2\sigma = \pi_1(\sigma^2)2\sigma = \frac{2\sigma}{\sigma^2\sqrt{2}} = \frac{\sqrt{2}}{\sigma}.$$

This is exactly equal to the prior we get when we start of from $I_2(\sigma)$, which would give

$$\pi_2(\sigma) = \sqrt{I_2(\sigma)} = \frac{\sqrt{2}}{\sigma}.$$

In many complicated models, it is difficult to compute the Fisher information and obtain an expression for the Jeffreys' prior. Moreover, in case the Fisher-information does not exist a clear problem arises. So this prior is definitely not "the solution" in general to the problem of selecting a prior in an objective way. Even saying it is objective is a bit vague: but it does resolve the reparametrisation issue for prior selection in some statistical models.

For integrable, non-negative k we have (using Jeffreys' prior)

$$\int k(\psi) f_\Psi(\psi)\,\mathrm{d}\psi = \int k(\psi) f_\Theta\left(g^{-1}(\psi)\right)\left|\frac{\mathrm{d}}{\mathrm{d}\psi} g^{-1}(\psi)\right|\,\mathrm{d}\psi = \int k(g(\theta)) f_\Theta(\theta)\,\mathrm{d}\theta,$$

where the second equality follows from the substitution $\theta = g^{-1}(\psi)$. This implies that for measurable subsets $A \subset \mathbb{R}$ we have

$$\frac{\int f_{X|\Psi}(x \mid \psi)\mathbf{1}_{g(A)}(\psi) f_\Psi(\psi)\,\mathrm{d}\psi}{\int f_{X|\Psi}(x \mid \psi) f_\Psi(\psi)\,\mathrm{d}\psi} = \frac{\int f_{X|\Psi}(x \mid g(\theta))\mathbf{1}_{g(A)}(g(\theta)) f_\Theta(\theta)\,\mathrm{d}\theta}{\int f_{X|\Psi}(x \mid g(\theta)) f_\Theta(\theta)\,\mathrm{d}\theta}.$$

The left-hand-side equals

$$\int_{\psi \in g(A)} f_{\Psi|X}(\psi \mid x)\,\mathrm{d}\psi = \mathbb{P}(\Psi \in g(A) \mid X),$$

while the right-hand-side equals

$$\int_{\theta \in A} f_{\Theta|X}(\theta \mid x)\,\mathrm{d}\theta = \mathbb{P}(\Theta \in A \mid X).$$

Hence Jeffreys' prior ensures that

$$\boxed{\mathbb{P}(\Psi \in g(A) \mid X) = \mathbb{P}(\Theta \in A \mid X)}.$$

Example 4.25 (Jeffreys' prior for location families)**.** Suppose Y has density f_Y. Let $b \in \mathbb{R}$. If $X = Y + \theta$, then $f_X(x \mid \theta) = f_Y(x - \theta)$. The family of densities $\{f_X(\cdot \mid \theta), \theta \in \mathbb{R}\}$ is called a **location-family** of densities. By example 2.13

$$I(\theta) = \int \frac{f_Y'(u)^2}{f_Y(u)}\,\mathrm{d}u.$$

it follows that Jeffreys' prior is given by $f_\Theta(\theta) \propto 1$, the (improper) uniform prior on $\mathbb{R}$.

Example 4.26. Suppose $X_1, \ldots, X_n \mid \Theta = \theta \overset{\text{ind}}{\sim} N(\theta, 1)$. Let $X = (X_1, \ldots, X_n)$. This is a location family and using Jeffreys' prior we get

$$f_{\Theta|X}(\theta \mid X) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^n (X_i - \theta)^2\right).$$

If $X = (X_1, \ldots, X_n)$, then the posterior satisfies $\Theta \mid X \sim N\left(\bar{X}_n, 1/n\right)$. The posterior mean equals $\mathrm{E}[\Theta \mid X] = \bar{X}_n$ and a 95%-credible set is given by

$$[\bar{X}_n - 1.96/\sqrt{n}, \bar{X}_n + 1.96/\sqrt{n}].$$

The classical 95%-confidence interval for θ is the same, but the interpretation of both intervals is completely different!

Example 4.27 (Jeffreys' prior for scale families)**.** Suppose Y has density f_Y. Let $\theta > 0$. If $X = \theta Y$, then $f_X(x \mid \theta) = \frac{1}{\theta} f_Y\left(\frac{x}{\theta}\right)$. Cf. example 1.17. By example 2.13

$$I(\theta) = \frac{1}{\theta^2} \int_0^\infty \left(1 + \frac{u f_Y'(u)}{f_Y(u)}\right)^2 f_Y(u)\, \mathrm{d}u.$$

it follows that Jeffreys' prior is given by $f_\Theta(\theta) \propto 1/\theta$, which is improper.

Example 4.28 (Jeffreys' prior for the Poisson distribution)**.** Suppose $X \sim Pois(\theta)$, then $I(\theta) = 1/\theta$ and Jeffreys' prior is given by

$$f_\Theta(\theta) \propto 1/\sqrt{\theta}$$

which is improper. The posterior is nonetheless proper: from

$$f_{\Theta|X}(\theta \mid x) \propto e^{-\theta} \frac{\theta^x}{x!} \frac{1}{\sqrt{\theta}} \propto e^{-\theta} \theta^{x-1/2}.$$

it follows that $\Theta \mid X \sim Ga(X + 1/2, 1)$.

Example 4.29 (Jeffreys' prior for the binomial and negative binomial distribution)**.** Suppose $X \sim Bin(n, \theta)$, then $I(\theta) = n/(\theta(1-\theta))$ and

$$f_\Theta(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}}.$$

Hence, $\Theta \sim Be(1/2, 1/2)$. In this experiment, the number of trials is fixed.

Suppose $X \sim NegBin(r, \theta)$, then $I(\theta) = r/(\theta^2(1-\theta))$ and

$$f_\Theta(\theta) \propto \frac{1}{\theta\sqrt{1-\theta}}$$

which is improper. In this experiment, the number of trials is random.

Suppose the statistician gets reported 3 successes and 12 failures among 15 independent Bernoulli trials. What is the posterior mean for θ using Jeffreys' prior?

- If we planned to do 15 trials and got 3 successes, the experiment is Binomial with $n = 15$. The likelihood is proportional to $\theta^3(1-\theta)^{12}$ and the posterior is proportional to

  $$\theta^{2.5}(1-\theta)^{11.5}$$

  This corresponds to the $Be(3.5, 12.5)$ distribution which has mean $3.5/16 \approx 0.22$.

- If we continued sampling until we had 12 failures, the experiment is Negative-Binomial with $r = 12$. The likelihood is again proportional to $\theta^3(1-\theta)^{12}$ and the posterior is proportional to

  $$\theta^2(1-\theta)^{11.5}$$

  This corresponds to the $Be(3, 12.5)$ distribution which has mean $3/15.5 \approx 0.19$.

This example implies that Jeffreys' prior does not satisfy the **likelihood principle** (Cf. chapter 3). For this reason, Jeffreys' priors are not universally accepted by Bayesians. A further complication is that it is not always easy to calculate the Jeffreys' prior for a given statistical model. As a final note, especially in multivariate cases, Jeffreys' prior can perform unsatisfactory. The interested reader can consult for example Berger et al. [2015]. The suggested fix is to use **reference priors**. We do not go into details here.

Exercise 4.15 * [YS exercise 3.9.] Let $X_1, \dots, X_n \mid \mu, \sigma \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Let $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ and $S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. Assume

$$p(\mu, \sigma) \propto \sigma^{-1} \mathbf{1}_{\mathbb{R} \times (0,\infty)}((\mu, \sigma)).$$

Show that the marginal posterior of $\sqrt{n}(\mu - \bar{X})/S \sim t_{n-1}$ (the t distribution with $n-1$ degrees of freedom) and find the marginal posterior of σ.

Exercise 4.16 Suppose $X \sim Ber(\theta)$. Use Jeffreys' prior to compute the posterior probability that $\theta \in (0, 1/2)$. Next assume $X \sim Ber(\sqrt{\eta})$. Use Jeffreys' prior to compute the posterior probability that $\theta \in (0, 1/4)$. Compare the results and explain.

Exercise 4.17 Show that if we parametrise the Poisson distribution instead of θ by $\sqrt{\theta}$, the Jeffreys prior is uniform. Construct a general transformation $\theta \mapsto g(\theta)$ such that the Jeffreys prior is uniform (possibly improper). You may assume that θ is one-dimensional.
Hint: Use Lemma 2.17 (on Fisher information under reparametrisation) to construct a general transformation such that Jeffreys' prior is uniform.

Exercise 4.18 Suppose we have data $\{X_i, i = 1, \dots, n\}$ and assume the following hierarchical model.

$$X_i \mid \Theta_i = \theta_i \overset{\text{ind}}{\sim} N\left(\theta_i, 1\right)$$
$$\Theta_1, \dots, \Theta_n \mid T = \tau \overset{\text{iid}}{\sim} N\left(0, \tau^2\right)$$
$$f_T(\tau) \propto \frac{1}{\tau}$$

The prior on τ is improper and motivated as a Jeffreys' prior (note that the $N\left(0, \tau^2\right)$ belongs to a scale family). Investigate whether the posterior for τ is a proper density.

4.5 Hierarchical Bayesian models

Hierarchical Bayesian models are no different from ordinary Bayesian models, but owe their name to the hierarchical way in which the prior is built. Hierarchical modelling is a fundamental concept in Bayesian statistics. Their role is well summarised in the opening section of Teh and Jordan [2010].

> Hierarchical modeling is a fundamental concept in Bayesian statistics. The basic idea is that parameters are endowed with distributions which may themselves introduce new parameters, and this construction recurses. A common motif in hierarchical modeling is that of the conditionally

> independent hierarchy, in which a set of parameters are coupled by making their distributions depend on a shared underlying parameter. These distributions are often taken to be identical, based on an assertion of exchangeability and an appeal to de Finetti's theorem. Hierarchies help to unify statistics, providing a Bayesian interpretation of frequentist concepts such as shrinkage and random effects. Hierarchies also provide ways to specify non-standard distributional forms, obtained as integrals over underlying parameters. They play a role in computational practice in the guise of variable augmentation. These advantages are well appreciated in the world of parametric modeling, and few Bayesian parametric modelers fail to make use of some aspect of hierarchical modeling in their work.

In the following definition we use Bayesian notation.

Definition 4.30. A **hierarchical Bayesian model** is a Bayesian statistical model, $(p(x \mid \theta), p(\theta))$, where the prior distribution $p(\theta)$ is decomposed into

$$p(\theta \mid \theta_1),\ p(\theta_1 \mid \theta_2), \ldots, p(\theta_{n-1} \mid \theta_n),\ p(\theta_n)$$

such that

$$p(\theta) = \int p(\theta \mid \theta_1)\, p(\theta_1 \mid \theta_2) \cdots p(\theta_{n-1} \mid \theta_n) p(\theta_n)\, \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_n.$$

The parameters θ_i are called the hyperparameters of level i $(1 \leq i \leq n)$.

The popularity of such models is partly due to their flexibility in modelling complex dependencies in the data, but also due to the existence of computational algorithms to draw from the posterior (most notably Markov Chain Monte Carlo methods, see chapter 5).

Example 4.31. Suppose in a medical treatment there are I treatments groups. Denote by X_{ij} the response of subject j in treatment group i. Within each group, we model the data as exchangeable:

$$X_{ij} \mid \Theta_i = \theta_i \overset{\text{ind}}{\sim} N\left(\theta_i, 1\right)$$
$$\Theta_1, \ldots, \Theta_I \overset{\text{ind}}{\sim} N\left(\nu, \tau^2\right)$$

for known values of ν and τ^2. The hyperparameters are ν and τ^2 in this case. Additional layers in the hierarchical model can be introduced by introducing priors on ν and/or τ^2 as well. If this is pursued, dependence is created among all X_{ij}.

Example 4.32. Suppose that a survey is conducted in I cities. Each person surveyed is asked a yes-no question. Denote by X_{ij} the response of person j in city i. Set $X_{ij} = 1$ if the answer is "yes" and $X_{ij} = 0$ if it is "no". If we model the data within a city as exchangeable then a possible model is given by

$$X_{ij} \mid \Theta_i = \theta_i \overset{\text{ind}}{\sim} Ber\left(\theta_i\right)$$
$$\Theta_1, \ldots, \Theta_I \overset{\text{ind}}{\sim} Be\left(\alpha, \beta\right)$$

for hyperparameters α and β.

Example 4.33. In the baseball data example introduced in section 3.4.1 in Young and Smith [2005], we observe y_i which is the number of home-runs out of n_i times at bat $(i = 1, \ldots, n,\ n = 17)$ from pre-season data. Two models are proposed:

1. A model where the response is transformed and next modelled by a Normal distribution. Define $X_i = \sqrt{n_i}\arcsin\left(2\frac{Y_i}{n_i} - 1\right)$, then the model is given by:

$$\begin{aligned} X_i \mid M_i = \mu_i &\overset{\text{ind}}{\sim} N\left(\mu_i, 1\right) \\ M_i \mid \Theta = \theta, T = \tau &\overset{\text{iid}}{\sim} N\left(\theta, \tau^2\right) \\ f_{(\Theta,T)}(\theta,\tau) &\propto \tau^{-1-2\alpha^*} e^{-\beta^*/\tau^2} \end{aligned}$$

2. A model where the responses are Binomially distributed.

$$\begin{aligned} Y_i \mid M_i = \mu_i &\overset{\text{ind}}{\sim} Bin\left(n_i, p_i\right) \quad \text{with} \quad p_i = (1+e^{-\mu_i})^{-1} \\ M_i \mid \Theta = \theta, T = \tau &\overset{\text{iid}}{\sim} N\left(\theta, \tau^2\right) \\ f_{(\Theta,T)}(\theta,\tau) &\propto \tau^{-1-2\alpha^*} e^{-\beta^*/\tau^2} \end{aligned}$$

To avoid excessive notational overhead one sometimes writes $x_i \mid \mu_i \overset{\text{ind}}{\sim} N\left(\mu_i, 1\right)$ instead of $X_i \mid M_i = \mu_i \overset{\text{ind}}{\sim} N\left(\mu_i, 1\right)$ (for example). Then, all quantities are written down in lower-case.

4.6 Empirical Bayes

Although the name suggests this is a Bayesian method, it's not. Suppose we are given the following setup

$$\begin{aligned} X \mid \Theta = \theta &\sim f_{X|\Theta}(\cdot \mid \theta) \\ \Theta &\sim f_\Theta(\theta;\eta), \end{aligned}$$

where η is the hyperparameter. A truly Bayesian analysis requires specification of the value for η. If there is insufficient prior information on η, the idea of empirical Bayes methods is to estimate η from f_X. This can for example be done in a way similar to maximum likelihood way by defining

$$\hat{\eta} = \underset{\eta}{\operatorname{argmax}}\, f_X(x;\eta). \tag{4.12}$$

The "posterior" obtained by the empirical Bayes method is the "ordinary" posterior, with $\hat{\eta}$ substituted for η. Empirical Bayes methods are neither classical nor Bayesian. It is observed that estimators obtained in this way often are "good" in terms of classical optimality criteria, such as minimaxity (details follow in section 6.3 of chapter 6).

Example 4.34. Suppose $X_1, \ldots, X_n \mid \Theta \overset{\text{iid}}{\sim} Pois\,(\Theta)$ and $\Theta \sim Ga\,(a,b)$ apriori. The hyperparameter is given by (a,b). The marginal density of the data is given by

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;a,b) &= \int \prod_{i=1}^{n} f_{X_i}(x_i \mid \theta) f_\Theta(\theta)\,\mathrm{d}\theta \\ &= \int_0^\infty e^{-n\theta}\frac{\theta^S}{C}\frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}\,\mathrm{d}\theta \\ &= \frac{b^a}{C\Gamma(a)}\int_0^\infty \theta^{S+a-1}e^{-(b+n)\theta}\,\mathrm{d}\theta = \frac{b^a}{C\Gamma(a)}\frac{\Gamma(S+a)}{(b+n)^{S+a}} \end{aligned}$$

with $S = \sum_{i=1}^n X_i$ and $C = \prod_{i=1}^n (X_i!)$. The final equality follows upon noting that the integrand is proportional to the density of the $Ga\,(S+a, b+n)$-density. The empirical Bayes estimator for (a,b) is defined by

$$(\hat{a}^{(EB)}, \hat{b}^{(EB)}) = \underset{(a,b)\in(0,\infty)^2}{\operatorname{argmax}} \left(a \log b - \log \Gamma(a) + \log \Gamma(S+a) - (S+a)\log(b+n)\right).$$

It is not entirely clear whether such a maximiser exists and for sure it cannot be calculated in closed form. If we fix $b = 1$ for example (corresponding to a $Exp\,(b)$-prior), the problem gets easier.

Example 4.35. Suppose $X \mid \Theta = \theta \sim N\,(\theta, 1)$ and $\Theta \sim N\,(0, A)$ apriori. Note that this prior is conjugate for the given statistical model (i.e. the posterior has a normal distribution). The "hyperparameter" for this model is A. The posterior mean is given by $AX/(A+1)$. The marginal density of X is computed as follows:

$$\begin{aligned} f_X(x; A) &= \int f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)\, \mathrm{d}\theta \\ &= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} \frac{1}{\sqrt{2\pi A}} e^{-\frac{\theta^2}{2A}}\, \mathrm{d}\theta = \frac{1}{\sqrt{2\pi(1+A)}} \exp\left(-\frac{x^2}{2(1+A)}\right). \end{aligned}$$

This shows that $X \sim N\,(0, 1+A)$. The empirical Bayes method postulates that we find A by maximising $f_X(X; A)$ over $A \geq 0$. This gives

$$\hat{A}^{(EB)} = \underset{A\geq 0}{\operatorname{argmax}}\, f_X(X; A) = \max(X^2 - 1, 0).$$

Therefore the empirical Bayes estimator obtained in this way is given by

$$\frac{\hat{A}^{(EB)}}{1+\hat{A}^{(EB)}} X = \begin{cases} 0 & \text{if } X^2 \leq 1 \\ (1 - 1/X^2)X & \text{if } X^2 > 1 \end{cases}.$$

Example 4.36. Suppose we define a family of priors for a fixed $\varepsilon \in [0,1]$ by letting

$$f_\Theta^{(\varepsilon,q)}(\theta) = (1-\varepsilon)\pi(\theta) + \varepsilon q(\theta).$$

Here π is a fixed density and q can be any density. The idea is that ε is close to zero and π is the prior density the statistician has in mind. The family of densities $\{f_\Theta^{(\varepsilon,q)}\}$ is called a **contamination family**. We will choose q in an empirical Bayes fashion. Note that

$$\begin{aligned} f_X(x) &= (1-\varepsilon) \int f_{X|\Theta}(x \mid \theta)\pi(\theta)\, \mathrm{d}\theta + \varepsilon \int f_{X|\Theta}(x \mid \theta) q(\theta)\, \mathrm{d}\theta \\ &\leq (1-\varepsilon) \int f_{X|\Theta}(x \mid \theta)\pi(\theta)\, \mathrm{d}\theta + \varepsilon f_{X|\Theta}(x \mid \hat{\Theta}), \end{aligned}$$

where $\hat{\theta}$ is the maximum likelihood estimate for θ. It follows that the empirical Bayes prior is given by $f_\Theta(\cdot) = (1-\varepsilon)\pi(\cdot) + \varepsilon\delta_{\hat{\Theta}}(\cdot)$. This is a mixture of π and a Dirac-measure at $\hat{\Theta}$.

Exercise 4.19 Suppose $X_1, \ldots, X_n$ are conditionally independent given $\Theta = \theta$ with the $N(\theta, 1)$ distribution. Assume $\Theta \sim N\left(0, \tau^2\right)$ apriori. Here τ^2 is a hyperparameter.

1. Show that
$$\Theta \mid X_1, \ldots, X_n \sim N\left(\frac{\tau^2 \sum_1^n X_i}{n\tau^2 + 1}, \frac{\tau^2}{n\tau^2 + 1}\right).$$
and conclude that the posterior mean is given by
$$\frac{\tau^2}{n\tau^2 + 1} n\bar{X}_n.$$

2. Derive that $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ equals
$$\frac{1}{(\sqrt{2\pi})^n \sqrt{n\tau^2 + 1}} \exp\left(-\frac{1}{2}\left[\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n + 1/\tau^2}\right]\right).$$
Hint: first write down the joint density of $(\Theta, X_1, \ldots, X_n)$ and then integrate out θ, then first show that this integral equals
$$(2\pi)^{-\frac{n+1}{2}} \frac{1}{\tau} e^{-\frac{1}{2}\sum_{i=1}^n x_i^2} \int \exp\left(S\theta - \frac{1}{2}\frac{\theta^2}{u_n}\right) \mathrm{d}\theta,$$
where $u_n^{-1} = n + \tau^{-2}$ and $S = \sum_{i=1}^n x_i$. Then show that this integral equals $\sqrt{2\pi u_n} \exp\left(\frac{1}{2} S^2 u_n\right)$.

3. Derive an estimator for the hyperparameter $\eta := \tau^2$ as in (4.12), i.e. as the maximiser of the marginal likelihood. Verify that this estimator is given by
$$\hat{\tau}^2 = \max\left\{0, \frac{\left(\sum_1^n x_i\right)^2 - n}{n^2}\right\}.$$

4. Conclude that an empirical Bayes estimator is given by
$$\hat{\Theta}^{EB} = \frac{\hat{\tau}^2}{n\hat{\tau}^2 + 1} n\hat{X}.$$

For the setting of the previous exercise we can now compare the following three estimators for θ:

- the posterior mean
$$\frac{\tau^2}{1 + n\tau^2} n\bar{X}_n$$
- the derived empirical Bayes estimator
$$\frac{\hat{\tau}^2}{1 + n\hat{\tau}^2} n\bar{X}_n, \qquad \text{with} \qquad \hat{\tau}^2 = \max\left(0, \frac{(n\bar{X}_n)^2 - n}{n}\right)$$
- the maximum likelihood estimator
$$\bar{X}_n.$$

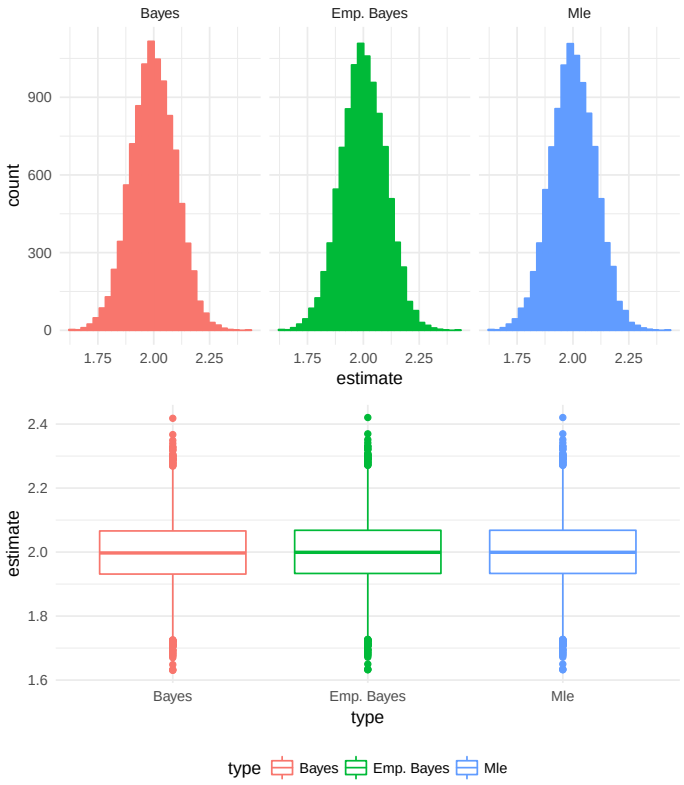

Figure 4.2: Comparison of posterior mean (Bayes) with $\tau^2 = 10$, empirical Bayes (Emp. Bayes) and maximum likelihood estimator by a Monte Carlo study. Each Monte Carlo sample is sampled as $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$ with $\theta = 2$.

We compare the performance of these estimators by a Monte Carlo study where we took $\tau^2 = 10$ in the posterior mean and 10^4 Monte Carlo samples. Each Monte Carlo sample is sampled as $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$. Figures 4.2 and 4.3 show the results for $\theta = 2$ and $\theta = 0$ respectively. In case $\theta = 2$ all three estimators perform roughly the same. In case $\theta = 0$ the situation is rather different: the empirical Bayes estimator is exactly equal to zero whenever $(n\bar{X}_n)^2 - n < 0$, i.e. when $|\bar{X}_n| < 1/\sqrt{n}$. This comes at the cost of slightly worse behaviour than the posterior mean and maximum likelihood estimator at ± 0.15 approximately.

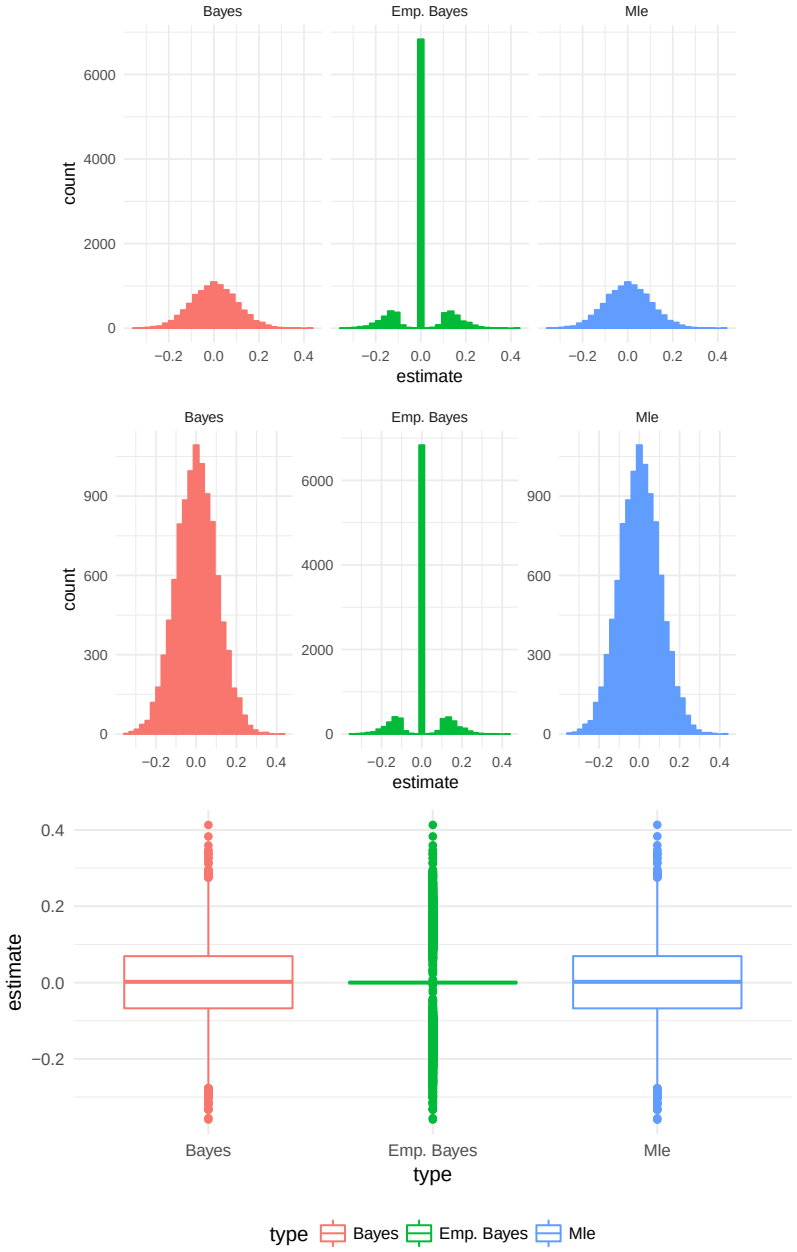

Figure 4.3: Comparison of posterior mean (Bayes) with $\tau^2 = 10$, empirical Bayes (Emp. Bayes) and maximum likelihood estimator by a Monte Carlo study. Each Monte Carlo sample is sampled as $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$ with $\theta = 0$. Note the difference in vertical scale of the middle and lower figures.

Exercise 4.20 Assume $X_1, \ldots, X_p$ are independent conditional on $\Theta_1, \ldots, \Theta_p$. Suppose $X_i \mid \Theta_i = \theta_i \sim Unif\left(0, \theta_i\right)$. Consider estimation of the parameters $\Theta_1, \ldots, \Theta_p$ based on data $X_1, \ldots, X_p$.

(a) Model $\Theta_1, \ldots, \Theta_p$ as independent with common density

$$f_\Theta(\theta) = \theta\lambda^2 e^{-\lambda\theta}\mathbf{1}_{[0,\infty)}(\theta).$$

Find the posterior mean for Θ_i $(1 \leq i \leq p)$.

(b) Verify that $\mathbb{E}[X_i] = 1/\lambda$ and explain why $\hat{\Lambda} = 1/\bar{X}_n$ is an intuitively reasonable estimator for λ.

Hint; use $\mathbb{E}[X_i] = \mathbb{E}[\mathbb{E}[X_i \mid \Theta_i]]$.

(c) Determine λ by marginal maximum likelihood.

(d) Combine parts (a) and (b) (or (c)) to find empirical Bayes estimators for $\Theta_1, \ldots, \Theta_p$.

Exercise 4.21 [YS exercise 3.13.] Let $X \sim Bin(n, \theta)$ and consider a conjugate $Be(\alpha, \beta)$ prior distribution for θ.

1. Show that if we reparametrise from (α, β) to (μ, M), where $\mu = \alpha/(\alpha + \beta)$ and $M = \alpha + \beta$, the marginal distribution of X if of *beta-binomial* form:

$$\mathbb{P}(X = x) = \frac{\Gamma(M)}{\Gamma(M\mu)\Gamma(M(1-\mu))}\binom{n}{x}\frac{\Gamma(x + M\mu)\Gamma(n - x + M(1-\mu)}{\Gamma(n + M)}.$$

   Hint: first show

$$\mathbb{P}(X = x) = \frac{1}{B(\mu M, (1-\mu)M)}\int_0^1 \binom{n}{x}\theta^{x+\mu M-1}(1-\theta)^{n-x+(1-\mu)M-1}\,\mathrm{d}\theta,$$

   where B denotes the Beta-function.

2. Verify that the marginal expectation and variance of X/n are respectively

$$\mathbb{E}[X/n] = \mu \qquad \mathrm{Var}(X/n) = \frac{\mu(1-\mu)}{n}\left(1 + \frac{n-1}{M+1}\right).$$

   In calculating $\mathrm{Var}(X/n)$ it is handy to use the law of total variance:

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathrm{E}[Y|\Theta]) + \mathrm{E}[\mathrm{Var}(Y|\Theta)].$$

3. Consider a model in which (X_i, θ_i), $i = 1, \ldots, k$ are a sequence of independent, identically distributed random variables, for which only the X_i are observable. Suppose that $X_i \mid \Theta_i = \theta_i \sim Bin\left(n, \theta_i\right)$ and $\Theta_i \sim Be(\mu, M)$ in terms of the above parametrisation of the beta distribution.

   (a) Show that

$$\mathbb{E}[\Theta_i \mid X_i] = \frac{X_i + M\mu}{M + n}.$$

   (b) How can you find empirical Bayes estimators for θ_i?

4.7 Bayesian asymptotics

This section is based upon section 4.1 in Ghosh et al. [2006]. In this section we study the posterior distribution from the frequentist point of view. Recall that for observed data x, the posterior measure is denoted by Π_x (which is determined by sampling probabilities P_θ and a prior measure μ_Θ). Now if we assume the frequentist point of view where $X \sim \mathrm{P}_{\theta_0}$, how does the posterior Π_X behave? Note that for each realisation x of X, the posterior Π_x is a probability measure on the parameter set Ω. So in fact what we are considering here is a distribution on probability measures.

Suppose $X \equiv \mathbf{X}_n = (X_1, \ldots, X_n)$ and that when n increases the amount of information about the parameter increases. The typical setting would be where $X_1, \ldots, X_n$ are independent with a common distribution depending on θ. **Frequentist validation of Bayesian procedures** consists of verifying that under

the assumption $X \sim \mathrm{P}_{\theta_0}$, the posterior $\Pi_{\mathbf{X}_n}$ converges to δ_{θ_0}, as $n \to \infty$ [5]. That is, the posterior measure concentrates at the "true" parameter θ_0 asymptotically. If this is the case, we say the posterior is **consistent**.

Theorem 2.33 shows that under certain conditions, maximum likelihood estimators are asymptotically normal with asymptotic variance the inverse of the Fisher information. It turns out that, under mild conditions, the posterior distribution is the same in the large sample limit.

4.7.1 Consistency

We start with an example.

Example 4.37. Take $X_1, \dots, X_n \mid \Theta = \theta \overset{\text{ind}}{\sim} Ga(2, \theta)$. Then

$$f_{\mathbf{X}_n|\Theta}(\mathbf{X}_n \mid \theta) = \prod_{i=1}^{n} \theta^2 X_i e^{-\theta X_i}.$$

If $\Theta \sim Exp(\lambda)$, then

$$f_{\Theta|\mathbf{X}_n}(\theta \mid \mathbf{X}_n) \propto \theta^{2n} e^{-(\lambda + S_n)\theta},$$

where $S_n = \sum_{i=1}^{n} X_i$. It follows that $\Theta \mid \mathbf{X}_n \sim Ga\left(2n+1, \lambda + S_n\right)$ and

$$\mathrm{E}[\Theta \mid \mathbf{X}_n] = \frac{2n+1}{\lambda + S_n} \quad \text{and} \quad \mathrm{Var}(\Theta \mid \mathbf{X}_n) = \frac{2n+1}{(\lambda + S_n)^2}.$$

In fact, since in general the $Ga(m, \beta)$ distribution is close to the $N\left(m/\beta, m/\beta^2\right)$ distribution if m is large, we have

$$\Theta \mid \mathbf{X}_n \overset{\approx}{\sim} N\left(\frac{2n+1}{\lambda + S_n}, \frac{2n+1}{(\lambda + S_n)^2}\right)$$

$\overset{\approx}{\sim}$ denoting informally "approximately distributed as". Hence the posterior is asymptotically normal. The parameters of this normal distribution are stochastic, as they depends on $\mathbf{X}_n$.

A frequentist would start of with the model

$$X_1, \dots, X_n \overset{\text{iid}}{\sim} Ga\left(2, \theta_0\right)$$

for some fixed unknown "true" (or "data-generating") value θ_0. It is a natural question to aks oneself how the posterior distribution behaves, assuming the frequentist data-generating model. Clearly, $S_n/n \to 2/\theta_0$ with P_{θ_0}-probability 1. This implies that under the frequentist model the distribution of $\sqrt{n}(\Theta - \theta_0) \mid \mathbf{X}_n$ asymptotically agrees with the $N\left(0, \theta_0^2/2\right)$ distribution. In theorem 4.39 we will make this statement precise in a more general setup.

The following lemma shows that posterior consistency is obtained in this example. In the following, we use the notation

$$m(x) = \int \theta \Pi_x(\,\mathrm{d}\theta) \qquad v(x) = \int (\theta - m(x))^2 \Pi_x(\,\mathrm{d}\theta).$$

Lemma 4.38. Suppose $m(\mathbf{X}_n) = \mathrm{E}[\Theta \mid \mathbf{X}_n] \xrightarrow{\mathrm{p}} \theta_0$ and $v(\mathbf{X}_n) = \mathrm{var}(\Theta \mid \mathbf{X}_n) \xrightarrow{\mathrm{p}} 0$ with P_{θ_0}-probability 1. Then the sequence of posterior distributions is consistent: for any $\varepsilon > 0$, with P_{θ_0} probability 1

$$\lim_{n \to \infty} \Pi_{\mathbf{X}_n}(B(\theta_0, \varepsilon)) = 1,$$

where $B(\theta_0, \varepsilon) = \{\theta : |\theta - \theta_0| \leq \varepsilon\}$.

[5] Convergence may refer to for example weak convergence or convergence in total variation, the details are out of scope for this course.

Proof. First fix x. By the triangle inequality

$$|\theta - \theta_0| \leq |\theta - m(x)| + |m(x) - \theta_0|.$$

Hence

$$\Pi_x\left(\{\theta : \ |\theta - m(x)| < \epsilon/2 \text{ and } |m(x) - \theta_0| < \epsilon/2\}\right) \leq \Pi_x\left(\{\theta : \ |\theta - \theta_0| < \epsilon\}\right).$$

Taking the complement (and using the Morgan's law), we get

$$\begin{aligned} 1 - \Pi_x(B(\theta_0, \varepsilon)) &\leq \Pi_x\left(B(m(x), \epsilon/2)^c\right) + \mathbf{1}\{|\theta_0 - m(x)| > \epsilon/2\} \\ &\leq \frac{v(x)}{\epsilon^2/4} + \mathbf{1}\{|\theta_0 - m(x)| > \epsilon/2\} \end{aligned}$$

where we used Chebyshev's inequality at the second inequality. Now substitute $\mathbf{X}_n$ for x and consider the limit $n \to \infty$ under P_{θ_0}. Then $v(\mathbf{X}_n)$ converges in probability to zero, by assumption. Taking the expectation under θ_0 of the second term, we see that term tends to zero in L^1, as we assumed $m(\mathbf{X}_n) - \theta_0$ to convergence in probability to zero. As convergence in L^1 is stronger than convergence in probability, the second term converges in probability to zero. We conclude that $1 - \Pi_x(B(\theta_0, \varepsilon))$ converges to zero under P_{θ_0}. □

Exercise 4.22 Suppose $X_1, \ldots, X_n$ are conditionally independent given $\Theta = \theta$ with common $Pois\,(\theta)$-distribution. Show that the posterior is consistent when the prior distribution on Θ is the Gamma-distribution.
Hint: show that

$$\mathrm{E}_{\theta_0}[\Theta \mid X_1, \ldots, X_n] \to \theta_0, \qquad \mathrm{Var}_{\theta_0}(\Theta \mid X_1, \ldots, X_n) \to 0.$$

The following exercise shows that if the parameter set is finite, posterior consistency can be proved under mild conditions.

Exercise 4.23 Suppose

$$X_1, \dots, X_n \mid \Theta = \theta \overset{\text{iid}}{\sim} \mathrm{P}_\theta$$

where $\theta \in \Omega = \{\theta_1, \dots, \theta_p\}$ (hence, the parameter set is finite). Suppose the prior $\mu_\Theta(\theta_k) = \xi_k$, with of course $\sum_{k=1}^p \xi_k = 1$. Assume that $\mathrm{P}_\theta \ll \nu$, with ν Lebesgue measure, and denote its density by $\mathrm{d}P_\theta / \mathrm{d}\nu = f(\cdot \mid \theta)$. In the following we denote $X = (X_1, \dots, X_n)$, and by x a realisation of X.

1. Derive that the (Bayesian) marginal measure of X is given by

$$\mu_X(B) = \int_B \int_\Omega \prod_{i=1}^n f(x_i \mid \theta)\, \mathrm{d}\mu_\Theta(\theta)\, \mathrm{d}x_1 \dots, \mathrm{d}x_n.$$

2. Show that for a measurable $A \subset \Omega$ the posterior Π_x is given by

$$\Pi_x(A) = \frac{\int_A \prod_{i=1}^n f(x_i \mid \theta)\, \mathrm{d}\mu_\Theta(\theta)}{\int_\Omega \prod_{i=1}^n f(x_i \mid \theta)\, \mathrm{d}\mu_\Theta(\theta)}.$$

3. Show that for $\ell \in \{1, \dots, p\}$ we have the representation

$$\Pi_x(\{\theta_\ell\}) = \frac{\xi_\ell L(\theta_\ell, x)}{\sum_{k=1}^p \xi_k L(\theta_k, x)}$$

   and identify $L(\theta, x)$.

4. Show that

$$\Pi_x(\{\theta_\ell\}) = \frac{\xi_\ell}{\sum_{k=1}^p \xi_k e^{-n z_{k,\ell}(x)}}$$

   with

$$z_{k,\ell}(x) = \frac{1}{n} \log \frac{L(\theta_\ell, x)}{L(\theta_k, x)}.$$

5. Now take a frequentist point of view where we assume the data are generated with θ_ℓ. Assume the model is identifiable. Consider the estimator $T = \Pi_X(\{\theta_\ell\})$ (note that the stochasticity is in X). Show that T converges in probability to 1 as $n \to \infty$. In other words, the posterior concentrates on $\{\theta_\ell\}$ asymptotically.

   Hint: law of large numbers, Kullback-Leiber divergence

4.7.2 Asymptotic normality of the posterior

We first present a quick heuristic derivation. Let $\tilde{\Theta}_n$ be the posterior mode of the logposterior. A Taylor expansion of the log-posterior about $\tilde{\Theta}_n$ gives:

$$\begin{aligned}
\log f_{\Theta|\mathbf{X}_n}(\theta \mid \mathbf{X}_n) &= \log f_{\Theta|\mathbf{X}_n}(\tilde{\Theta}_n \mid \mathbf{X}_n) + (\theta - \tilde{\Theta}_n) \frac{\partial}{\partial \theta} \log f_{\Theta|\mathbf{X}_n}(\theta \mid \mathbf{X}_n)\Big|_{\theta=\tilde{\Theta}_n} \\
&\quad - \frac{1}{2}(\theta - \tilde{\Theta}_n)^2 \tilde{I}_n + \cdots \\
&\approx \log f_{\Theta|\mathbf{X}_n}(\tilde{\Theta}_n \mid \mathbf{X}_n) - \frac{1}{2}(\theta - \tilde{\Theta}_n)^2 \tilde{I}_n,
\end{aligned}$$

where

$$\tilde{I}_n = -\frac{\partial^2}{\partial\theta^2} \log f_{\Theta|\mathbf{X}_n}(\theta \mid \mathbf{X}_n)\Big|_{\theta=\tilde{\Theta}_n}.$$

Hence

$$f_{\Theta|\mathbf{X}_n}(\theta \mid \mathbf{X}_n) \propto \exp\left(-\frac{1}{2}(\theta - \tilde{\Theta}_n)^2 \tilde{I}_n\right).$$

which means that the posterior distribution of Θ is approximately $N\left(\tilde{\Theta}_n, \tilde{I}_n^{-1}\right)$.

Denote the k-th derivative of the loglikelihood with respect to θ by by $l^{(k)}(\theta \mid \mathbf{X}_n)$.

Define the following assumptions:

(A1) The set $C = \{x : f_{X_1|\Theta}(x \mid \theta) > 0\}$ is the same for all θ.

(A2) $\log f_{X_1|\Theta}(x \mid \theta)$ is thrice differentiable with respect to θ in a neighbourhood $(\theta_0 - \delta, \theta_0 + \delta)$ of θ_0 The expectations $\mathrm{E}_{\theta_0} l^{(1)}(\theta_0 \mid X_1)$ and $\mathrm{E}_{\theta_0} l^{(2)}(\theta_0 \mid X_1)$ are both finite and

$$\sup_{\theta\in(\theta_0-\delta,\theta_0+\delta)} |l^{(3)}(\theta \mid X_1)| \leq M(x) \quad \text{and} \quad \mathrm{E}_{\theta_0} M(X_1) < \infty.$$

(A3) Interchange of the order of integration with respect to P_{θ_0} and differentiation at θ_0 is justified, so that

$$\mathrm{E}_{\theta_0} l^{(1)}(\theta_0 \mid X_1) = 0 \quad \text{and} \quad \mathrm{E}_{\theta_0} l^{(2)}(\theta_0 \mid X_1) = -\mathrm{E}_{\theta_0}\left(l^{(1)}(\theta_0 \mid X_1)\right)^2.$$

Furthermore $I(\theta_0; X_1) = E_{\theta_0}\left(l^{(1)}(\theta_0 \mid X_1)\right)^2 < \infty$.

(A4) For any $\delta > 0$

$$\mathrm{P}_{\theta_0}\left(\forall\,\varepsilon > 0, \exists N \in \mathbb{N} \text{ s.t. } \forall\, n \geq N \quad \sup_{|\theta-\theta_0|>\delta} (l(\theta \mid \mathbf{X}_n) - l(\theta_0 \mid \mathbf{X}_n)) < -n\varepsilon\right) = 1$$

The condition that is hardest to check is (A4). A common assumption for deriving frequentist asymptotic results involve the behaviour of $l(\theta \mid \mathbf{X}_n)$ in the neighbourhood of θ_0. Since Bayesian estimators involve integration over the whole of Ω, it is also necessary to control $l(\theta \mid \mathbf{X}_n)$ at a distance from θ_0. Condition (A4) turns this requirement in a formal mathematical condition. Note that

$$\frac{l(\theta \mid \mathbf{X}_n) - l(\theta_0 \mid \mathbf{X}_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f_{X_i|\Theta}(x_i \mid \theta)}{f_{X_i|\Theta}(x_i \mid \theta_0)}.$$

If the sequence $\hat{\Theta}_n$ is strongly consistent for θ_0. In that case, it can be shown that $\hat{\Theta}_n$ satisfies the score equation $l^{(1)}(\theta \mid \mathbf{X}_n) = 0$.

The following theorem is known as the **Bernstein-Von Mises** theorem.

Theorem 4.39 (Theorem 4.2 in Ghosh et al. [2006])**.** In addition to (A1)-(A4), assume

- $\theta_0 \in \Omega$ which is an open subset of $\mathbb{R}$;
- the prior f_Θ is continuous and positive at θ_0;
- $\hat{\Theta}_n$ is a strongly consistent solution of the score equation.

If $\Psi_n = \sqrt{n}(\Theta - \hat{\Theta}_n)$, then

$$\lim_{n\to\infty} \int \left| f_{\Psi_n|\mathbf{X}_n}(\psi \mid \mathbf{X}_n) - \phi(\psi; 0, I(\theta_0)^{-1}) \right| \mathrm{d}\psi = 0$$

with P_{θ_0}-probability 1.

The result holds as well if $I(\theta_0)$ is replaced by $\hat{I}_n/n$, where

$$\hat{I}_n = -\frac{\partial^2}{\partial\theta^2} \log f_{\mathbf{X}_n|\Theta}(\mathbf{X}_n \mid \theta)\Big|_{\theta=\hat{\Theta}_n}.$$

This theorem is useful since

- It shows that if we have a large number of observations the prior does not matter (the prior is "washed away"/"overridden" by the data). Note that we assume the number of parameters in the statistical model does not grow with n, as sometimes happens to be the case in hierarchical models.
- The normal approximation to the posterior justifies summarising the posterior by its mean and standard-deviation.
- It offers computational simplicity: whereas computation of the posterior mean is hard, computation of the posterior mode is usually much easier.

*Proof of theorem 4.39**. [6] In the proof we write $l_n(\theta)$ instead of $l(\theta \mid \mathbf{X}_n)$. For the posterior we have

$$f_{\Theta|\mathbf{X}_n}(\theta \mid \mathbf{X}_n) \propto \prod_{i=1}^{n} \frac{l_n(\theta)}{l_n(\hat{\Theta}_n)} f_\Theta(\theta) = \exp(L_n(\theta) - L_n(\hat{\Theta}_n)) f_\Theta(\theta).$$

Let $\Psi_n = \sqrt{n}(\Theta - \hat{\Theta}_n)$, then

$$f_{\Psi_n|\mathbf{X}_n}(\psi \mid \mathbf{X}_n) \propto \exp(l_n(\hat{\Theta}_n + \psi/\sqrt{n}) - l_n(\hat{\Theta}_n)) f_\Theta(\hat{\Theta}_n + \psi/\sqrt{n}) =: h_n(\psi).$$

(the Jacobian term equals $1/\sqrt{n}$ and is absorbed into the proportionality constant). If we define $C_n = \int h_n(\psi) d\psi$, then $f_{\Psi_n|\mathbf{X}_n}(\psi \mid \mathbf{X}_n) = C_n^{-1} h_n(\psi)$. We wish to show that

$$\int \left| C_n^{-1} h_n(\psi) - \sqrt{\frac{I(\theta_0)}{2\pi}} e^{-\frac{1}{2}\psi^2 I(\theta_0)} \right| \mathrm{d}\psi$$

tends to zero. This integral can be bounded by

$$C_n^{-1} \int \left| h_n(\psi) - f_\Theta(\theta_0) e^{-\frac{1}{2}\psi^2 I(\theta_0)} \right| \mathrm{d}\psi + \int \left| C_n^{-1} f_\Theta(\theta_0) e^{-\frac{1}{2}\psi^2 I(\theta_0)} - \sqrt{\frac{I(\theta_0)}{2\pi}} e^{-\frac{1}{2}\psi^2 I(\theta_0)} \right| \mathrm{d}\psi$$

Suppose

$$B_n := \int \left| h_n(\psi) - f_\Theta(\theta_0) e^{-\frac{1}{2}\psi^2 I(\theta_0)} \right| \mathrm{d}\psi \to 0$$

Then $C_n \to \int f_\Theta(\theta_0) e^{-\frac{1}{2}\psi^2 I(\theta_0)} d\psi = f_\Theta(\theta_0)\sqrt{2\pi I(\theta_0)}$ which implies both terms in the bound tend to zero.

[6] The proof is not part of the exam.

Hence, it suffices to show $B_n \to 0$. We do this by separately considering the integrals over $A_1 = \{\psi : |\psi| > \delta_0\sqrt{n}\}$ and $A_2 = \{\psi : |\psi| \leq \delta_0\sqrt{n}\}$. Denote these integrals by I and II respectively. In the following, the limits are understood to hold with P_{θ_0} probability 1.

Bounding I: The integral over domain A_1 is bounded by

$$\int_{A_1} h_n(\psi)\,\mathrm{d}\psi + \int_{A_1} f_\Theta(\theta_0) e^{-\frac{1}{2}\psi^2 I(\theta_0)}\,\mathrm{d}\psi.$$

It is easy to see that the second integral tends to zero. For the first integral, if $\psi \in A_1$ we have $l_n(\hat{\Theta}_n + \psi/\sqrt{n}) - l_n(\hat{\Theta}_n) < -\varepsilon n$ for n sufficiently large and hence the integral tends to zero if $n \to \infty$.

Bounding II: If suffices to show that $\int_{A_2} h_n(\psi)d\psi \to 0$. Using that $l_n^{(1)}(\hat{\Theta}_n) = 0$, a Taylor expansion of the loglikelihood around $\hat{\Theta}_n$ gives

$$l_n(\hat{\Theta}_n + \psi/\sqrt{n}) - l_n(\hat{\Theta}_n) = -\frac{1}{2}\psi^2 \hat{I}_n + R_n(\psi),$$

where $R_n(\psi) = \frac{1}{6}(\psi/\sqrt{n})^3 l_n^{(3)}(\Theta_n')$ and $\hat{\Theta}_n'$ lies in between $\hat{\Theta}_n$ and $\hat{\Theta}_n + \psi/\sqrt{n}$. For $\psi \in A_2$ and n sufficiently large

$$|R_n(\psi)| \leq \frac{1}{6}\frac{\delta_0/\sqrt{n}}{\sqrt{n}}\left(\frac{\psi}{\sqrt{n}}\right)^2 \left|l_n^{(3)}(\Theta_n')\right| \leq \frac{1}{6}\delta_0\psi^2 \frac{1}{n}\sum_{i=1}^{n} M(X_i)$$

By choosing δ_0 sufficiently small, this can be bounded by $\frac{1}{4}\psi^2\hat{I}_n$. This implies

$$\exp\left(l_n(\hat{\Theta}_n + \psi/\sqrt{n}) - l_n(\hat{\Theta}_n)\right) \leq \exp\left(-\frac{1}{4}\psi^2\hat{I}_n\right) \leq \exp\left(-\frac{1}{4}\psi^2 I(\theta_0)\right),$$

for n sufficiently large. The final inequality follows from $\hat{I}_n \to I(\theta_0)$. It is easy to see that the integral $\exp\left(-\frac{1}{4}\psi^2 I(\theta_0)\right)$ over A_2 tends to zero, as $n \to \infty$. □

The following corollary is proved in Schervish [1995] (Theorem 7.101). It shows that posterior probabilities converge in probability under P_{θ_0}. for any (Borel) subset B of $\mathbb{R}$

Corollary 4.40. Define $\Lambda_n = \sqrt{\hat{I}_n}(\Theta - \hat{\Theta}_n)$. Under "regularity conditions",

$$\mathbb{P}(\Lambda_n \in B \mid \mathbf{X}_n) \xrightarrow{\mathrm{p}} \Phi(B) \quad \text{under } P_{\theta_0}, \text{ as } n \to \infty.$$

Here $\Phi(B)$ denotes the probability that a standard normal random variable lies in B.

Chapter 5

Bayesian computation

The posterior distribution is virtually always intractable. The field of Bayesian computation is centred on computational techniques to approximate the posterior distribution or sample from it.

5.1 The Metropolis-Hastings algorithm

Consider a dominated Bayesian statistical model, as in Section 4.1.3. Assume that the posterior density with respect to the dominating measure ξ is given by

$$f_{\Theta|X}(\theta \mid x) = \frac{f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)}{\int f_{X|\Theta}(x \mid \theta) f_\Theta(\theta) \xi(\mathrm{d}\theta)}.$$

Except for exceptional simple cases the integral in the denominator, which is the normalising constant $f_X(x)$, is intractable. Its evaluation possibly poses a high-dimensional integration problem, in case the dimension of hte parameter is large. Markov Chain Monte Carlo methods is a collection of techniques that can be used to obtain (dependent) samples from the posterior distribution.The main algorithm is the Metropolis-Hastings (MH) algorithm.

Definition 5.1. A **Markov chain Monte Carlo (MCMC)** method for sampling from a distribution π is any method producing an ergodic Markov chain whose stationary distribution is π.

For ease of exposition, first we consider the problem of sampling from the probability mass function π, supported on a countable set $\mathcal{X}$, where

$$\pi(x) = \mathbb{P}(X = x), \qquad x \in \mathcal{X}.$$

We will refer to this "density" as the **target density**.

Warning on notation: Whereas our motivation for studying MCMC methods is drawing from the posterior, i.e. drawing samples from $f_{\Theta|X}(\cdot \mid x)$, from now on we will consider the generic problem of drawing samples from a density $\pi(x)$, where x takes values in the (measurable) space $\mathcal{X}$. The reason for this is that MCMC is a stochastic simulation method, which also has wide applicability outside Bayesian statistics (for example in statistical mechanics).

The MH-algorithm requires as input an irreducible[1] Markov chain on $\mathcal{X}$, say with transition probabilities $q(x, y)$, $x, y \in \mathcal{X}$ (so each row is a vector consisting of numbers in $[0, 1]$ that sum to 1; in case $\mathcal{X}$ is finite-dimensional, q is just a matrix). Hence $q(x, y)$ is the probability of moving a point x one time instance further

[1] All states communicate: with positive probability state j can be reached from state i.

to the point y. Put differently, for each x, $y \mapsto q(x, y)$ is a probability vector, it plays the role of a conditional probability. As such, q is referred to as the **proposal density**. The output of this algorithm is a Markov chain $\{X_n\}$ that has π as invariant distribution. Under weak additional assumptions

$$\frac{1}{N}\sum_{n=1}^{N} g(X_n) \xrightarrow{\text{a.s.}} \mathrm{E}_\pi\, g(X), \qquad N \to \infty,$$

for π-integrable functions g (a precise statement is given by Theorem 5.8). From this description it is apparent that there is huge freedom in choosing q.

Definition 5.2. The **Metropolis-Hastings (MH) algorithm** is the algorithm by which a Markov chain is constructed which evolves $x_n = x$ to x_{n+1} by the following steps

1. propose y from a proposal density $q(x, \cdot)$;
2. Compute
$$\alpha(x, y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\frac{q(y, x)}{q(x, y)}\right).$$
3. Set
$$x_{n+1} = \begin{cases} y & \text{with probability } \alpha(x, y) \\ x & \text{with probability } 1 - \alpha(x, y) \end{cases}.$$

For computing $\alpha(x, y)$ it suffices to know π up to a proportionality constant. Within Bayesian statistics, this is a very attractive property of the algorithm, as it avoids computing $f_X(x)$.

To understand the algorithm, note that the transition probabilities of the Markov chain defined by the MH-algorithm are given by

$$p(x, y) = \begin{cases} q(x, y)\alpha(x, y) & \text{if} \quad x \neq y \\ q(x, x) + \sum_{z \neq x} q(x, z)(1 - \alpha(x, z)) & \text{if} \quad x = y \end{cases}.$$

Hence, the MH-acceptance rule adjusts the transition probabilities from q to p. Then for $y \neq x$

$$\begin{aligned} \pi(x)p(x, y) &= \pi(x)q(x, y)\min\left(1, \frac{\pi(y)}{\pi(x)}\frac{q(y, x)}{q(x, y)}\right) \\ &= \min\left(\pi(x)q(x, y), \pi(y)q(y, x)\right). \end{aligned}$$

As the expression on the right-hand-side is symmetric in x and y we obtain

$$\pi(x)p(x, y) = \pi(y)p(y, x) \qquad \text{for all } x \text{ and } y.$$

This relation is generally referred to as **detailed balance**. Summing over x on both sides gives

$$\pi(y) = \sum_x \pi(x)p(x, y).$$

This reveals that π is invariant for the chain: if we draw x according to π and let the chain evolve, the distribution at all future times will be exactly π.

5.1.1 A general formulation of the Metropolis-Hastings algorithm*

In this section we give a more general formulation of the MH-algorithm, relaxing the assumption that $\mathcal{X}$ is countable. We closely follow the exposition in chapter 5 of Dashti and Stuart [2013]. Let π_0 be a reference measure on a separable Banach space $\mathcal{X}$ equipped with the Borel σ-algebra $\mathcal{B}$. We assume that $\pi \ll\ll \pi_0$ and we are interested in sampling from π. Let P denote a Markov transition kernel (Cf. Definition 4.1).

Definition 5.3. The Markov chain with transition kernel P is **invariant** with respect to π if for each $x \in \mathcal{X}$

$$\int_{\mathcal{X}} \pi(\mathrm{d}x) P(x, \cdot) = \pi(\cdot)$$

as measures on $(\mathcal{X}, \mathcal{B})$. The Markov chain is said to satisfy **detailed balance** with respect to π if

$$\pi(\mathrm{d}x) P(x, \mathrm{d}y) = \pi(\mathrm{d}y) P(y, \mathrm{d}x)$$

as measures on $(\mathcal{X} \times \mathcal{X}, \mathcal{B} \otimes \mathcal{B})$. The resulting Markov chain is then said to be **reversible** with respect to π.

By integrating the detailed balance relation with respect to x it follows that if a Markov chain is reversible with respect to π then it is invariant with respect to π.

The MH-algorithm accepts-rejects proposals from a Markov kernel Q to produce a Markov chain with kernel P which is reversible with respect to π. Thus, the chain evolves $x_n = x$ to x_{n+1} by the following steps.

1. Propose y from a Markov kernel Q: $y \sim Q(x, dy)$.
2. Compute $\alpha(x, y)$ and set

$$x_{n+1} = \begin{cases} y & \text{with probability } \alpha(x, y) \\ x & \text{with probability } 1 - \alpha(x, y) \end{cases}.$$

Proposition 5.4. The kernel of the MH-chain is given by

$$P(x, \mathrm{d}y) = Q(x, \mathrm{d}y)\alpha(x, y) + \delta_x(\mathrm{d}y) \int_{\mathcal{X}} (1 - \alpha(x, y)) Q(x, \mathrm{d}y). \tag{5.1}$$

Proof. Suppose $X_n = x$. To evolve the chain to time $n + 1$ we independently draw $U \sim Unif(0, 1)$ and $Y_{n+1} \sim Q(x, \cdot)$. The algorithm prescribes that X_{n+1} either equals X_n or Y_{n+1}, depending on the event $\{U < \alpha(X_n, Y_{n+1})$. Hence

$$\begin{aligned} P(x, B) &= \mathbb{P}(X_{n+1} \in B \mid X_n = x) \\ &= \mathbb{P}(Y_{n+1} \in B, U < \alpha(X_n, Y_{n+1}) \mid X_n = x) + \mathbb{P}(X_n \in B, U \geq \alpha(X_n, Y_{n+1}) \mid X_n = x) \end{aligned}$$

The first term can be rewritten to

$$\int_B \int \mathbf{1}_{\{u < \alpha(x,y)\}} \mathrm{d}u \, Q(x, \mathrm{d}y) = \int_B \alpha(x, y) Q(x, \mathrm{d}y).$$

The second term can be written as

$$\mathbf{1}_B(x) \int_{\mathcal{X}} \int \mathbf{1}_{\{u \geq \alpha(x,y)\}} \mathrm{d}u \, Q(x, \mathrm{d}y) = \mathbf{1}_B(x) \int_{\mathcal{X}} (1 - \alpha(x, y)) Q(x, \mathrm{d}y).$$

□

Given the kernel Q, a key question is how $\alpha : \mathcal{X} \times \mathcal{X} \to [0,1]$ should be chosen to ensure $P(x, \mathrm{d}y)$ satisfies detailed balance with respect to π. The following theorem is due to Tierney [1998] (Theorem 2).

Theorem 5.5. Define the measures

$$\nu(\mathrm{d}x, \mathrm{d}y) = \pi(dx)Q(x, \mathrm{d}y)$$
$$\nu^T(\mathrm{d}x, \mathrm{d}y) = \pi(\mathrm{d}y)Q(y, \mathrm{d}x)$$

on $(\mathcal{X} \times \mathcal{X}, \mathcal{B} \otimes \mathcal{B})$. Assume ν and ν^T are absolutely continuous and that

$$\nu(\mathrm{d}x, \mathrm{d}y) = r(x, y)\nu^T(\mathrm{d}x, \mathrm{d}y).$$

Then the kernel P defined in (5.1) satisfies detailed balance with respect to π if and only if

$$r(x, y)\alpha(x, y) = \alpha(y, x), \quad \nu - a.s.$$

In particular, the choice $\alpha_{MH}(x, y) = \min(1, r(y, x))$ will imply detailed balance.

The final statement can easily be verified:

$$\alpha_{MH}(x, y)r(x, y) = \min(r(x, y), r(y, x)r(x, y)) = \min(r(x, y), 1) = \alpha_{MH}(y, x).$$

This is a general formulation of the MH-algorithm which even applies to infinite dimensional settings. The result gets less abstract and easier to comprehend in case there is a common dominating measure. Suppose there is a measure μ such that

$$Q(x, \mathrm{d}y) = q(x, y)\mu(\mathrm{d}y) \quad \text{and} \quad \pi(\mathrm{d}x) = \pi(x)\mu(\mathrm{d}x).$$

Then it is easily verified that

$$\nu(\mathrm{d}x, \mathrm{d}y) = \underbrace{\frac{\pi(x)q(x, y)}{\pi(y)q(y, x)}}\, \nu^T(\mathrm{d}x, \mathrm{d}y).$$

The underbraced part then equals $r(x, y)$. This corresponds exactly to the algorithm defined in Definition 5.2.

5.1.2 Convergence of the Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is constructed such that it has a specified measure π as invariant distribution. We need a little bit more to ensure that the algorithm will eventually yield samples from π. Luckily, the additional assumptions are mild. The results in this section apply to the case where $\mathcal{X} \subset \mathbb{R}^d$. For simplicity we assume that the measure π has a density with respect to Lebesgue measure, which is, with a slight abuse of notation, also denoted π.

Definition 5.6. A Markov chain is **π-irreducible** if, for any initial state, it has positive probability of entering any set to which π assigns positive probability.

Proposition 5.7. Assume that the proposal kernel Q has density q. If $q(x, y) > 0$ for all $x, y \in \mathrm{supp}(\pi)$, then the induced MH-chain is π-irreducibility.

The assumption is natural, as the induced MH-chain can only reach points that are proposed according to Q.

The following theorem gives sufficient conditions for convergence of MH-Markov chains.

Theorem 5.8 (Robert and Casella [2004], Theorem 7.4)**.**] Suppose the Metropolis-Hastings chain $(X_i, i \geq 1)$ is π-irreducible.

1. If $\int |h(x)|\pi(x)\,\mathrm{d}x < \infty$, then
$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(X_i) = \int h(x)\pi(x)\,\mathrm{d}x \qquad \pi - \text{a.s.}$$
2. If, in addition, $(X_i, i \geq 1)$ is aperiodic, then
$$\lim_{n\to\infty} \left\| \int P^n(x,\cdot)\,\mathrm{d}\mu_0(x) - \pi(\cdot) \right\|_{TV} = 0,$$
for any initial distribution μ_0.[2]

A sufficient condition for aperiodicity is that events $\{X_{i+1} = X_i\}$ have positive probability. This is equivalent to
$$\mathbb{P}(\pi(X_t)q(X_t, Y_t) \leq \pi(Y_t)q(Y_t, X_t)) < 1$$
where $X_t \sim \pi$ and $Y_t \sim Q(X_t, \cdot)$. That is to say, there should be positive probability of rejecting a proposal when the "current" iterate is drawn from π.

5.2 Examples of proposal kernels

There are numerous ways to construct the proposal kernel and the real challenge is to find a good one (in the sense that the support of the target density π is explored within the time the algorithm has run). We give a few examples of proposal kernels, where we assume a "continuous density" π.

1. **Random walk proposals**. Here we choose a tuning parameter $\sigma > 0$ and set $y = x + \sigma Z$ where Z has a symmetric distribution, for example the $N(0,1)$ or $Unif(0,1)$-distribution. In the former case, we have $Q(x, \mathrm{d}y) = \phi(y; x, \sigma^2)\,\mathrm{d}y$. On the one hand, fixing σ to a value too small, the the chain may take a long time before entering the stationary region, though having a high average acceptance rate. On the other hand, if σ is too large, most proposals will be rejected. Therefore, there is a need for tuning the algorithm to a specified average acceptance probability. For certain specific models, it has been proved that an average acceptance probability of about 0.25 is optimal. [3] As a rule of thumb many researchers advise the number to be in $(0.25, 0.5)$.
2. **Independent proposals**. Here we take $Q(x, \mathrm{d}y) = \bar{q}(y)\,\mathrm{d}y$ with $\bar{q}$ a probability density. Hence the proposal is independent of the current state. In this case
$$\alpha(x, y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\frac{\bar{q}(x)}{\bar{q}(y)}\right).$$
The acceptance probability is maximised for $\bar{q} = \pi$ which is of course intractable. This does show that ideally $\bar{q}$ resembles π.

[2] The **total variation distance** between two probability measures P and Q equals $\|P - Q\|_{TV} = \sup_A |\mathrm{P}(A) - Q(A)|$ (where the supremum is over all measurable sets). In case P and Q admit densities p and q with respect to a common dominating measure ν, then $\|P - Q\|_{TV} = \frac{1}{2}\int |p(x) - q(x)|\nu(\mathrm{d}x)$.

[3] Optimal refers to the Markov chain for which $\mathbb{E}\|X_n - X_{n-1}\}^2$ is maximal, where the expectation is over all proposed moves (including rejected ones), when the Markov chain has reached its stationary regimen. There are various possible notions for defining optimality, we refer to Chaper 4 in Brooks et al. [2011] for additional information. Citing from that source: "Best is to find reasonably large proposed moves which are reasonably likely to be accepted."

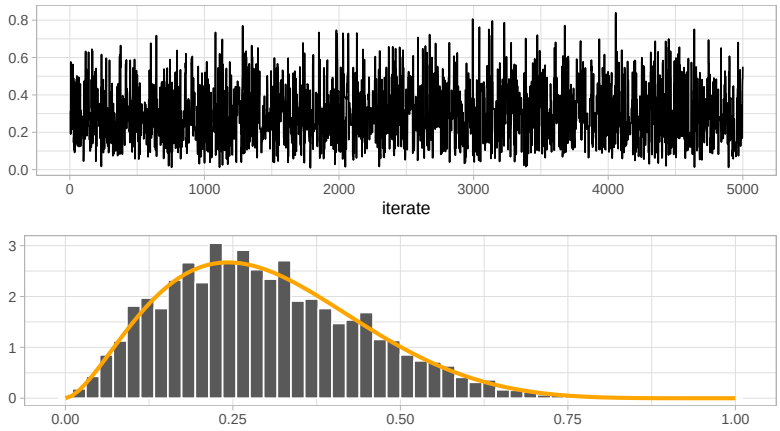

Figure 5.1: Output of the MH algorithm with independent $Unif\,(0,1)$ proposals. Top: trace plot. Bottom: histogram.

3. **Langevin adjusted proposals**. Here we choose a tuning parameter $h > 0$ and set

$$y = x + \frac{1}{2}h\nabla \log \pi(x) + \sqrt{h}Z,$$

where $Z \sim N\,(0,1)$. Hence

$$Q(x,\,\mathrm{d}y) = \phi(y; x + (h/2)\nabla \log \pi(x), h)\,\mathrm{d}y.$$

The rationality behind this choice is that π is invariant for the Langevin diffusion

$$\mathrm{d}X_t = \frac{1}{2}A\nabla \log \pi(X_t)\,\mathrm{d}t + \sqrt{A}\,\mathrm{d}W_t.$$

Here W is a Wiener process. The proposal follows upon Euler discretisation of this stochastic differential equation. The MH-acceptance probability corrects for the discretisation error made.

The derivation of the Langevin adjusted proposal is an example of a general strategy for finding MH-algorithms: construct a stochastic process that has π as invariant distribution. Ideally this process can be simulated without error. Otherwise, the MH-acceptance rule will correct for any (discretisation) error in simulating the process to ensure the resulting Markov chain has π as its invariant distribution.

Example 5.9. Suppose we wish to simulate from the $Be\,(a,b)$-distribution. Of course, there exist direct ways for simulating *independent* realisations of the beta distribution. We use the MH-algorithm to generate dependent draws from the Beta-distribution. First, we use an independence-sampler, where the proposals are independent draws from the $Unif\,(0,1)$-distribution with $a = 2.7$ and $b = 6.3$. The results are in figure 5.1. Next, we also use random-walk proposals, where given the current state x we propose

$$x' := x + Unif\,(-\eta, \eta)\,,$$

with η a tuning parameter.

In figures 5.2, 5.3 and 5.4 we show traceplots and histograms of the iterates when $\eta = 10$, $\eta = 1$ and $\eta = 0.1$ respectively. Clearly, $\eta = 1$ is about right, whereas $\eta = 10$ and $\eta = 0.1$ propose too large and small steps respectively.

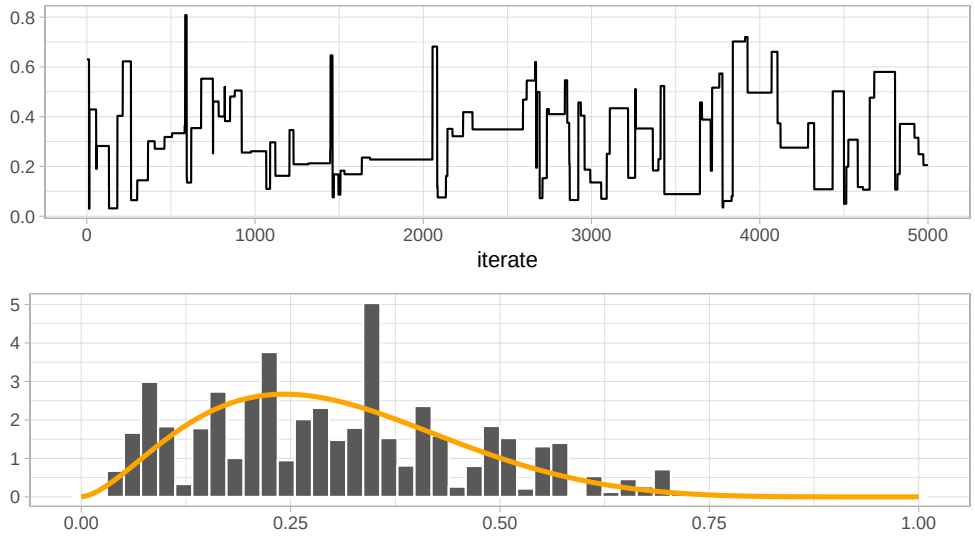

Figure 5.2: Output of the MH algorithm with random walk proposals with $\eta = 10$. Top: trace plot. Bottom: histogram. Average acceptance probability equals 0.023.

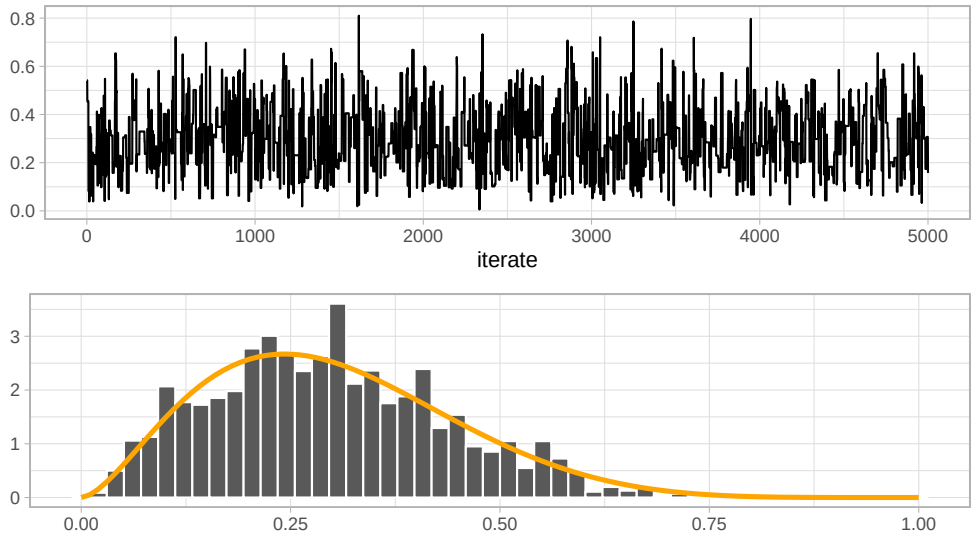

Figure 5.3: Output of the MH algorithm with random walk proposals with $\eta = 1$. Top: trace plot. Bottom: histogram. Average acceptance probability equals 0.224.

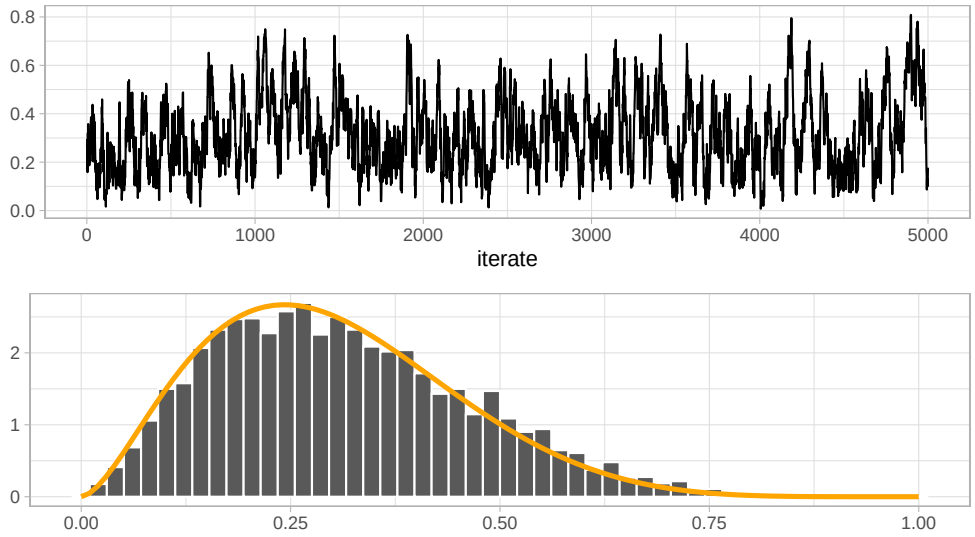

Figure 5.4: Output of the MH algorithm with random walk proposals with $\eta = 0.1$. Top: trace plot. Bottom: histogram. Average acceptance probability equals 0.844.

5.3 Cycles, mixtures and Gibbs sampling

In case the support of the target density is a subset of $\mathbb{R}^d$ and d is large, it may be difficult to find good proposal kernels. In that case, cycles and mixtures of MH-kernels can be utilised. The underlying idea is that some of the kernels may focus on particular subsets of $\text{supp}(\pi)$.

- **Cycle**: Suppose Q_1 and Q_2 are invariant for π, then so is

$$Q(x, \mathrm{d}y) = \int Q_1(x, \mathrm{d}z) Q_2(z, \mathrm{d}y).$$

  The extension to cycling with more than two kernels is obvious.

- **Mixture**: If each of the kernels $Q_i, i = 1, \ldots, p$ is invariant for π, then so is

$$Q(x, \mathrm{d}y) = \sum_{i=1}^{p} w_i Q_i(x, \mathrm{d}y).$$

  This means that with probability w_i we generate a proposal from kernel Q_i.

As an application of cycling kernels, we consider **One-at-a-time MH**. Suppose we wish to generate samples from $\pi(x)$ and that we split x into two parts: $x = (x_1, x_2)$. Assume we have

- a proposal density $Q_1(x_1, \mathrm{d}y_1 \mid x_2)$ for the 1st component (x_2 fixed);
- a proposal density $Q_2(x_2, \mathrm{d}y_2 \mid x_1)$ for the 2nd component (x_1 fixed).

For ease of exposition, we assume $Q_1(x_1, \mathrm{d}y_1 \mid x_2)$ and $\pi(\mathrm{d}x_1, x_2)$ are dominated by a common dominating measure and the corresponding densities are denoted by $q_1(x_1, y_1 \mid x_2)$ and $\pi(x_1, x_2)$ (and similarly for Q_2).

Suppose we have (x_1, x_2), then we evolve the chain by the following steps.

1. • draw $y_1 \sim q_1(x_1, y_1 \mid x_2)$
   • accept with probability

$$\alpha_1 = \min\left(1, \frac{\pi(y_1, x_2)}{\pi(x_1, x_2)} \frac{q_1(y_1, x_1 \mid x_2)}{q_1(x_1, y_1 \mid x_2)}\right),$$

   else set $y_1 = x_1$.

2. • draw $y_2 \sim q_2(x_2, y_2 \mid y_1)$
   • accept with probability

$$\alpha_2 = \min\left(1, \frac{\pi(y_1, y_2)}{\pi(y_1, x_2)} \frac{q_2(y_2, x_2 \mid y_1)}{q_2(x_2, y_2 \mid y_1)}\right),$$

   else set $y_2 = x_2$.

A special case is obtained by taking

$$q_1(x_1, y_1 \mid x_2) = \pi(y_1 \mid x_2)$$
$$q_2(x_2, y_2 \mid x_1) = \pi(y_2 \mid x_1)$$

Then

$$\alpha_1 = \min\left(1, \frac{\pi(y_1, x_2)}{\pi(x_1, x_2)} \frac{\pi(x_1 \mid x_2)}{\pi(y_1 \mid x_2)}\right) = \min\left(1, \frac{\pi(x_1 \mid x_2)}{\pi(x_1, x_2)} \frac{\pi(y_1, x_2)}{\pi(y_1 \mid x_2)}\right) = 1.$$

and similarly for α_2. Hence in this case the acceptance probabilities equal 1. This algorithm is known as the **Gibbs sampler**. It prescribes to iteratively sample from the "full conditionals".

Exercise 5.1. Suppose we wish to simulate from

$$\pi(x) = 0.7\phi(x; 2, 0.1) + 0.3\phi(x; 6, 1).$$

Note that this is a mixture density. Of course, there is a simple direct way to sampling from this density, but suppose we don't know this and wish to apply the MH-algorithm. Consider

- random walk MH, where $Z \sim Unif\,(-1, 1)$ and $\sigma = 1$;
- Langevin adjusted proposals, with $h = 1$.

1. Implement both methods and experiment how they perform on simulated data.
2. Repeat for the bivariate normal distribution, where the marginals are standard Normal and the correlation between the components is 0.9. For random walk MH, you can for example updates both components iteratively like in the Gibbs sampler.

5.4 Applying MCMC methods to the Baseball data

In the baseball example of Young and Smith [2005] we observe y_i which is the number of home-runs out of n_i times at bat ($i = 1, \ldots, n$, $n = 17$) from pre-season data (see also Example 1.30). Two models are proposed:

Model 1. Let

$$X_i = \sqrt{n_i} \arcsin\left(2\frac{Y_i}{n_i} - 1\right)$$

and assume the following hierarchical model:

$$\begin{aligned} X_i \mid \mu_i &\overset{\text{ind}}{\sim} N\left(\mu_i, 1\right) \\ \mu_i \mid \theta, \tau &\sim N\left(\theta, \tau^2\right) \\ p(\theta, \tau) &\propto \tau^{-1-2\alpha^*} e^{-\beta^*/\tau^2} \end{aligned}$$

Note that τ is a scaling parameter, and Jeffrey's prior is obtained when $\alpha^* \downarrow 0$ and $\beta^* \downarrow 0$.

Model 2. The hierarchical model:

$$\begin{aligned} Y_i \mid \mu_i &\sim Bin\left(n_i, p_i\right) \quad \text{with} \quad p_i = (1 + e^{-\mu_i})^{-1} \\ \mu_i \mid \theta, \tau &\sim N\left(\theta, \tau^2\right) \\ p(\theta, \tau) &\propto \tau^{-1-2\alpha^*} e^{-\beta^*/\tau^2} \end{aligned}$$

See also Young and Smith [2005], sections 3.4.1, 3.8 and 3.9.
Denote $\mu = (\mu_1, \ldots, \mu_n)$, $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_n)$.

5.4.1 MCMC algorithm for model 1

The Gibbs sampler can be used to sample from the posterior density $p(\mu, \theta, \tau^2 \mid X)$.

1. Initialise (θ, τ^2).
2. For $i = 1, \ldots, n$, update $\mu_i \mid \theta, \tau^2, X_i$: sample
$$\mu_i \sim N\left(\frac{\theta + \tau^2 X_i}{1 + \tau^2}, \frac{\tau^2}{1 + \tau^2}\right).$$
3. Update $\theta \mid \mu, \tau^2, X$: sample
$$\theta \sim N\left(\bar{\mu}_n, \tau^2/n\right).$$
4. Update $\tau^2 \mid \theta, X, \mu$: sample
$$\tau^{-2} \sim Ga\left(n/2 + \alpha^*, \beta^* + \frac{1}{2}\sum_{i=1}^{n}(\mu_i - \theta)^2\right).$$
5. Repeat steps 2,3 and 4 a large number of times.

Derivation of update for τ^{-2}: Set $v = \tau^{-2}$. The prior for v is given by

$$p(v) \propto v^{\alpha^*+1/2} e^{-\beta^* v} v^{-3/2} \propto v^{\alpha^*-1} e^{-\beta^* v}$$

(the term $v^{-3/2}$ comes from the Jacobian). Hence, the prior on v is $Ga(\alpha^*, \beta^*)$. Then (using Bayesian notation)

$$p(v \mid \mu, \theta, Y) \propto v^{n/2} \exp\left(-\frac{1}{2} v \sum_{i=1}^{n}(\mu_i - \theta)^2\right) p(v) \propto v^{\alpha^*+n/2-1} \exp\left(-v\left[\beta^* + \sum_{i=1}^{n}(\mu_i - \theta)^2\right]\right).$$

Therefore, conditional on (μ, θ, Y), v is Gamma distributed with shape parameter $\alpha^* + n/2$ and rate parameter $\beta^* + \sum_i (\mu_i - \theta)^2$.

> **Exercise 5.1** Check the update steps for $\{\mu_i\}$, θ.

5.4.2 MCMC algorithm for model 2

In model 2, steps 1, 3 and 4 are the same as for model 1. For step 2 (updating μ), note that

$$p(\mu_i \mid Y_i, \tau^2, \theta) \propto \frac{e^{Y_i \mu_i}}{(1+e^{\mu_i})^{n_i}} \exp\left(-\frac{(\mu_i-\theta)^2}{2\tau^2}\right).$$

As there is no direct way to simulate μ_i we propose to use a Metropolis-Hastings step here.

5.4.3 Some simulation results based on model 2

We implemented the MCMC-algoritm for model 2, taking $\alpha^\star = \beta^\star = 0.001$ as hyper-parameters. In the upper panel of Figure 5.5 we compare the maximum likelihood estimators for p_i with the mean of the predictive distribution. Note that for a measurable function g (we use Bayesian notation)

$$\begin{aligned}
\mathrm{E}[g(y_{new}) \mid y] &= \int g(y_{new}) p(y_{new} \mid y)\, \mathrm{d}y_{new} \\
&= \int g(y_{new}) \int p(y_{new}, \theta \mid y)\, \mathrm{d}\theta\, \mathrm{d}y_{new} \\
&= \int \left(\int g(y_{new}) p(y_{new} \mid \theta, y)\, \mathrm{d}y_{new} \right) p(\theta \mid y)\, \mathrm{d}\theta \\
&= \int \mathrm{E}[g(y_{new}) \mid \theta] p(\theta \mid y)\, \mathrm{d}\theta.
\end{aligned}$$

At the third equality we use Fubini's theorem; the final equality holds true provided that y_{new} and y are independent conditionally on θ. From this derivation its follows that the predictive mean in case of n_{new} trials equals $n_{new}\,\mathrm{E}[\theta \mid y]$ (take g the identity map).

In the lower panel of Figure 5.5 we compare the estimated number of home-runs based on pre-season data with the actual number of homeruns. Fro this, wee computed the sum of the squared deviations. For the mean of the predictive distribution and MLE these numbers are 2008 and 9051 respectively. This means that overall the Bayesian approach shows a large improvement.

Hence, overall, the estimates based on the posterior mean appear to be better than those based on maximum likelihood **when considering all players together**. Note in particular the difference in estimates for the players Sosa and Vaughn. Improving the estimators by learning from other players is often referred to as **borrowing strength** from others.

Exercise 5.2 Assume the following hierarchical model:

$$\begin{aligned}
X_i \mid \theta_i &\overset{\text{ind}}{\sim} Pois(\theta_i) \\
\theta_i \mid \beta &\overset{\text{ind}}{\sim} Exp(\beta) \\
\pi(\beta) &\propto 1
\end{aligned}$$

Here $i = 1, \ldots, n$. Derive the Gibbs-sampler for drawing from the posterior of $(\theta_1, \ldots, \theta_n, \beta)$.

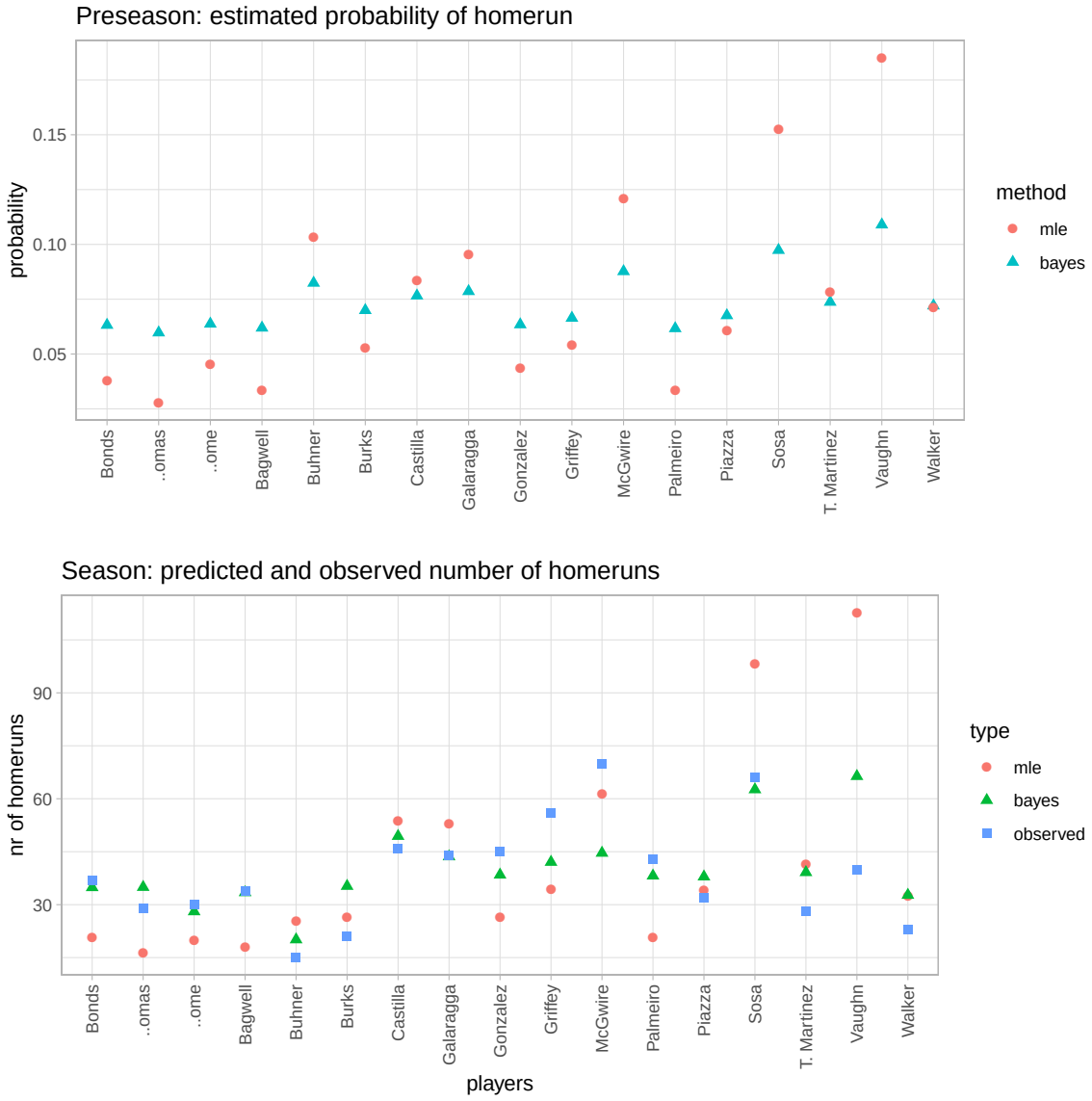

Figure 5.5: Comparison of maximum likelihood estimators and posterior mean estimators for p_i (based on model 2). Top: posterior mean and maximum likelihood estimator based on pre-season data. Bottom: estimated number of homeruns during the season based on both the mean of the predictive distribution and the maximum likelihood estimator. In blue the observed number of homeruns during the season have been added.

Exercise 5.3 [YS exercise 3.14.] Let $X_1, \ldots, X_n$ be IID $N(\mu, 1/\tau)$, conditional on (μ, τ). Suppose independent priors are assigned to μ and τ with $\mu \sim N\left(\xi, \kappa^{-1}\right)$ and $\tau Ga(\alpha, \beta)$.

1. Find the form of the joint posterior distribution of $(\mu, \tau) \mid X$, where $X = (X_1, \ldots, X_n)$. Note that this is not of standard form. Show that the conditional (posterior) distributions (also known as **full conditionals** are of simple forms:

$$\mu \mid \tau, X \sim N\left(\frac{\tau \sum_{i=1}^n X_i + \kappa\xi}{\tau n + \kappa}, \frac{1}{\tau n + \kappa}\right),$$

$$\tau \mid \mu, X \sim Ga\left(\alpha + n/2, \beta + (1/2)\sum_{i=1}^n (X_i - \mu)^2\right).$$

2. How can the derived distributions be used to devise an MCMC algorithm to draw from the posterior of (μ, τ)?

5.5 Applying Gibbs sampling for missing data problems*

Suppose we observed pairs (X_i, Y_i), $1 \leq i \leq n$, where $Y_i \in \{0, 1\}$. For new values x_*, we aim to predict the corresponding value of y_*. This is a **classification** problem. Using Bayesian notation, the following model is often used

$$p(x, y \mid \theta) = p(y \mid x, \theta)p(x). \tag{5.2}$$

Suppose now that the set of indices $\{1, \ldots, n\}$, can be partitioned into two (disjoint) sets I^{obs} and I^{mis}. If $i \in I^{obs}$, then we observe both x_i and y_i, whereas if $i \in I^{mis}$ only x_i is observed. Hence, for part of the data the label is unknown. Naturally, one can try to predict y_*, and estimate θ, using just the data with index in I^{obs}. However, in this way, we do not use the information contained in x data indexed by I^{mis}. Define

$$x^{mis} = \{x_i,\, i \in I^{mis}\} \qquad x^{obs} = \{x_i,\, i \in I^{obs}\} \qquad x = \{x_i,\, i \in \{1, \ldots, n\}\}$$

and similarly y^{mis}, y^{obs} and y.

We derive a Gibbs sampler to sample from $p(\theta, y^{mis} \mid y^{obs}, x)$. This turns out to be more natural than sampling from $p(\theta \mid y^{obs}, x)$. As can be seen from this, we augment the data with y^{mis} and for this reason this approach is often referred to as **data-augmentation**. Gibbs sampling boils down to iteratively sampling from

1. $p(\theta \mid y, x)$ and
2. $p(y^{mis} \mid \theta, x, y^{obs})$.

Step (1) is just as if we had no missing labels. In step (2) we sample y^{mis} conditional on the value of θ. The details for both steps are as follows:

1. Here we sample from

$$p(\theta \mid y, x) \propto p(y, x, \theta) = p(x, y \mid \theta)p(\theta) \propto p(y \mid x, \theta)p(\theta).$$

2. In this step we sample from

$$p(y^{mis} \mid \theta, x, y^{obs}) \propto p(y, x, \theta) \propto p(y \mid x, \theta).$$

If the y_i are independent conditionally on x, then this is proportional to $p(y^{mis} \mid x^{mis}, \theta)$.

Note that specifying the marginal distribution of x, $p(x)$, is not necessary due to the choice of model in (5.2).

Logistic regression corresponds to the particular choice

$$p(y \mid x, \theta) = \prod_{i=1}^{n} p(y_i \mid x_i, \theta) = \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-\theta x_i^T \theta}} \right)^{y_i} \left(1 - \frac{1}{1 + e^{-\theta x_i^T \theta}} \right)^{1-y_i}.$$

5.6 Variational Inference*

Variational Inference (VI) is a method to approximate the posterior that has become popular over the past two decades in the machine-learning community. Its popularity is due to the fact that it has been successfully applied in a host of applications, where MCMC is computationally too demanding. The basic idea is to approximate the posterior density $\pi(\theta) := f_{\Theta|X}(\theta \mid x)$ by some class of tractable densities $\mathcal{Q}$. Then the variational approximation to the posterior density is defined by

$$q^\star = \operatorname*{argmin}_{q \in \mathcal{Q}} KL(q, \pi). \tag{5.3}$$

This turns the inference problem into an optimisation problem. As the posterior is intractable, so is $KL(q, \pi)$, but we have the decomposition

$$KL(q, \pi) = \mathrm{E}_q \log \frac{q(\Theta)}{\pi(\Theta)} = \mathrm{E}_q \log q(\Theta) - \mathrm{E}_q \log f_{\Theta,X}(\Theta, x) + \log f_X(x).$$

Now if we define the **Evidence Lower BOund**, by (the dependence on x is suppressed in the notation)

$$ELBO(q) = \mathrm{E}_q \log f_{\Theta,X}(\Theta, x) - \mathrm{E}_q \log q(\Theta),$$

then

$$q^\star = \operatorname*{argmax}_{q \in \mathcal{Q}} ELBO(q).$$

The "trick" is now to take the class $\mathcal{Q}$ sufficiently large to get a good approximation, while still being able to compute $ELBO(q)$ for $q \in \mathcal{Q}$.

We can decompose the ELBO as follows

$$\begin{aligned} ELBO(q) &= \mathrm{E}_q \log f_\Theta(\Theta) + \mathrm{E}_q \log f_{X|\Theta}(x \mid \Theta) - \mathrm{E}_q \log q(\Theta) \\ &= \mathrm{E}_q \ell(\Theta \mid x) - KL(q, f_\Theta), \end{aligned}$$

where

$$\ell(\Theta \mid x) = \log f_{X|\Theta}(x \mid \Theta)$$

is the loglikelihood. So the VI-approximation balances two terms

- the first term encourages densities that place their mass on θs that explain the observed data;
- the second term encourages densities close to the prior.

So the variational objective mirrors the usual balance between likelihood and prior.

Probably the most popular class of approximating densities is the *mean field variational family*, where it is assumed that if $\theta \in \mathbb{R}^m$ then

$$q(\theta) = \prod_{j=1}^{m} q_j(\theta_j).$$

This simply means that the joint density factorises into its marginals. Even for the mean field family, deriving an expression for $ELBO(q)$ usually requires lengthy tricky calculations. For examples we refer to chapter 21 in Murphy [2012].

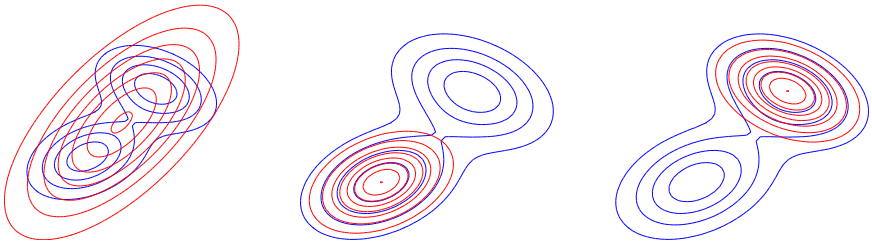

Figure 5.6: The blue curves are the contours of the true density π. The red curves are approximations. Left: using $KL(\pi, q)$ leads to a q that tries to cover both modes. Middle and right: using $KL(q, \pi)$ forces q to choose one of the modes. This is what is done using Variational Inference.

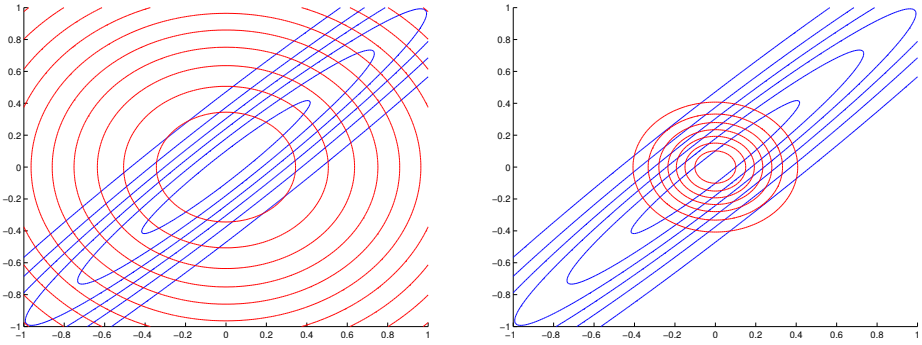

Figure 5.7: The blue curves are the contours of the true density π. The red curves are approximations. Left: using $KL(\pi, q)$ leads to a q that tries to cover the full support of π. Middle and right: using $KL(q, \pi)$ forces q to choose one of the modes. This is what is done using Variational Inference.

Exercise 5.2. Show that $\log f_X(x) \geq ELBO(q)$ Hence, $\log f_X(x)$, sometimes called the log-evidence in machine learning, is lower bounded by $ELBO(q)$. This explains the nomenclature.

Remark 5.10. Instead of $q^\star$ as in (5.3) an alternative is to consider

$$q^\circ = \operatorname*{argmin}_{q \in \mathcal{Q}} KL(\pi, q). \tag{5.4}$$

Note however that for intractable π (what we are dealing with), the expectation in $KL(\pi, q)$ is over π which is henceforth intractable. It is easy to see that $KL(q, \pi)$ is infinite if $\pi(\theta) = 0$ while $q(\theta) > 0$ (in a neighbourhood). We say that $KL(q, \pi)$ is zero-forcing for q and hence $q^\star$ *will typically underestimate the support of* π. Some terminology: $q^\star$ and q° are known as the information projection and reverse information projection of π on $\mathcal{Q}$.

To clearly see the difference between the two types of projection, consider Figures 5.6 and 5.7 which are taken from chapter 21 in Murphy [2012]. Here, the blue curves represent contour lines of the density of π, whereas the red-curves represent contour lines of the best approximation which in this case is taken to be a distribution with elliptic contours (i.e. a multivariate normal distribution).

5.7 Probabilistic programming languages

All that is needed for MCMC are the likelihood and prior (that is, the joint distribution of all variables). Probabilistic programming languages allow the user to obtain samples from the posterior by only specifying a generative hierarchical model. By additionally specifying which variables are observed, the language itself extract the joint distribution of all variables and runs a MCMC-algorithm to deliver samples from the posterior. Presently, *STAN* (named after Stanislaw Ulam) appears to be most popular though *Turing* is the Julia-language has recently also attracted more users. The workhorse algorithm is Hamiltonian Monte Carlo, the details of which are outside the scope of this course. A very important computational method in this algorithm (which in fact also is crucial for training of neural networks and in deep-learning) is **automatic differentiation**.

5.8 Expectation Maximisation (EM) algorithm*

The EM-algorithm is not really part of what is typically known as Bayesian computation, but it is closely related to data-augmentation. It is an iterative algorithm for numerically approximating the maximum likelihood estimator or posterior mode. Suppose we wish to find the maximiser of $\theta \mapsto \log f_X(x;\theta)$ (the loglikelihood), for data x. Suppose $f_X(x;\theta)$ is hard to evaluate, or intractable, but this is not the case for $f_{X,Z}(x,z;\theta)$ which admits f_X as marginal. Here, Z is a latent (unobserved) variable. The steps of the EM-algorithm are as follows:

1. Choose an initial value $\theta^{(0)}$ and set $i = 1$
2. Let
$$\theta^{(i)} = \operatorname*{argmax}_\theta Q(\theta, \theta^{(i-1)}), \tag{5.5}$$
where
$$Q(\theta,\eta) = \mathbb{E}_\eta[\log f_{X,Z}(x,Z;\theta) \mid X = x] = \int f_{Z|X}(z \mid x;\eta) \log f_{X,Z}(x,z;\theta)\,\mathrm{d}z.$$
3. Check for convergence of either the loglikelihood or parameter values. If the convergence criterion is not satisfied, increase i by 1 and return to step (2).

Commonly, (x,z) are referred to as the "full data". The second step requires to compute the expected full loglikelihood, under the distribution of $Z \mid X$ with the current iterate of the parameter.

There are various ways to motivate the EM-algorithm. Here, we follow the exposition in Särkkä [2013] (Section 12.2.3). Let q be a probability density. By Jensen's inequality

$$\begin{aligned}\log f_X(x;\theta) &= \log \int f_{X,Z}(x,z;\theta)\,\mathrm{d}z \\ &= \log \int q(z)\frac{f_{X,Z}(x,z;\theta)}{q(z)}\,\mathrm{d}z \geq \int q(z)\log\frac{f_{X,Z}(x,z;\theta)}{q(z)}\,\mathrm{d}z =: F(q,\theta).\end{aligned}$$

As direct maximisation of the left-hand-side is difficult, the EM-algorithm proceeds by iteratively maximising the lower bound on the right-hand-side. That is, it iterates the steps

1. find $q^{(i+1)} = \operatorname{argmax}_q F(q,\theta^{(i)})$;
2. find $\theta^{(i+1)} = \operatorname{argmax}_\theta F(q^{(i+1)},\theta)$.

The solution to the first optimisation problem is given by

$$q_{\theta^{(i)}}^{(i+1)}(z) = f_{Z|X}(z \mid x; \theta^{(i)}),$$

where we add the subscript to q to highlight its dependence on $\theta^{(i)}$ (and dropped dependence on x as the data are fixed). Next, note that subsequently the second step satisfies

$$\operatorname*{argmax}_{\theta} F(q^{(i+1)}, \theta) = \operatorname*{argmax}_{\theta} \int q_{\theta^{(i)}}^{(i+1)}(z) \log f_{X,Z}(x, z; \theta) \, \mathrm{d}z = \operatorname*{argmax}_{\theta} Q(\theta, \theta^{(i)})$$

which is exactly as in (5.5). A Monte-Carlo version of the EM-algorithm is obtained by approximating $Q(\theta, \eta)$ by $B^{-1} \sum_{b=1}^{B} \log f_{X,Z}(x, Z_b; \theta)$, where $Z_1, \ldots, Z_B \overset{\text{iid}}{\sim} f_{Z|X}(\cdot \mid x; \eta)$.

For a full discussion and conditions for the algorithm to converge to the MLE we refer to Section 9.4 in Bishop [2006] and Section 7.2 in Groeneboom and Jongbloed [2014].

Exercise 5.3. Verify that the maximum aposterior (MAP) estimator can also be approximated using the EM-algorithm, by changing (5.5) to $\theta^{(i)} = \operatorname{argmax}_{\theta} Q(\theta, \theta^{(i-1)}) + \log f_{\Theta}(\theta)$.

Chapter 6

Statistical decision theory

In this chapter we show that many statistical methods, such as point estimation and hypothesis testing, are part of a general framework provided by statistical decision theory. We discuss optimality of statistical decision rules and connect these to both classical and Bayesian statistics.

6.1 Introduction

Statistical decision theory is about making decisions under uncertainty. It provides a unifying framework to think about what makes a "good" estimator, test, or more generally a statistical procedure. within this framework one often speaks of "decision rules". Tests, confidence sets and estimators all turn out to be examples of decision rules. The conceptual framework is due to A. Wald (1939). You might wonder, "What's new?". Whereas "conventional statistics" is only directed towards the use of sampling information (data) in making inferences about θ, decision theory combines the sampling information with knowledge of the consequences of our decisions. Using a loss function and the data X, a decision rule maps the data X to an action a. Depending on the objective of the decision, actions can be "estimate the parameter θ using estimator $\hat{\Theta}$", "reject the hypothesis that $\theta \in \Omega_0 \subset \Omega$", or "change the treatment of all patients within one month".

Definition 6.1. Suppose X takes values in $\mathbb{R}^k$ and $X \sim P_\theta$, where $\theta \in \Omega$. Denote by $\mathbb{A}$ the set of allowable actions. Let $\mathcal{F}_\mathbb{A}$ be a σ-field on $\mathbb{A}$. The measurable space $(\mathbb{A}, \mathcal{F}_\mathbb{A})$ is called the **action space**. A **decision rule** is a measurable function $d : (\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k)) \to (\mathbb{A}, \mathcal{F}_\mathbb{A})$. If a decision rule d is chosen, we take action $d(X) \in \mathbb{A}$ when X is observed.

The construction or selection of decision rules cannot be done without any criterion about the preference of decision rules.

Definition 6.2. The **loss function** L is a measurable function $L : \Omega \times \mathbb{A} \to [0, \infty)$. If $X = x$ is observed then the loss for using decision rule d is given by $L(\theta, d(x))$ when the parameter value equals θ.

Definition 6.3. The **risk function** of the decision rule d is defined by

$$R(\theta, d) = \mathrm{E}_\theta\, L(\theta, d(X)) = \int L(\theta, d(x))\, \mathrm{dP}_\theta(x).$$

Example 6.4. In *estimation problems* $\mathbb{A} = \Omega$. A decision rule is usually called an estimator in this case. Common loss functions include:

- L_2-loss: $L(\theta, a) = \|\theta - a\|^2$, also known as the **Mean Squared Error**. If $\Omega \subset \mathbb{R}$ then

$$R(\theta, d) = \mathrm{E}_\theta(d(X) - \theta)^2 = \mathrm{Var}_\theta\, d(X) + \left(\mathrm{E}_\theta[\theta - d(X)]\right)^2.$$

  which is the bias-variance decomposition.
- L_1-loss: $L(\theta, a) = \|\theta - a\|$.
- Large deviation loss: choose $c > 0$ and define

$$L_c(\theta, a) = \mathbf{1}\{\|\theta - a\| > c\}.$$

Example 6.5. The following example is taken from Berger [1985] (chapter 1). Suppose a drug company has to decide whether or not to market a new pain reliever and suppose that there are two factors affecting the decision

1. η: proportion of people for which the drug will be effective;
2. θ: proportion of the market that the drug will capture.

For ease of exposition we only take θ into account. Sample information for θ can be obtained by interviewing people. Suppose the sample size equals n and let X denote the number of people that will buy the drug. If people decide independently on buying, then we may assume $X \sim Bin(n, \theta)$. The maximum likelihood estimator for θ is given by X/n.

Suppose now that overestimation is considered twice as costly as underestimation. Should we still use the MLE as estimator? Within a decision-theoretic perspective, we can deal with this problem by specifying an appropriate loss-function. In this case we can take

$$L(\theta, a) = \begin{cases} \theta - a & \text{if} \quad \theta > a \\ 2(a - \theta) & \text{if} \quad \theta \leq a \end{cases},$$

where $a \in [0, 1]$. Further optimality criteria for choosing a decision rule (estimator in this case) are required for deciding which rule to use.

Example 6.6. In the *hypothesis testing* problem

$$H_0 \ : \ \theta \in \Omega_0 \qquad H_1 \ : \ \theta \in \Omega_1,$$

a decision rule is called a test. The action space can be taken $\mathbb{A} = \{a_0, a_1\}$ with $a_0 = \{\text{accept } H_0\}$ and $a_1 = \{\text{accept } H_1\}$. We can make two errors: erroneously accepting H_0 and erroneously accepting H_1. In case we find these errors equally important, the following loss function is appropriate:

$$L(\theta, a_0) = \begin{cases} 0 & \text{if} \quad \theta \in \Omega_0 \\ 1 & \text{if} \quad \theta \in \Omega_1 \end{cases} \qquad L(\theta, a_1) = \begin{cases} 1 & \text{if} \quad \theta \in \Omega_0 \\ 0 & \text{if} \quad \theta \in \Omega_1 \end{cases}$$

This loss function is called **zero-one loss**. For a given decision rule d we have

$$R(\theta, d) = L(\theta, a_0)\mathrm{P}_\theta(d(X) = a_0) + L(\theta, a_1)\mathrm{P}_\theta(d(X) = a_1)$$

as d can only take 2 values. Hence

$$R(\theta, d) = \begin{cases} \mathrm{P}_\theta(d(X) = a_1) & \text{if} \quad \theta \in \Omega_0 \\ \mathrm{P}_\theta(d(X) = a_0) & \text{if} \quad \theta \in \Omega_1 \end{cases}.$$

These probabilities correspond to type I and type II errors in classical hypothesis testing. If we identify a_0 with 0 and a_1 with 1, then we might also have chosen $\mathbb{A} = \{0, 1\}$. In that case the decision rule is of the form $\mathbf{1}_C(X)$, the set C being the critical region of the test.

Within the framework of decision theory, statistical inference has an interpretation as a random game with two players: "nature" and the "statistician". In this game we fix

- a statistical model with sampling probabilities P_θ;
- an action space $\mathbb{A}$;
- a loss function L.

Then the game is played as follows:

1. nature selects (plays) $\theta_0 \in \Omega$;
2. the statistician observes $X \sim \mathrm{P}_{\theta_0}$ and plays $a \in \mathbb{A}$ in response, where $a = d(X)$ is determined by the statistician's decision rule;
3. the statistician has to pay nature $L(\theta, a)$.

We summarise the key components of statistical decision theory:

- Observed data $X \in \mathcal{X}$, where $\mathcal{X}$ is the sample space.
- A statistical model $\{\mathrm{P}_\theta,\ \theta \in \Omega\}$, where Ω is the parameter space.
- An action space $\mathbb{A}$: the set of all actions available to the experimenter.
- A loss function $L(\theta, a)$ defined on $\Omega \times \mathbb{A}$: the cost of taking action $a \in \mathbb{A}$ when the parameter is θ.

The goal is to find an "optimal" decision rule (in a sense to be made precise shortly), which is a mapping $d : \mathcal{X} \to \mathbb{A}$.

6.2 Comparing decision rules

Background reading material is in Young and Smith [2005], Sections 2.2 and 2.3.

Ideally, one would use a decision rule with minimal risk for all θ. Such a rule however hardly ever exists. Usually a particular decision rule d only has minimal risk for a subset Ω_1 of Ω whereas on Ω_1^c another rule behaves better. This is illustrated by the following example.

Example 6.7. Suppose $X \sim N(\theta, 1)$, $\theta \in \mathbb{R}$ and we wish to estimate θ. We consider the rules $d(X) = X$ and $d'(X) = 2$. Obviously, for most values of θ, d' is a silly choice, but in case $\theta = 2$ it is perfect. Under L_2-loss

$$R(\theta, d) = 1 \qquad R(\theta, d') = (2 - \theta)^2$$

from which we see that d' has strictly smaller risk than d if $\theta \in (1, 3)$.

Definition 6.8. A decision rule d is as good as another rule d' if

$$R(\theta, d) \leq R(\theta, d') \qquad \text{for all } \theta \in \Omega.$$

It is better if strict inequality holds for at least one θ. In that case we say that d' is dominated by d and call d' **inadmissible**. A decision rule that is not inadmissible is called **admissible**.

Strictly speaking, one should speak of admissibility with respect to a given loss function. Admissibility rules out certain decision rules (that henceforth should not be used), but is a very weak condition. In example 6.7 for example d' is admissible though few statisticians would consider this to be a good choice. Maximum likelihood estimators need not be admissible, as the following example shows.

Example 6.9. Suppose $X_1, \ldots, X_n \mid \Theta = \theta \overset{\text{ind}}{\sim} Exp(\theta)$. The maximum likelihood estimator is given by $d(X) = 1/\bar{X}_n$. For L_2-loss it is inadmissible. To see why, we first note that $\sum_{i=1}^n X_i \sim Ga(n, \theta)$ which implies

$$\mathrm{E}_\theta\, d(X) = n \int_0^\infty \frac{1}{z} \frac{\theta^n}{\Gamma(n)} z^{n-1} e^{-\theta z}\, \mathrm{d}z = \frac{n}{n-1}\theta.$$

Therefore, the decision rule $d'(X) = \frac{n-1}{n} d(X)$ is unbiased for estimating θ. Since

$$R(\theta, d') = \operatorname{Var} d'(X) = \left(\frac{n-1}{n}\right)^2 \operatorname{Var} d(X) < \operatorname{Var} d(X) \leq R(\theta, d).$$

d' improves upon d.

Another famous example of an inadmissible estimator is given in the next subsection.

6.2.1 Shrinkage and James-Stein estimation

Stein [1956] considered the model

$$X_i \overset{\text{ind}}{\sim} N\left(\theta_i, 1\right), \quad \text{for} \quad i = 1, \ldots, p.$$

It is a trivial exercise to derive that the maximum likelihood estimator equals $\hat{\Theta} = X = (X_1, \ldots, X_p)$. Under quadratic loss, i.e. $L(\theta, a) = \|\theta - a\|^2$, we then have

$$\mathbb{E}\left[\sum_{i=1}^p (\hat{\Theta}_i - \theta_i)^2\right] = \sum_{i=1}^p \mathbb{E}(X_i - \theta_i)^2 = p.$$

Theorem 6.10. Define the **James-Stein estimator** by

$$\hat{\Theta}^{(JS)} = \left(1 - \frac{\alpha}{\|X\|^2}\right) X.$$

If $0 < \alpha < 2(p-2)$, then $\hat{\Theta}^{(JS)}$ strictly dominates $\hat{\Theta}$ for the loss function $L(\theta, a) = \|\theta - a\|^2$.

Proof. If we let $\theta = (\theta_1, \ldots, \theta_n)$, then

$$\|\theta - \hat{\Theta}^{(JS)}\|^2 = \left\|\theta - X + \frac{\alpha}{\|X\|^2} X\right\|^2 = \|\theta - X\|^2 + \frac{2\alpha}{\|X\|^2} X^T(\theta - X) + \frac{\alpha^2}{\|X\|^2}.$$

We take expectations on both sides of this equation. Define for each i the function $h_i : \mathbb{R}^p \to \mathbb{R}$ by

$$h_i(x) = \frac{x_i}{\sum_{j=1}^p x_j^2}.$$

We apply lemma 6.11 to obtain

$$-\mathrm{E}\left[\frac{X^T(\theta - X)}{\|X\|^2}\right] = \sum_{i=1}^p \mathrm{E}\left[(X_i - \theta_i) h_i(X)\right] = \sum_{i=1}^p \mathrm{E}\left[\frac{\partial h_i(X)}{\partial X_i}\right] = (p-2)\,\mathrm{E}\left[\frac{1}{\|X\|^2}\right]$$

This implies

$$\mathrm{E}\,\|\theta - \hat{\Theta}^{(JS)}\|^2 = p - \underbrace{\left(2\alpha(p-2) - \alpha^2\right)}\,\mathrm{E}\left[\frac{1}{\|X\|^2}\right].$$

The term in braces is strictly positive when $0 < \alpha < 2(p-2)$. □

In the proof we used **Stein's lemma**.

Lemma 6.11 (One dimensional version)**.** Suppose $X \sim N\left(\mu, \sigma^2\right)$. If $h : \mathbb{R} \to \mathbb{R}$ is absolutely continuous and $\mathrm{E}\,|h'(X)| < \infty$, then

$$\mathrm{E}\,[(X-\mu)h(X)] = \sigma^2\,\mathrm{E}\,h'(X).$$

Proof. First assume $\mu =$ and $\sigma = 1$. Without loss of generality we can assume $h(0) = 0$. By Fubini's theorem:

$$\begin{aligned}\int_0^\infty xh(x)e^{-x^2/2}\,\mathrm{d}x &= \int_0^\infty \left[\int_0^x h'(y)\,\mathrm{d}y\right] xe^{-x^2/2}\,\mathrm{d}x \\ &= \int_0^\infty h'(y)\left[\int_y^\infty xe^{-x^2/2}dx\right]\mathrm{d}y = \int_0^\infty h'(y)e^{-y^2/2}\,\mathrm{d}y.\end{aligned}$$

A similar calculation shows that the resulting equality also holds when integrating over $(-\infty, 0)$. Adding $\int_0^\infty xh(x)e^{-x^2/2}\,\mathrm{d}x$ and $\int_{-\infty}^0 xh(x)e^{-x^2/2}\,\mathrm{d}x$ and dividing by $\sqrt{2\pi}$ we get $\mathrm{E}\,Zh(Z) = \mathrm{E}\,h'(Z)$ if $Z \sim N(0,1)$. For the general case: let $X = \mu + \sigma Z$. Then

$$\mathrm{E}(X-\mu)h(X) = \sigma\,\mathrm{E}\,Zh(\mu+\sigma Z) = \sigma^2\,\mathrm{E}\,h'(\mu+\sigma Z) = \sigma^2\,\mathrm{E}\,h'(X).$$

Here we used the result for $\mu = 0$ and $\sigma = 1$ at the second equality sign. □

The JS-estimator has an empirical Bayes interpretation. For that, consider the model

$$\begin{aligned}X_i \mid \Theta_i = \theta_i &\overset{\text{ind}}{\sim} N\left(\theta_i, 1\right) \\ \Theta_1, \ldots, \Theta_p &\overset{\text{ind}}{\sim} N\left(0, A\right)\end{aligned}$$

Define $\Theta = (\Theta_1, \ldots, \Theta_p)$. The Bayes estimator under L_2-loss is given by

$$\Theta^{\text{Bayes}} = \mathbb{E}\left[\Theta \mid X\right] = \left(1 - \frac{1}{A+1}\right) X.$$

Replacing $1/(A+1)$ with an unbiased estimator turns out to give precisely the James-Stein estimator.

Lemma 6.12. If $X \sim N\left(0, (A+1)I_p\right)$, then if $p \geq 3$,

$$\mathbb{E}\left[\frac{p-2}{\|X\|^2}\right] = \frac{1}{A+1}.$$

Proof. It is a well known that $Y := \|X\|^2/(A+1) \sim \chi^2_p$. To prove the result, rewrite

$$\frac{p-2}{\|X\|^2} = \frac{1}{A+1}\frac{p-2}{Y}$$

Hence, it suffices to prove that $(p-2)\mathbb{E}[1/Y] = 1$ which follows from algebra. □

The way is which A is estimated is not prescribed by the empirical Bayes method. So we can alternatively use the method of maximum likelihood to estimate A. It is not too hard to establish that $X_1, \ldots, X_n$ are independent with the $N(0, 1 + A)$ distribution. An estimator for A can henceforth be obtained as maximiser of

$$A \mapsto (2\pi(1 + A))^{-p/2} \exp\left(-\frac{\|X\|^2}{2(1 + A)}\right).$$

over $A \in [0, \infty)$. This problem has solution

$$\hat{A} = \begin{cases} \|X\|^2/p - 1 & \text{if } \|X\|^2 > p \\ 0 & \text{else} \end{cases}.$$

This implies an empirical Bayes estimator is given by

$$\hat{\Theta}_{EB} = \frac{\hat{A}}{1 + \hat{A}} X = \left(1 - \frac{p}{\|X\|^2}\right)_+ X.$$

Here $(x)_+ = \max(x, 0)$. This estimator is known as the **truncated James-Stein estimator**.

Although the JS-estimator strictly dominates the MLE it is inadmissible itself. The use of seemingly unrelated data in forming estimators is sometimes referred to as **borrowing strength from others**. Bayesian hierarchical modelling incorporates the borrowing strength idea automatically by introducing dependence via additional layers in the prior specification. If the sole purpose is estimating one specific θ_i, then X_i is admissible (and minimax, a concept that we will define in the next section).

6.3 Minimax and Bayes decision rules

The risk function provides a ranking of performance between decision functions for every fixed θ. If we consider d to be better than d' if $R(\theta, d) < R(\theta, d')$ for all θ, then most decision rules are not comparable. See for instance example 6.7. As a solution, two principles for comparing decision rules have been proposed: the minimax- and Bayes risk principle.

Definition 6.13. **Minimax principle**: in comparing decision rules, the rule d with the smallest value of $\sup_{\theta \in \Omega} R(\theta, d)$ is best. A decision rule d is called **minimax** if for all d'

$$\sup_{\theta \in \Omega} R(\theta, d) \leq \sup_{\theta \in \Omega} R(\theta, d').$$

Alternatively, d is minimax if

$$\sup_{\theta \in \Omega} R(\theta, d) = \inf_{d'} \sup_{\theta \in \Omega} R(\theta, d').$$

This criterion chooses the decision rule which behaves best in the worst-case scenario.

The second notion of optimality is called Bayes optimality. Here we assign weights to different values of θ by means of a prior density f_Θ. The Bayes risk is defined as a weighted average of the risk with respect to this prior.

Definition 6.14. **Bayes risk principle**: in comparing decision rules, choose the rule d with the smallest value of

$$r(\mu_\Theta, d) = \int R(\theta, d) \, \mathrm{d}\mu_\Theta(\theta),$$

where μ_Θ is a probability measure on Ω. The number $r(\mu_\Theta, d)$ is known as the **Bayes risk** of d with respect to μ_Θ. Any decision rule that minimises the Bayes risk is called a **Bayes decision rule**.

Both $\sup_{\theta\in\Omega} R(\theta, d)$ and $r(\mu_\Theta, d)$ are nonnegative real numbers that depend on the chosen decision rule, but not on x (as the risk function is defined by taking an expectation over X under P_θ).

Definition 6.15. The **posterior risk** of d using prior measure μ_Θ is defined as

$$r_x(\mu_\Theta, d) = \int L(\theta, d(x))\Pi_x(\,\mathrm{d}\theta).$$

Note that the posterior risk depends on x. Whereas the risk averages the loss over X (and hence depends on θ), the posterior risk averages the loss over the posterior of Θ (and hence depends on x).

Lemma 6.16. The decision rule d is Bayes with respect to the prior μ_Θ if and only if it minimises the posterior risk.

Proof. This follows from

$$\begin{aligned} r(\mu_\Theta, d) &= \int R(\theta, d)\mu_\Theta(\,\mathrm{d}\theta) = \int \left(\int L(\theta, d)\,\mathrm{dP}_\theta(x) \right) \mu_\Theta(\,\mathrm{d}\theta) \\ &= \int \left(\int L(\theta, d)\Pi_x(\,\mathrm{d}\theta) \right) \mu_X(\,\mathrm{d}x) = \int r_x(\mu_\Theta, d)\mu_X(\,\mathrm{d}x). \end{aligned}$$

We can interchange integrals as the integrand is nonnegative. □

Remark 6.17. In classical statistics, an estimator is considered good if it is close to the true value on average or in the long-run (**pre-trial**): the loss is averaged by taking an expectation over X. In Bayesian statistics the loss is averaged by taking an expectation over $\Theta \mid X$. This is an expectation after seeing the data and is sometimes called **post-trial**.

From here, we can perhaps most clearly see how fundamentally different classical statistics and Bayesian statistics are (as opposed to how Bayesian methods are incorrectly introduced by saying that it is like adding a prior to the framework of classical statistics). Classical statistics bases inference on

$$\int_{\mathcal{X}} L(d(x), \theta)\mathrm{P}_\theta(\,\mathrm{d}x),$$

while Bayesian statistics bases inference on

$$\int_\Omega L(d(x), \theta)\Pi_x(\,\mathrm{d}\theta).$$

Hence, in classical statistics a sample space $\mathcal{X}$ is required and inference depends on it. Now the Bayesian perspective does neither require a sample space nor a hypothetical population. "This analysis does not use the sample space except insofar as $\mathcal{X}$ may affect the likelihood $p(x \mid \theta)$ regarded as a function of θ for fixed x. Once x is observed and the likelihood is available, $\mathcal{X}$ is irrelevant." (Lindley [1990], page 46). This is of course in full agreement with the likelihood principle.

Exercise 6.1 [YS exercise 2.1.] Let $X \sim Unif\,(0, \theta)$, where $\theta > 0$ is unknown. Let the action space be $[0, \infty)$ and the loss function $L(\theta, d) = (\theta - d)^2$, where d is the action chosen. Consider the decision rules

$$d_\mu(x) = \mu x, \qquad \mu \geq 0.$$

For what value of μ is d_μ unbiased? Show that $\mu = 3/2$ is a necessary condition for d_μ to be admissible.

Exercise 6.2 The risks for five decision rules $\delta_1, \ldots, \delta_5$ depend on the value of a positive-valued parameter θ. The risks are given in the table below

	δ_1	δ_2	δ_3	δ_4	δ_5
$0 \leq \theta < 1$	10	10	7	6	8
$1 \leq \theta < 2$	8	11	8	5	10
$2 \leq \theta$	15	11	12	14	14

1. Which decision rules are at least as good as δ_1 for all θ?
2. Which of the five stated decision rules are admissible?
3. Which of the five stated decision rules is minimax?
4. Suppose θ has a uniform distribution on $[0, 5]$. Which is the Bayes rule and what is the Bayes risk for that rule?

Exercise 6.3 [YS exercise 3.5.] At a critical stage in the development of a new aeroplane, a decision must be taken to continue or to abandon the project. The financial viability of the project can be measured by a parameter $\theta \in (0, 1)$, the project being profitable if $\theta > 1/2$. Data x provide information about θ. We assume x to be a realisation of $X \sim \mathrm{P}_\theta$.

- If $\theta < 1/2$, the cost to the taxpayer of continuing the project is $1/2 - \theta$ (in units of billion dollars), whereas if $\theta > 1/2$ it is zero (since the project will be privatised if profitable).
- If $\theta > 1/2$ the cost of abandoning the project is $\theta - 1/2$ (due to contractual arrangements for purchasing the aeroplane from the French), whereas if $\theta < 1/2$ it is zero.

1. Derive an expression for the posterior risk.
2. Derive the Bayes decision rule in terms of the posterior mean of θ by choosing the decision rule that minimises the posterior risk.
3. The Minister of Aviation has prior density $6\theta(1-\theta)$ for θ. The Prime Minister has prior density $4\theta^3$. The prototype aeroplane is subjected to trials, each independently having probability θ of success, and the data x consist of the total number of trials required for the first successful result to be obtained. For what values of x will there be serious ministerial disagreement?

Exercise 6.4 Suppose we have a single observation, X, which comes from a distribution with density function f_θ, with $\theta \in \{0, 1\}$ and we want to test

$$H_0 : f(x) = f_0(x) = 2(1-x)\mathbf{1}_{[0,1]}(x)$$

against

$$H_1 : f(x) = f_1(x) = 2x\mathbf{1}_{[0,1]}(x)$$

1. Using Neyman-Pearson, show that the best critical region for the likelihood ratio test of H_0 versus H_1 is given by $X \geq B$ for some constant B.
2. Consider now choosing B using decision theory. Suppose the losses incurred by a type II error is four times the loss of a type I error. Consider decision rules d_B which choose H_1 if $X \geq B$.
   (a) Write down the loss function when considering the action space is $\mathcal{A} = \{a_0, a_1\}$ with $a_0 = \{\text{accept } H_0\}$ and $a_1 = \{\text{accept } H_1\}$.
   (b) Calculate the risks $R(0, d_B)$ and $R(1, d_B)$ as functions of B. Use this to find the value of B which gives the minimax rule.
   (c) Calculate the Bayes risk, when the prior probabilities are $1/4$ and $3/4$ for H_0 and H_1 respectively, and find the value of B which gives the Bayes rule.

Having introduced admissibility, minimax rules and Bayes rules it is interesting to see how these concepts relate. This is a broad field from which we can only discuss a few main results. Typical questions that are part of statistical decision theory include:

1. Are minimax rules always admissible? (no)
2. Are Bayes rules always admissible? (usually)
3. Are all admissible rules Bayes for some prior? (complete class theorem)
4. Are Bayes rules minimax? (extended Bayes rules which are equaliser rules are minimax)
5. Are minimax rules Bayes for some prior π? (requires existence of a least favourable prior).

In brackets we have given hints or partial answers to these questions (some concepts appearing in these answers will be discussed as we proceed along this chapter). One relation is easy: the Bayes risk is always smaller than (or equal) the maximum risk:

$$r(\mu_\Theta, d) = \int R(\theta, d)\mu_\Theta(\,\mathrm{d}\theta) \leq \sup_\theta R(\theta, d).$$

In section 6.8 we will investigate the role of sufficient statistics to the list of posed questions.

Related to question 3 we introduce the following definition.

Definition 6.18.

1. A class of decision rules $\mathcal{D}$ is **complete** if for any decision rule $d \notin \mathcal{D}$ there exists a decision rule $d' \in \mathcal{D}$ that dominates d.

2. A class of decision rules $\mathcal{D}$ is **essentially complete** if for any decision rule $d \notin \mathcal{D}$ there exists a decision rule $d' \in \mathcal{D}$ such that $R(\theta, d') \leq R(\theta, d)$ for all θ.

3. A class of decision rules $\mathcal{D}$ is said to be **minimal (essentially) complete** if $\mathcal{D}$ is complete and if no proper subset of $\mathcal{D}$ is (essentially) complete.

Clearly, any decision rule that is outside a complete class is inadmissible.

6.4 Admissibility of Bayes rules

Bayes rules are usually admissible. We present a few results.

Theorem 6.19. Assume Ω is finite. If the prior measure μ_Θ satisfies $\mu_\Theta(\{\theta\}) > 0$ for all $\theta \in \Omega$, then a Bayes rule with respect to μ_Θ is admissible.

Exercise 6.5 Prove theorem 6.19.

Theorem 6.20. If a Bayes rule is unique, then it is admissible.

Theorem 6.21. Suppose Ω is a subset of the real line. Assume that the risk functions $R(\theta, d)$ are continuous in θ for all decision rules d. Suppose that for any $\varepsilon > 0$ and any θ the interval $(\theta - \varepsilon, \theta + \varepsilon)$ has positive probability under the prior μ_Θ. Then a Bayes rule with respect to μ_Θ is admissible.

Proof. Suppose that $\bar{d}$ is a Bayes rule with respect to μ_Θ, but not admissible. Then there exists another decision rule d' such that

$$\begin{cases} R(\theta, d') \leq R(\theta, \bar{d}), & \forall \theta \in \Omega, \\ R(\theta_0, d') < R(\theta_0, \bar{d}), & \text{for some } \theta_0 \in \Omega. \end{cases}$$

From the continuity of $R(\theta, d)$, it follows that there exists $\epsilon > 0$ such that $R(\theta, d') < R(\theta, \bar{d})$, for all $\theta \in I := (\theta_0 - \epsilon, \theta_0 + \epsilon)$. Hence

$$\int_I R(\theta, d')\mu_\Theta(\,\mathrm{d}\theta) < \int_I R(\theta, \bar{d})\mu_\Theta(\,\mathrm{d}\theta).$$

Moreover, it also holds that

$$\int_{\Omega \setminus I} R(\theta, d')\mu_\Theta(\,\mathrm{d}\theta) \leq \int_{\Omega \setminus I} R(\theta, \bar{d})\mu_\Theta(\,\mathrm{d}\theta),$$

As a result,

$$r(\mu_\Theta, d') = \int_\Omega R(\theta, d')\mu_\Theta(\,\mathrm{d}\theta) < \int_\Omega R(\theta, \bar{d})\mu_\Theta(\,\mathrm{d}\theta) = r(\mu_\Theta, \bar{d}),$$

which contradicts that $\bar{d}$ is a Bayes rule. □

In case the loss function is strictly convex, admissibility of Bayes rules follows from the following theorem.

Theorem 6.22. Assume the action space $\mathbb{A}$ is a convex subset of $\mathbb{R}^m$and that all P_θ are absolutely continuous with respect to each other. If $L(\theta, \cdot)$ is strictly convex for all θ, then for any probability measure Π on Ω, the Bayes rule d_Π is admissible.

Proof. Suppose d_Π is not admissible. Then there exists d_0 such that $R(\theta, d_0) \leq R(\theta, d_\Pi)$ with strict inequality for some θ. Define a new decision rule

$$d_1(x) = \frac{d_\Pi(x) + d_0(x)}{2},$$

then $d_1 \in \mathbb{A}$ by convexity. For all θ we have

$$\begin{aligned} R(\theta, d_1) &= \int L(\theta, (d_\Pi(x) + d_0(x))/2) \mathrm{P}_\theta(\,\mathrm{d}x) \\ &\leq \int \frac{1}{2}\left[L(\theta, d_\Pi(x)) + L(\theta, d_0(x))\right] \mathrm{P}_\theta(\,\mathrm{d}x) \\ &= \frac{1}{2}\left[R(\theta, d_\Pi) + R(\theta, d_0)\right] \leq R(\theta, d_\Pi). \end{aligned}$$

For a fixed value of θ, the first inequality will be strict unless $\mathrm{P}_\theta(d_\Pi(X) = d_0(X)) = 1$. Since all P_θ are absolutely continuous with respect ot each other, it follows that $\mathrm{P}_\theta(d_\Pi(X) = d_0(X)) = 1$ for one θ if and only if $\mathrm{P}_\theta(d_\Pi(X) = d_0(X)) = 1$ for all θ.

Therefore, the first inequality will be strict unless $\mathrm{P}_\theta(d_\Pi(X) = d_0(X)) = 1$ for all θ. The latter condition is equivalent to $d_\Pi(X)$ and $d_0(X)$ having the same distribution. This cannot be the case as this would violate the supposition that d_1 dominates d_Π. Hence, we have $R(\theta, d_1) < R(\theta, d_\pi)$ for all θ.

By integrating over the prior we get $r(\Pi, d_1) < r(\Pi, d_\Pi)$ which gives the desired contradiction. Hence d_Π is admissible. □

6.5 Bayes rules in various settings

In this section we derive Bayes rules for point estimation, interval estimation and hypothesis testing.

6.5.1 Bayesian point estimation

Suppose $\Omega \subset \mathbb{R}^k$ and L_2 loss function. The posterior risk equals

$$r_x(\mu_\Theta, d) = \int \|\theta - d\|^2 \Pi_x(\,\mathrm{d}\theta)$$

and the Bayes rule minimises this expression with respect to d. Taking partial derivatives with respect to each element d_i of the vector d and equating to zero, we easily derive that the Bayes rule is given by $d = \int \theta \Pi_x(\,\mathrm{d}\theta)$ which is the posterior mean.

If $\Omega \subset \mathbb{R}$ and we take

$$L(\theta, a) = \begin{cases} c_1(\theta - a) & \text{if } a \leq \theta \\ c_2(a - \theta) & \text{if } a > \theta \end{cases}. \tag{6.1}$$

then it follows that if a minimises the posterior risk then it satisfies the equation

$$\Pi_x((-\infty, a)) = \frac{c_1}{c_1 + c_2} \tag{6.2}$$

Hence, the Bayes rule for this loss is the $c_1/(c_1+c_2)$-quantile of the posterior. In particular, taking $c_1 = c_2 = 1$ the resulting point estimator is the **posterior median**.

Exercise 6.6 Prove (6.2). *Hint: consider the posterior risk and differentiate the latter with respect to a.*

Finally, consider large deviation loss, which is defined by $L_c(\theta, a) = \mathbf{1}\{\|\theta - a\| > c\}$ for a fixed $c > 0$. The Bayes rule is the value of a that has the largest posterior mass in a ball with radius c. If the prior is approximately flat, then this a will be close to the value maximising the likelihood. For this loss function minimising the posterior risk is equivalent to maximising

$$d \mapsto \int_{\theta:\ \|\theta - d\| \leq c} f_{\Theta|X}(\theta \mid x)\, \mathrm{d}\theta.$$

Upon letting $c \downarrow 0$ the Bayes rule equals that value of θ for which $f_{\Theta|X}(\theta \mid x)$ is maximal (if such a value exists). When it exists, we call $\operatorname{argmax}_{\theta \in \Theta} f_{\Theta|X}(\theta \mid x)$ the **posterior mode**.

Example 6.23. This is a continuation of example 6.5. Suppose we decide to take a $Unif\,(0,1)$-prior on Θ. Then $\Theta \mid X = x \sim Be\,(x+1, n-x+1)$. The loss-function that takes into account that overestimation is twice as costly as underestimation is of the form 6.1 with $c_1 = 1$ and $c_2 = 2$. The Bayes rule is henceforth the $1/3$-th quantile of the posterior.

6.5.2 Bayesian interval estimation*

It is common practise in statistical analysis to report interval estimates. These are meant to give a range of plausible values of the unknown quantities and thereby quantify uncertainty. Interval estimates should trade off two competing goals: intervals should be small and close to the true value. As an example: suppose $\Omega \subseteq \mathbb{R}$ is the parameter set and we wish to choose an interval (a_1, a_2) within that space. The half-width of this interval is given by $(a_2 - a_1)/2$. One way to quantify the distance by which the parameter θ is outside the interval is given by

$$(a_1 - \theta)_+ + (\theta - a_2)_+.$$

Here $x_+ = \max(x, 0)$. As a loss-function we can therefore take a weighted combination:

$$L(\theta, a) = L_1 \frac{a_2 - a_1}{2} + L_2 \left[(a_1 - \theta)_+ + (\theta - a_2)_+\right].$$

It can be shown that the optimal values for a_1 and a_2 are the $L_1/(2L_2)$ and $1 - L_1/(2L_2)$ quantiles of the posterior distribution of θ respectively. This provides a decision-theoretic justification for the common practise of computing equal tail posterior probability regions. Taking $L_1 = \alpha L_2$ (with $\alpha \in (0,1)$, typically being taken equal to 0.05), we see that α_1 and a_2 are the $\alpha/2$ and $1 - \alpha/2$ posterior quantiles respectively. Clearly, from a decision-theoretic point of view the penalty on the width of the interval is less severe than that on missing θ.

6.5.3 Bayesian hypothesis testing and Bayes factors

The hypothesis testing problem was already quickly discussed on page 114 and the essential ideas are already contained in exercise 6.4. Here, we consider the case where $\Omega = \Omega_0 \cup \Omega_1$ with $\Omega_1 \cap \Omega_2 = \emptyset$ and the action space is given by $\{a_0, a_1\}$ (with a_i corresponding to accepting the hypothesis $H_i : \theta \in \Omega_i$). Assume the following simple loss function

	$\theta \in \Omega_0$	$\theta \in \Omega_1$
a_0	0	L_0
a_1	L_1	0

The posterior risk is given by

$$r_x(\mu_\Theta, a) = \int_\Omega L(\theta, a)\Pi_x(\,\mathrm{d}\theta) = \begin{cases} L_0\Pi_x(\Omega_1) & \text{if} \quad a = a_0 \\ L_1\Pi_x(\Omega_0) & \text{if} \quad a = a_1 \end{cases}.$$

The Bayes action is a_0 if $L_0\Pi_x(\Omega_1) < L_1\Pi_x(\Omega_0)$. That is, we take action a_0 if

$$\frac{L_0}{L_1} < \frac{\Pi_x(\Omega_0)}{\Pi_x(\Omega_1)}. \tag{6.3}$$

For hypothesis testing, assume the prior μ_Θ is of the form

$$\mu_\Theta(A) = \pi_0\mu_\Theta^{(0)}(A \cap \Omega_0) + \pi_1\mu_\Theta^{(1)}(A \cap \Omega_1)$$

Here $\pi_0 + \pi_1 = 1$ and π_0 is the prior probability that hypothesis H_0 is true. Furthermore, $\mu_\Theta^{(0)}$ and $\mu_\Theta^{(1)}$ are are prior (probability) measures supported on Ω_0 and Ω_1 respectively, implying that μ_Θ is a probability measure on Ω. If we assume $\mathrm{P}_\theta \ll \nu$ and denote $L(\theta; x) = \frac{\mathrm{dP}_\theta}{\mathrm{d}\nu}(x)$, then

$$\frac{\Pi_x(\Omega_0)}{\Pi_x(\Omega_1)} = \frac{\int_{\Omega_0} L(\theta; x)\mu_\Theta^{(0)}(\,\mathrm{d}\theta)}{\int_{\Omega_1} L(\theta; x)\mu_\Theta^{(1)}(\,\mathrm{d}\theta)} = \frac{\pi_0}{\pi_1} \times B_{0,1}(x) \tag{6.4}$$

with the **Bayes factor** $B_{0,1}(x)$ defined by

$$B_{0,1}(x) = \frac{\int_{\Omega_0} L(\theta; x)\mu_\Theta^{(0)}(\,\mathrm{d}\theta)}{\int_{\Omega_1} L(\theta; x)\mu_\Theta^{(1)}(\,\mathrm{d}\theta)}.$$

Equation (6.4) if often summarised by

$$\boxed{\text{posterior odds} = \text{prior odds} \times \text{Bayes factor}}.$$

Jeffreys advocated the use of the Bayes factor as a direct and intuitive measure of evidence to be used in alternative to, say, p-values, for evidence against a hypothesis. A good discussion on Bayes factors can be found in section 4.4 of Young and Smith [2005]. Note that the formulation of the prior in terms of *measures* rather than *densities* easily incorporates point-null hypothesis testing: if $\Omega_0 = \{\theta_0\}$ then $\mu_\Theta^{(0)} = \delta_{\theta_0}$ (Dirac mass at θ_0). Equation (6.4) then implies

$$\Pi_x(\{\theta_0\}) = \left(1 + \frac{\pi_1}{\pi_0 B_{0,1}(x)}\right)^{-1}$$

so that we can really talk about the posterior probability of θ_0.

Exercise 6.7 Verify the preceding calculation. Check that if $\pi_0 = \pi_1 = 1/2$ and $B_{0,1}(x) = 1$, then $\Pi_x(\{\theta_0\}) = 1/2$.

If both $\Omega_0 = \{\theta_0\}$ and $\Omega_1 = \{\theta_1\}$ (point null versus a single alternative testing), we have $\mu_\Theta^{(0)} = \delta_{\theta_0}$ and $\mu_\Theta^{(1)} = \delta_{\theta_1}$. In this case the Bayes factor is the likelihood ratio and we have

$$\frac{\Pi_x(\Omega_0)}{\Pi_x(\Omega_1)} = \frac{\pi_0}{\pi_1} \times \frac{L(\theta_0; x)}{L(\theta_1; x)}.$$

Compared to the Neyman-Pearson lemma, the data enter in exactly the same way (via the likelihood ratio), but the decision to accept/reject a hypothesis is made on completely different criteria.

The difference between classical hypothesis testing and Bayesian hypothesis testing are clearly illustrated by the following example, an instance of what is known as **Lindley's paradox**.

Example 6.24. Assume $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Exp(\theta)$. Suppose we wish to test $H_0 : \theta = 1$ versus $H_0 : \theta \neq 1$. Let $S_n = \sum_{i=1}^n X_i$. Under the null hypothesis $T := (S_n - n)/\sqrt{n}$ has asymptotically a standard Normal distribution. Hence the test with testfunction $\phi(X_1, \ldots, X_n) = \mathbf{1}\{|T| \geq \xi_{\alpha/2}\}$ has significance level α in the large sample limit (Of course, for this testing problem there is no need to use asymptotics as $S_n \sim Ga(n, \theta)$, but it turns out to be convenient for the remainder of this example). Note that if we observe

$$s_n = n + \xi_{\alpha/2}\sqrt{n} \tag{6.5}$$

the test will have significance level α and the decision will be to reject, *for any* value of n.

We now also view the problem from the Bayesian point of view using a prior measure on θ. Let π_0 and π_1 be nonnegative and add to one. Define

$$\mu_\Theta(A) = \pi_0 \delta_{\{1\}}(A) + \pi_1 \int_A \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \mathbf{1}_{[0,\infty)}(\theta)\, \mathrm{d}\theta,$$

for A a Borel set in $\mathbb{R}$ and where $\delta_{\{1\}}(A) = \mathbf{1}_A(1)$ (writing the measure, instead of density f_Θ is somewhat more convenient due to the occurrence of the Dirac mass). The posterior odds are given by

$$\frac{\pi_1}{\pi_0} \frac{\int_0^\infty \theta^n e^{-\theta s} \Gamma(a)^{-1} b^a \theta^{a-1} e^{-b\theta}\, \mathrm{d}\theta}{e^{-s}} =: \frac{\pi_1}{\pi_0} B_n,$$

with

$$B_n = e^s \frac{\Gamma(n+a)}{\Gamma(a)} \frac{b^a}{(b+s)^{n+a}}.$$

To simplify, we take $a = b = 1$ so that $B_n = e^s(n!)(1+s)^{-n-1}$. Applying Sterling's formula, $n! \sim \sqrt{2\pi}\sqrt{n} n^n e^{-n}$ we get

$$B_n \sim e^{s-n} \sqrt{\frac{2\pi}{n}} \left(\frac{1+s}{n}\right)^{-(n+1)}.$$

Next, we take $s = s_n$, with s_n as defined in (6.5). This gives

$$B_n \sim \sqrt{\frac{2\pi}{n}} e^{\xi_{\alpha/2}\sqrt{n}} \left(1 + \frac{1}{n} + \frac{\xi_{\alpha/2}}{\sqrt{n}}\right)^{-(n+1)}.$$

This behaves asymptotically as $\sqrt{2\pi/n}$ and therefore tends to zero. So in case the observations satisfy (6.5) the frequentist test rejects for any value of n whereas the posterior probability of the alternative hypothesis tends to zero. Therefore, the conclusions from both methods are rather different.

Exercise 6.8 Suppose $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ and assume σ^2 is known. We consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.

1. Derive the critical region of the UMP-test with significance level α.
2. Assume a Bayesian setup in which $X_1, \ldots, X_n \mid \Theta = \theta \overset{\text{ind}}{\sim} N(\theta, \sigma^2)$ and Θ gets assigned a prior distribution. Take a flat prior on Θ: $f_\Theta(\theta) \propto 1$. Compute the posterior probability that H_0 is true.
3. Compare the computed posterior probability with the p-value and comment on this.

6.5.4 Application: classification under 0/1 loss

The Maximum a Posteriori estimator is given by

$$\Theta_{MAP} = \underset{\theta\in\Omega}{\operatorname{argmax}}\, f_{\Theta|X}(\theta \mid X) = \underset{\theta\in\Omega}{\operatorname{argmax}} \left(\log f_{X|\Theta}(X \mid \theta) + \log f_{\Theta}(\theta)\right).$$

Exercise 6.9 Verify the second equality sign.

Now consider the binary classification problem where $\mathcal{A} = \Omega = \{0, 1\}$. Then

$$\Theta_{MAP} = \mathbf{1}\left\{\frac{f_{X|\Theta}(x \mid 1)}{f_{X|\Theta}(x \mid 0)} \geq \frac{f_{\Theta}(0)}{f_{\Theta}(1)}\right\},$$

in which we recognise both the likelihood ratio (Bayes factor) and the prior ratio.

As an example, we consider the binary detection problem from the communications literature. Here, an emitter outputs either "0" or "1" with apriori probabilities $f_\Theta(0)$ and $f_\Theta(1)$ respectively. Each digit is transmitted through a noisy channel that adds to it a $N\left(0, \sigma^2\right)$ random variable. This is the simplest, but most common, model for channel noise in digital communications. As this model stipulates that $X \mid \Theta = \theta \sim N\left(\theta, \sigma^2\right)$ we get

$$\Theta_{MAP} = \mathbf{1}\{2X - 1 \geq 2\sigma^2 \log\left(f_\Theta(0)/f_\Theta(1)\right)\}.$$

Exercise 6.10

1. Verify this formula.
2. (*)Now suppose σ^2 is unknown. Set $u = 1/\sigma^2$ and consider u to be a realisation of the random variable U, which gets assigned the $Ga(\alpha, \beta)$-distribution. Find the MAP.
   Hint: $f_{\Theta|X}(\theta \mid x) = \int_0^\infty f_{\Theta,U|X}(\theta, u \mid x)\, \mathrm{d}u$.
3. (*) Derive the Bayes rule when the loss for incorrectly deciding "1" equals 1, but the loss for incorrectly deciding "0" equals $c > 0$.

We extend the binary classification problem to the case of a known finite number of classes. **Linear discriminant analysis** is concerned with the following statistical model: suppose $X = (X_1, \ldots, X_n)$ satisfies

$$X = \theta\, \mathbb{1} + Z, \qquad \text{with} \quad Z \sim N(0, C).$$

Here $\mathbb{1}$ denotes the vector with n times a 1 and Z is assumed to be *multivariate* Normally distributed, with covariance matrix C. If C is diagonal, then the X_i are independent, else not. We assume $\theta \in \Omega = \{\theta_1, \ldots, \theta_M\}$. The MAP estimator is defined by

$$\Theta_{MAP} = \underset{1\leq i\leq M}{\operatorname{argmax}} \left(-\frac{1}{2}\|X - \theta_i\|_C^2 + \log f_\Theta(\theta_i)\right),$$

where $\|x - \theta\|_C^2 = (x - \theta)' C^{-1}(x - \theta)$ is the squared *Mahalanobis*-distance. If we define

$$\alpha_i = C^{-1}\theta_i \qquad \text{and} \qquad \beta_i = f_\Theta(\theta_i) - \frac{1}{2}\|\theta_i\|^2,$$

then

$$\Theta_{MAP} = \underset{1\leq i\leq M}{\operatorname{argmax}} \left(X'\alpha_i + \beta_i\right).$$

6.5.5 Bayesian prediction

In practical problems the consequences of a decision may depend on future observations, say Y, and not directly on the parameter θ. In this setting, we have a model where the conditional distribution of $(X, Y) \mid \Theta$ is specified, augmented by a prior on Θ.

Definition 6.25. The **predictive loss function**[1] L is a measurable function from $\mathcal{X} \times \mathbb{A}$ to $[0, \infty)$. If $(X, Y) = (x, y)$ is observed then the loss for using decision rule d is given by $L(y, d(x))$.

Just like the definition of posterior risk is obtained by averaging the loss over the posterior of Θ (Cf. definition 6.15), we now define the predictive risk by averaging the loss over the predictive distribution of Y.

Definition 6.26. The **predictive risk** with respect to the prior μ_Θ is defined by

$$r_{pred}(\mu_\Theta, d \mid x) = \int L(y, d(x)) f_{Y|X}(y \mid x) \, \mathrm{d}y.$$

Here the predictive density $f_{Y|X}$ is defined in equation (4.8).

6.6 Finite decision problems

A finite decision problem is one in which the parameter space $\Omega = \{\theta_1, \ldots, \theta_k\}$ is finite. In such cases the notions of admissible, minimax and Bayes decision rules can be given a geometric interpretation. We start with two examples

Example 6.27. This example is extracted from Exercise 4.1 in YS. Suppose the random variable X has one of two possible densities:

$$f_X(x \mid \theta) = \theta e^{-\theta x}, \qquad x \in (0, \infty), \qquad \theta \in \{1, 2\}.$$

This is clearly a finite decision problem, the parameter set Ω being equal to $\{1, 2\}$.

We consider the family of decision rules

$$d_\mu(x) = \begin{cases} 1 & \text{if } x \geq \mu \\ 2 & \text{if } x < \mu \end{cases},$$

indexed by $\mu \in [0, \infty]$. To completely specify the decision problem, define the loss function to be $L(\theta, d) = |\theta - d|$. The risk function is then given by

$$\begin{aligned} R(\theta, d_\mu) &= \mathrm{E}_\theta \, |\theta - d_\mu(X)| \\ &= \mathrm{E}_\theta \, |\theta - d_\mu(X)| \mathbf{1}_{\{X \geq \mu\}} + \mathrm{E}_\theta \, |\theta - d_\mu(X)| \mathbf{1}_{\{X < \mu\}} \\ &= |\theta - 1| \mathrm{P}_\theta(X \geq \mu) + |\theta - 2| \mathrm{P}_\theta(X < \mu) \\ &= |\theta - 1| e^{-\theta\mu} + |\theta - 2|(1 - e^{-\theta\mu}). \end{aligned}$$

Therefore, the maximum risk is given by

$$\max_{\theta \in \{1,2\}} R(\theta, d_\mu) = \max\{1 - e^{-\mu}, e^{-2\mu}\}.$$

[1] I am not aware of common terminology.

By sketching this function, considered as a function of μ, it becomes clear that the minimax estimator satisfies

$$1 = e^{-\mu} = e^{-2\mu}.$$

Letting $\xi = e^{-\mu}$ this becomes a quadratic equation in ξ with only one positive solution, which is equal to $\xi_{minmax} = (-1 + \sqrt{5})/2$. Hence

$$\mu_{\text{minmax}} = -\log \frac{-1 + \sqrt{5}}{2} \approx 0.48.$$

For a given prior on Ω, the Bayes rule can be determined. If $\{1\}$ gets prior probability π_1 then the Bayes risk equals

$$\pi_1 R(1, d_\mu) + (1 - \pi_1) R(2, d_\mu) = \pi_1(1 - \xi) + (1 - \pi_1)\xi^2.$$

Minimising this function over all $\xi > 0$ is straightforward and gives and $\mu_{Bayes} = -\log \xi_{Bayes}$.

Exercise 6.11 For what prior mass function for θ does the minimax rule coincide with the Bayes rule?

Example 6.28. Suppose $\Omega = \{\theta_0, \theta_1\}$ and the action set is given by $\mathbb{A} = \{a_0, a_1\}$. Assume the loss function

$\theta \setminus a$	a_0	a_1
θ_0	0	2
θ_1	1	0

Suppose the statistician gets to see $X \sim Ber(\theta)$, with $\theta \in \Omega$. Since both the sample space $\mathcal{X} = \{0, 1\}$ and the action space $\mathbb{A}$ are finite, we are able to write down all (non-randomised) decision rules. These are

$$d_1(X) = a_0 \qquad d_2(X) = a_1$$

$$d_3(X) = \begin{cases} a_0 & \text{if } X = 0 \\ a_1 & \text{if } X = 1 \end{cases} \qquad d_4(X) = \begin{cases} a_1 & \text{if } X = 0 \\ a_0 & \text{if } X = 1 \end{cases}.$$

For computing the risk function of these decision rules we use that

$$\begin{aligned} R(\theta_0, d) &= L(\theta_0, a_0)\mathrm{P}_{\theta_0}(d(X) = a_0) + L(\theta_0, a_1)\mathrm{P}_{\theta_0}(d(X) = a_1) \\ &= 2P_{\theta_0}(d(X) = a_1) \end{aligned}$$

and

$$\begin{aligned} R(\theta_1, d) &= L(\theta_1, a_0)\mathrm{P}_{\theta_1}(d(X) = a_0) + L(\theta_1, a_1)\mathrm{P}_{\theta_1}(d(X) = a_1) \\ &= P_{\theta_1}(d(X) = a_0) \end{aligned}$$

For computing the minimax decision rule in this example, it turns out that we need to consider randomised decision rules.

Randomised decision rules are decision rules that include an "external" randomisation (think for example of randomised tests). Here we give the definition of a randomised decision rule when a finite number of (nonrandom) decision rules is given to us (the general definition is somewhat involved).

Definition 6.29. Let $p_i \in [0,1]$, $i = 1, \dots, I$ be such that $\sum_{i=1}^{I} p_i = 1$. Suppose $d_1, \dots, d_I$ are decision rules. A **randomised decision rule** d^* is obtained by choosing rule d_i with probability p_i. The loss of the randomised decision rule d^* is defined by

$$L(\theta, d^*) = \sum_{i=1}^{I} p_i L(\theta, d_i).$$

On pages 12 and 13 of Young and Smith [2005] it is explained how minimax and Bayes decision rules can be found using the geometry of the risk set in case $k = 2$. Precise mathematical statements that justify these derivations are given in section 3.2.4 of Schervish [1995].

Definition 6.30. Suppose $\Theta = \{\theta_1, \theta_2, \dots, \theta_k\}$. The **risk set** of the decision rule d is defined by

$$\mathcal{R} = \{z \in \mathbb{R}^k : \ z_i = R(\theta_i, d), \text{ for some randomised decision rule } d \text{ and } i = 1, \dots, k\}.$$

Lemma 6.31. If $|\Omega| < \infty$, the risk set is convex.

Proof. Suppose $z, w \in \mathcal{R}$ and $\alpha \in [0,1]$. Suppose z corresponds to decision rule d_z and w to decision rule d_w. For $i = 1, 2$

$$\alpha z_i + (1-\alpha) w_i = \alpha R(\theta_i, d_z) + (1-\alpha) R(\theta_i, d_w) = R(\theta_i, d^*), .$$

with d^* the randomised decision rule

$$d^* = \begin{cases} d_z & \text{with probability } \alpha \\ d_w & \text{with probability } 1-\alpha \end{cases}.$$

It follows that $\alpha z + (1-\alpha) w \in \mathcal{R}$. □

Example 6.32 (Continuation of example 6.28)**.** The risk set is the convex hull of

$$\{(R(\theta_0, d_i), R(\theta_1, d_i), \ 1 \le i \le 4\}.$$

It is easy to verify that these points are given by

$$(0,1) \qquad (2,0) \qquad (2\theta_0, 1-\theta_1) \qquad (2(1-\theta_0), \theta_1).$$

Suppose $p_1, p_2, p_3, p_4 \in [0,1]$ with $p_1 + p_2 + p_3 + p_4 = 1$. The randomised decision rule d^* picks rule d_i with probability p_i ($i \in \{1,2,3,4\}$). Note that

$$\begin{bmatrix} R(\theta_0, d^*) \\ R(\theta_1, d^*) \end{bmatrix} = p_1 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + p_2 \begin{bmatrix} 2 \\ 0 \end{bmatrix} + p_3 \begin{bmatrix} 2\theta_0 \\ 1-\theta_1 \end{bmatrix} + p_4 \begin{bmatrix} 2(1-\theta_0) \\ \theta_1 \end{bmatrix}.$$

Next, we try to find d^* (i.e. p_1, p_2, p_3 and p_4) such that

$$\max\{R(\theta_0, d^*), R(\theta_1, d^*)\}$$

is minimised. To make the problem a bit more concrete, suppose

$$\theta_0 = 1/3 \qquad \theta_1 = 3/4.$$

Then the risk set contains the points

$$(0,1) \qquad (2,0) \qquad (2/3, 1/4) \qquad (4/3, 3/4).$$

A sketch of the risk set then reveals that d_4 is inadmissible (any probability assigned to this rule can better be assigned to d_3 to lower the maximal risk). Just as in example 6.27 the minimax rule is the rule for which $R(\theta_0, d^*) = R(\theta_1, d^*)$. From the form of the risk set it then follows that the minimax rule is of the form

$$p_1 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + p_3 \begin{bmatrix} 2/3 \\ 1/4 \end{bmatrix}$$

with $p_1 + p_3 = 1$. Both coordinates are equal for $p_1 = 5/17$, resulting in the minimax risk being equal to $8/17$. Note that this number is indeed smaller than the maximum risk of the best non-randomised estimator (which is d_3 with max-risk equal to $2/3$.

In the previous example the minimax estimator is a randomised estimator. The need of an external randomisation device (completely independent of the data) is unappealing to many statisticians, but a consequence of taking the maximum risk as optimality criterion.

Example 6.33 (Continuation of example 6.28)**.** We now turn to Bayes rules. Let π be the prior probability of $\{\theta_0\}$. We wish to find the rule d_i ($i \in \{1, 2, 3, 4\}$) for which

$$c = \eta R(\theta_0, d) + (1 - \eta) R(\theta_1, d)$$

is smallest. Rewriting this equation gives

$$R(\theta_1, d) = \frac{c}{1 - \eta} - \frac{\eta}{1 - \eta} R(\theta_0, d).$$

As example, suppose $\eta = 1/2$. Then the Bayes rule is the rule where the line with slope -1 touches the risk set. This is at d_3.

Exercise 6.12 Suppose $\eta = 3/4$ in the preceding example, what is the Bayes rule? Is it unique. Repeat for $\eta = 9/17$.

The previous exercise shows that Bayes rules need not be unique, and may be randomised. However, for any randomised rule there is a non-randomised rule with the same Bayes risk. Hence, within the Bayesian setup randomised rules are not needed, which can hardly be surprising when thinking about the likelihood principle.

Exercise 6.13 In example 6.28 assume $\theta_0 = \frac{1}{10}$, $\theta_1 = \frac{1}{2}$.

1. Make a sketch of the risk set.
2. Find the minimax estimator and the prior for which it is a Bayes rule.

Define the **lower quantant** at $r \in \mathbb{R}^k$ by

$$Q_r = \{x \in \mathbb{R}^k \ : \ x_i \leq r_i, \ \forall\, 1 \leq i \leq k\}.$$

For a convex set S define the set of **lower boundary points** by

$$\lambda(S) = \left\{ r \ : \ Q_r \cap \bar{S} = \{r\} \right\}.$$

Finally, define S to be **closed from below** if $\lambda(S) \subset S$. If S is closed, then it follows that S is closed from below. As an example in $\mathbb{R}^2$, take S to be an open ball. Then (make a sketch) in this case S is not closed from below. Any decision rule with risk vector that is in $\lambda(\mathcal{R})$ is admissible.

The following theorem shows that minimax rules exist and correspond to a Bayes rule with respect to a particular prior, which is then called least favourable.

Theorem 6.34 (Minimax theorem, Schervish [1995] page 172)**.** Suppose that the loss function is bounded below and Ω is finite. Then

$$\sup_{\mu} \inf_{d} r(\mu, d) = \inf_{d} \sup_{\theta} R(\theta, d)$$

and there exists a prior μ_0 for Θ such that

$$\inf_{d} r(\mu_0, d) = \sup_{\mu} \inf_{d} r(\mu, d). \tag{6.6}$$

The prior μ_0 such that (6.6) is called **least favourable**. If $\mathcal{R}$ is closed from below, then there is a minimax rule that is a Bayes rule with respect to μ_0.

We omit the proof, which is given in Schervish [1995].

Exercise 6.14 [YS exercise 2.5.] Bacteria are distributed at random in a fluid, with mean density θ per unit volume, for some $\theta \in H \subseteq [0, \infty)$. This means that

$$\mathrm{P}_\theta(\text{no bacteria in volume } v) = e^{-\theta v}.$$

We remove a sample of volume v from the fluid and test it for the presence or absence of bacteria. On the basis of this information we have to decide whether there are any bacteria in the fluid at all. An incorrect decision will result in a loss of 1, a correct decision in no loss.

(a) Suppose $H = [0, \infty)$. The problem can be cast into the decision framework as follows by defining the action space $\mathcal{A} = \{a_0, a_1\}$ with

$$\begin{aligned} a_0 &: \quad \text{there are bacteria in the fluid} \\ a_1 &: \quad \text{there are no bacteria in the fluid} \end{aligned}$$

The "state of nature" is $\theta \in H \subseteq [0, \infty)$. The loss function is given by

$$L(\theta, a_1) = \begin{cases} 0 & \text{if } \theta = 0 \text{ (conclude correctly no bacteria)} \\ 1 & \text{if } \theta > 0 \text{ (incorrectly conclude no bacteria)} \end{cases}.$$

$$L(\theta, a_0) = \begin{cases} 1 & \text{if } \theta = 0 \text{ (incorrectly conclude there are bacteria)} \\ 0 & \text{if } \theta > 0 \text{ (correctly conclude there are bacteria)} \end{cases}.$$

Finally, the statistician observes X with

$$X = \begin{cases} 0 & \text{test results is that there are bacteria} \\ 1 & \text{test results is that there no are bacteria} \end{cases}.$$

Describe all the non-randomised decision rules for this problem and calculate their risk. Which of these rules are admissible?

(b) Suppose $H = \{0, 1\}$. Identify the risk set

$$\mathcal{R} = \{(R(0, d), R(1, d)) : \ d \text{ a randomised rule}\} \subseteq \mathbb{R}^2$$

where $R(\theta, d)$ is the expected loss in applying d under P_θ. Show that the minimax rule is a randomised rule, where the randomisation consists of tossing a coin that lands heads with probability $(1 + e^v)^{-1}$.

(c) Now suppose again that $H = [0, \infty)$, just as under (a). Determine the Bayes decision rules and Bayes risk for the prior

$$\mu_\Theta(\mathrm{d}\theta) = \frac{1}{3}\delta_0(\mathrm{d}\theta) + \frac{2}{3}e^{-\theta}\,\mathrm{d}\theta.$$

Hint: The Bayes risk of the rule d is given by

$$\frac{1}{3}R(0, d) + \frac{2}{3}\int_0^\infty R(\theta, d)e^{-\theta}d\theta.$$

(d) If it costs additionally $v/24$ to test a sample of volume v (so $v/24$ is added to the loss-function), what is the optimal volume to test? What if the cost is $1/6$ per unit volume?

Exercise 6.15 [YS exercise 2.4] An unmanned rocket is being launched in order to place in orbit an important new communications satellite. At the time of launching, a certain crucial electronic component is either functioning or not functioning. In the control centre there is a warning light that is not completely reliable. If the crucial component is not functioning, the warning light goes on with probability $2/3$; if the component is functioning, it goes on with probability $1/4$. At the time of launching, an observer notes whether the warning light is on or off. It must then be decided immediately whether or not to launch the rocket.

There is no loss associated with launching the rocket with the component functioning, or aborting the launch when the component is not functioning. However, if the rocket is launched when the component is not functioning, the satellite will fail to reach the desired orbit. The Space Shuttle mission required to rescue the satellite and place it in the correct orbit will cost 10 billion dollars. Delays caused by the decision not to launch when the component is functioning result, through lost revenue, in a loss of 5 billion dollars.

Suppose that the prior probability that the component is not functioning is $\psi = 2/5$. If the warning light does not go on, what is the decision according to the Bayes rule? For what values of the prior probability ψ is the Bayes decision to launch the rocket, even if the warning light comes on?
Hints: The actions can be defined as

$$\begin{cases} a_0 & \text{do not launch rocket} \\ a_1 & \text{launch rocket} \end{cases}.$$

The "state of nature" is θ where

$$\theta = \begin{cases} 0 & \text{component is not functioning} \\ 1 & \text{component is functioning} \end{cases}.$$

The observation is

$$X = \begin{cases} 1 & \text{warning light turns on} \\ 0 & \text{warning light does not turn on} \end{cases}.$$

First write down the loss function and all non randomised decision rules. Next, compute the risk function.

6.6.1 Complete classes*

There is a famous example where a minimal complete class exists.

Theorem 6.35 (Neyman-Pearson fundamental lemma, Schervish [1995] theorem 3.87)**.** Let $\Omega = \mathbb{A} = \{0, 1\}$ and assume the loss function

$\theta \setminus a$	0	1
0	0	k_0
1	k_1	0

with $k_0 > 0$ and $k_1 > 0$. Let $f_i(x) = dP_i/d\nu(x)$ for $i = 0, 1$, where $\nu = P_0 + P_1$. Let C denote the class of all rules with test function of the following forms:

- For $k \in (0, \infty)$ and each function $\gamma : \mathcal{X} \to [0, 1]$

$$\phi_{k,\gamma}(x) = \begin{cases} 1 & \text{if } f_1(x) > k f_0(x) \\ \gamma(x) & \text{if } f_1(x) = k f_0(x) \\ 0 & \text{if } f_1(x) < k f_0(x) \end{cases}$$

- For $k = 0$,

$$\phi_0(x) = \begin{cases} 1 & \text{if } f_1(x) > 0 \\ 0 & \text{if } f_1(x) = 0 \end{cases}$$

- For $k = \infty$,

$$\phi_\infty(x) = \begin{cases} 1 & \text{if } f_0(x) = 0 \\ 0 & \text{if } f_0(x) > 0 \end{cases}$$

Then $\mathcal{C}$ is a minimal complete class.

Obviously, in the setting of this theorem there is no need to look for other test than the likelihood ratio test. As Parmigiani and Inoue [2009] put it (page 168):

> With this result we have come in a complete circle: the Neyman-Pearson theory was the seed that started Wald's statistical decision theory: minimal completeness is the ultimate rationality endorsement for a statistical approach within that theory – all and only the rules generated by the approach are worth considering. The Neyman-Pearson tests are a minimal complete class. Also, for each of these tests we can find a prior for which that test is the formal Bayes rule. What is left to argue about?

A more general version of the preceding theorem can be derived:

Theorem 6.36 (Complete class theorem, Schervish [1995] theorem 3.95)**.** Suppose $|\Omega| = k$, the loss function is bounded below, and the risk set is closed from below. Then the set of all Bayes rules is a complete class, and the set of admissible Bayes rules is a minimal complete class. These are also the rules whose risk functions are on the boundary of the risk set.

6.7 Minimax-Bayes connections

In this section we show that minimax are often limits of Bayes rules. We start with an extension of the concept of a Bayes rule.

Definition 6.37. The decision rule d_0 is called **extended Bayes** if for each $\varepsilon > 0$ there exists a (proper) prior μ_ε such that

$$r(\mu_\varepsilon, d_0) \leq \varepsilon + \inf_d r(\mu_\varepsilon, d).$$

A Bayes rule is always extended Bayes. The concept of extended Bayes rule involves a weaker requirement than Bayes rule because:

- the prior may depend on ε;
- the Bayes risk need only be attained by d_0 up to an amount ε.

Example 6.38. If $X \sim N(\theta, 1)$, $L(\theta, a) = (\theta - a)^2$, then $d_0 = X$ is not a Bayes rule by Proposition 6.52 ahead (note that d_0 is unbiased for estimating θ and that this proposition says that Bayes rules are necessarily biased). However, it is an extended Bayes rule, as we now show.

As $R(\theta, d_0) = \mathrm{E}_\theta((\theta - X)^2) = 1$, we have $r(\mu, d_0) = 1$ for any prior μ. Take the $N\left(0, \sigma^2\right)$-prior on θ, and denote it by μ_σ. It follows that $r(\mu_\sigma, d_0) = 1$.

As we assume the quadratic loss, Bayes rule is given by the posterior mean, which is given by

$$d_\sigma(X) = \frac{\sigma^2}{1+\sigma^2} X.$$

Cf. Example 4.4. This rule shrinks the observation X towards the prior mean (which is assumed zero here). The risk function of d_σ is given by

$$\begin{aligned} R(\theta, d_\sigma) &= \mathrm{E}_\theta\left[(\theta - \sigma^2(1+\sigma^2)^{-1}X)^2\right] \\ &= \theta^2 - \frac{2\sigma^2}{1+\sigma^2}\theta^2 + \frac{\sigma^4}{(1+\sigma^2)^2}(1+\theta^2) = \theta^2 \frac{1}{(1+\sigma^2)^2} + \frac{\sigma^4}{(1+\sigma^2)^2}. \end{aligned}$$

This implies

$$r(\mu_\sigma, d_\sigma) = \mathrm{E}_{\theta\sim\mu_\sigma} R(\theta, d_\sigma) = \frac{\sigma^2}{1+\sigma^2}.$$

For any $\sigma > 0$ we have

$$r(\mu_\sigma, d_0) = 1 = \left(1 - \frac{\sigma^2}{1+\sigma^2}\right) + r(\mu_\sigma, d_\sigma).$$

As d_σ minimises $r(\mu_\sigma, d)$ over all decision rules (it is the Bayes rule!), it follows that

$$r(\mu_\sigma, d_0) = 1 \leq \left(1 - \frac{\sigma^2}{1+\sigma^2}\right) \inf_d r(\mu_\sigma, d).$$

The claim now follows upon taking $\sigma = \sigma_\varepsilon$ such that $\varepsilon = 1/(1+\sigma_\varepsilon^2)$.

The idea of the proof is to take a class of prior distributions for which the posterior is tractable. The following example is similar in spirit.

Example 6.39. Suppose $X \sim Pois(\theta)$ and the loss function is given by $L(\theta, a) = (\theta - a)^2/\theta$. We prove that $d_0(X) = X$ is extended bayes. To see this, for any prior μ

$$r(\mu, d_0) = \int \mathrm{E}_\theta\left[\frac{(\theta - X)^2}{\theta}\right] \mu(\,\mathrm{d}\theta) = 1.$$

Take a sequence of priors μ_λ such that under μ_λ, $\Theta \sim Exp(\lambda)$. Then

$$f_{\Theta|X}(\theta \mid x) \propto \theta^x e^{-(1+\lambda)\theta}$$

and hence $\Theta \mid X \sim Ga(X+1, 1+\lambda)$. To find the Bayes rule for the given loss function, we need to choose d to minimise

$$d \mapsto \int (d-\theta)^2 \theta^{x-1} e^{-(1+\lambda)\theta}\,\mathrm{d}\theta.$$

But this minimiser is exactly the mean of the $Ga(X, 1+\lambda)$-distribution and thus the Bayes rule is given by $d_\lambda = X/(1+\lambda)$. To compute the Bayes risk of d_λ, first note that

$$R(\theta, d_\lambda) = \frac{1}{\theta}\,\mathrm{E}\left[\left(\frac{X}{1+\lambda} - \theta\right)^2\right] = \frac{1+\lambda^2\theta}{(1+\lambda)^2}$$

We leave it as an exercise to show that then

$$r(\mu_\lambda, d_\lambda) = 1/(1+\lambda).$$

The claimed result now follows from

$$r(\mu_\lambda, d_0) = 1 = \left(1 - \frac{1}{1+\lambda}\right) + \frac{1}{1+\lambda} \leq \frac{\lambda}{1+\lambda} + \inf_d r(\pi_\lambda, d)$$

and taking $\lambda_\varepsilon = \varepsilon/(1-\varepsilon)$.

Both of the decision rules d_0 in the previous examples are examples of rules with constant risk.

Definition 6.40. An **equaliser rule** is a decision rule with constant risk.

The following theorem shows that an extended Bayes rule with constant risk is minimax, generalising the two examples just given.

Theorem 6.41. Suppose d_0 is extended Bayes and $R(\theta, d_0)$ is constant for all θ. Then d_0 is minimax.

Here, the loss function is assumed fixed.

Proof. Suppose $R(\theta, d_0) = C$. Suppose d_0 is not minimax. Then there exists a rule d' for which $\sup_\theta R(\theta, d') < C$. So let $\sup_\theta R(\theta, d') = C - \varepsilon$ for some $\varepsilon > 0$. As d_0 is extended Bayes, we can find a prior μ_ε such that

$$r(\mu_\varepsilon, d_0) < \inf_d r(\mu_\varepsilon, d) + \varepsilon/2 \leq r(\mu_\varepsilon, d') + \varepsilon/2 \leq C - \varepsilon + \varepsilon/2 = C - \varepsilon/2.$$

Since d_0 is an equaliser rule $r(\pi_\varepsilon, d_0) = C$ and we have reached a contradiction. □

Example 6.42. Suppose that $X \sim N(\theta, 1)$ and the loss function is given by $L(\theta, a) = (\theta - a)^2$. Then $d_0 = X$ is extended Bayes. As $R(\theta, d_0) = 1$, it follows that X is minimax.

Theorem 6.43. Suppose

1. μ_n is a sequence of priors with d_n being the Bayes rule with respect to μ_n;
2. $\lim_{n\to\infty} r(\mu_n, d_n) = C$;
3. there exists a d_0 such that $R(\theta, d_0) \leq C$ for all $\theta \in \Omega$.

Then d_0 is minimax.

Proof. Suppose d_0 is not minimax. Then there exists a d' for which $\sup_\theta R(\theta, d') < C$. Hence there exists a $\varepsilon > 0$ such that $R(\theta, d') < C - \varepsilon$ for all θ. As $r(\mu_n, d_n) \to C$ we can find an n for which

$$|r(\mu_n, d_n) - C| < \varepsilon/2 \qquad C - \varepsilon/2 < r(\pi_n, d_n) < C + \varepsilon/2$$

So we have

$$r(\mu_n, d') \leq \sup_\theta R(\theta, d') < C - \varepsilon < C - \varepsilon/2 < r(\mu_n, d_n).$$

But then d_n cannot be Bayes with respect to μ_n which is a contradiction.

□

This theorem suggests a recipe for finding a minimax rule d

1. Make a guess on an equaliser d_0 with $R(\theta, d_0) = C$.
2. Propose a sequence of priors μ_n.
3. Find a Bayes rule d_n for each μ_n.

4. Check if $\lim_{n\to\infty} r(\mu_n, d_n) = C$.

Example 6.44. Suppose $X_i \overset{\text{ind}}{\sim} N\left(\theta_i, 1\right)$, $1 \leq i \leq m$. Define

$$\theta = (\theta_1, \ldots, \theta_m)$$
$$d_0(X) = X = (X_1, \ldots X_m)$$
$$L(\theta, a) = \sum_{i=1}^{m} (\theta_i - a_i)^2.$$

Let μ_n be the prior such that $\Theta \sim N_m(0, nI)$. The Bayes rule is

$$d_n(X) = nX/(n+1).$$

The Bayes risk satisfies

$$r(\pi_n, d_n) = mn/(n+1)$$

which converges to m as $n \to \infty$. As $R(\theta, d_0) = m$ is constant, it follows that X is minimax.

Example 6.45. Suppose $X_1, \ldots, X_n \overset{\text{ind}}{\sim} N(\theta, 1)$. For estimation of θ under L_2-loss, $d(X) = \bar{X}_n$ is minimax. To see this, first note that $\bar{X}_n$ is an equaliser rule

$$\mathrm{E}\left[(\bar{X}_n - \theta)^2\right] = 1/n.$$

Now take $\mu_k \sim N(0, k)$. Let d_k denote the Bayes rule corresponding to μ_k. Some computations show that

$$r(\mu_k, d_k) = \frac{k/n}{k + 1/n},$$

which tends to $1/n$ as $k \to \infty$.

Definition 6.46. A prior μ_0 for which $r(\mu, d_\mu)$ is maximised is called a **least favourable prior**:

$$r(\mu_0, d_{\mu_0}) = \sup_{\mu} r(\mu, d_\mu).$$

Theorem 6.47. Suppose μ is a prior distribution on Ω such that

$$r(\mu, d_\mu) = \sup_{\theta} R(\theta, d_\mu).$$

Then

1. d_μ is minimax.
2. If d_μ is unique Bayes with respect to μ, then d_μ is unique minimax.
3. μ is least favourable.

Proof. Let d be another rule. Then

$$\sup_{\theta} R(\theta, d_\mu) = r(\mu, d_\mu) \leq r(\mu, d) \leq \sup_{\theta} R(\theta, d).$$

Hence d_μ is minimax. If d_μ is *unique* Bayes, then the first inequality is strict, and then d_μ is unique minimax. Let μ^* be another prior distribution. Then

$$r(\mu^*, d_{\mu^*}) \leq r(\mu^*, d_\mu) \leq \sup_{\theta} R(\theta, d_\mu) = r(\mu, d_\mu).$$

□

Example 6.48. Suppose $X \mid \Theta = \theta \sim Bin(n, \theta)$ and squared error loss. Take $\Theta \sim Be(a, b)$, then under squared error loss the Bayes rule is given by

$$d(X) = \mathrm{E}[\Theta \mid X] = \frac{a + X}{a + b + n}.$$

If $a = b = \sqrt{n}/2$, then

$$R(\theta, d) = \mathrm{E}\left[(d(X) - \theta)^2\right] = \text{constant}.$$

It follows that $d(X)$ is the unique minimax rule and that the $Be\left(\sqrt{n}/2, \sqrt{n}/2\right)$-prior is least-favourable.

Exercise 6.16 [YS exercise 2.7.] In the context of a finite decision problem, decide whether each of the followingstatements is true, providing a proof or counterexample as appropriate.

1. The Bayes risk of a minimax rule is never greater than the minimax risk.
2. If a Bayes rule is not unique, then it is inadmissible.

Exercise 6.17 [YS exercise 3.4.] Suppose $X \mid \Theta = \theta \sim Bin(n, \theta)$ and $\Theta \sim Unif(0, 1)$. Consider loss function

$$L(\theta, d) = \frac{(\theta - d)^2}{\theta(1 - \theta)}.$$

Derive the Bayes rule. Is it minimax?
Hint: In order to show that the Bayes rule is minimax, show that the risk of the Bayes rule is constant and apply Theorem 6.41.

Exercise 6.18 * Let $\Theta = [0, 1)$, $\mathbb{A} = [0, 1]$ and $L(\theta, a) = (\theta - a)^2/(1 - \theta)$. Suppose X is a random variable with probability mass function

$$\mathrm{P}_\theta(X = x) = (1 - \theta)\theta^x, \qquad x = 0, 1, 2, \ldots$$

1. Write the risk function $R(\theta, d)$ for a decision rule d as a power series in θ.
2. Show that the only nonrandomised equaliser rule is $d(0) = 1/2$, $d(1) = d(2) = \cdots = 1$.
   Hint: Proving that the given rule is an equaliser rule should not be too hard. Proving that it is the only equaliser rule is somewhat harder. A first step consists of showing that for $\ell \geq 2$, an equaliser rule satisfies $d(\ell - 1) \geq d(\ell)$.
3. Suppose a prior distribution π on Θ is given and define $\mu_\ell = \mathbb{E}_\pi \Theta^\ell$. Show that if if $d(\ell) = \mu_{\ell+1}/\mu_\ell$ for $\ell = 0, 1, 2, \ldots$, then d is a Bayes rule for the prior π.
4. Show that the rule in (b) is extended Bayes, hence minimax.
   Hint: for d to be Bayes with respect to a prior π, we must have $d(\ell) = \mu_{\ell+1}/\mu_\ell$ according to (c). Verify that this reduces to $\mu_1 = \mu_2 = \mu_3 = \cdots = 1/2$. Next, verify that the prior with $\pi(0) = \pi(1) = 1/2$ has this property. As 1 is not in the parameter space, this is not a Bayes rule. To show that it is extended Bayes, consider the prior distributions π_ε with $\pi_\varepsilon(0) = \pi_\varepsilon(1 - \varepsilon) = 1/2$.

6.8 The role of sufficient statistics

The Rao-Blackwell theorem asserts that for convex loss functions we only need to consider decision rules that depend on sufficient statistics.

Theorem 6.49 (Rao-Blackwell theorem)**.** Suppose the action space $\mathbb{A}$ of a statistical decision problem is a convex subset of $\mathbb{R}^m$ and that for all $\theta \in \Theta$, $L(\theta, a)$ is a convex function of a. Suppose also that T is sufficient for θ and d_0 is a nonrandomised decision rule such that $\mathrm{E}_\theta[\|d_0(X)\|] < \infty$. Define

$$d_1(T) = \mathrm{E}_\theta[d_0(X) \mid T],$$

then $R(\theta, d_1) \leq R(\theta, d_0)$ for all θ.

Proof. From the conditional version of Jensen's inequality

$$L(\theta, d_1(T)) = L(\theta, \mathrm{E}_\theta[d_0(X) \mid T]) \leq \mathrm{E}_\theta[L(\theta, d_0(X)) \mid T].$$

Now take the expectation on both sides with respect to the distribution of T under $X \sim \mathrm{P}_\theta$

$$\mathrm{E}_\theta\, L(\theta, d_1(T)) \leq \mathrm{E}_\theta\, L(\theta, d_0(X))$$

which is the result. □

Note that, by sufficiency, $\mathrm{E}_\theta[d_0(X) \mid T]$ does not depend on θ. The result gives a direct recipe for improving estimators, as illustrated by the following examples.

Example 6.50. Suppose $X_1, \ldots, X_n$ are independent $N(\theta, 1)$ random variables. Suppose we wish to estimate $\eta = \mathbb{P}_\theta(X_1 \leq c) = \Phi(c - \theta)$ under quadratic loss

$$L(\theta, a) = (a - \Phi(c - \theta))^2.$$

A naive estimator that is unbiased for η is given by $S = n^{-1} \sum_{i=1}^n \mathbf{1}\{X_i \leq c\}$. However, this estimator does not depend on the sufficient statistic $T = \sum_{i=1}^n X_i$. As the loss function is convex in a we obtain the improved estimator by calculating

$$\mathbb{E}[S \mid T] = \mathbb{P}_\theta(X_1 \leq c \mid T) = \Phi\left(\frac{c - T/n}{\sqrt{(n-1)/n}}\right).$$

The second equality in fact does not depend on θ (by sufficiency). The final equality follows from general results on the multivariate normal distribution (you can skip the details of this computation if you wish).

Exercise 6.19 [YS exercise 6.3.] Independent factory-produced items are packed in boxes each containing k items. The probability that an item is in working order is θ with $0 < \theta < 1$. A sample of n boxes are chosen for testing, and X_i, the number of working items in the i-th box, is noted. Thus $X_1, \ldots, X_n$ are a sample from a binomial distribution, $Bin(k, \theta)$, with index k and parameter θ. It is required to estimate the probability, θ^k, that all items in a box are in working order. Find the minimum variance unbiased estimator, justifying your answer. Proceed along the following steps:

1. Show that $T = \sum_{i=1}^{n} X_i$ is sufficient for θ.
2. Show that $S = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i = k\}}$ is an unbiased estimator of θ^k.
3. Argue that $V = \mathbb{E}[S \mid T]$ is an unbiased estimator of minimum variance for θ.
4. Verify that
$$V = \frac{\binom{(n-1)k}{T-k}}{\binom{nk}{T}}$$

Example 6.51. Suppose interest lies in approximating the integral $I = \int h(x)\mathrm{P}(dx)$ where P is a probability measure and h a measurable function on $\mathbb{R}^k$ such that $\int |h(x)|\mathrm{P}(dx) < \infty$. Especially when k is large, Monte-Carlo simulation is a common way to approximate I. Hence, suppose $X_1, \ldots, X_n$ are draws from P, then a straightforward estimator is defined by

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} h(X_i).$$

If X can be decomposed as $X = (Y, Z)$ and the conditional expectation of $\mathrm{E}[h(X) \mid Y] =: g(Y)$ can be carried out in closed form then an alternative estimator for I is given by

$$\tilde{I} = \frac{1}{n} \sum_{i=1}^{n} g(Y_i).$$

Both estimators are unbiased. To compare variances, we have

$$\mathrm{Var}\, h(X) = \mathrm{Var}\, \mathrm{E}[h(X) \mid Y] + \mathrm{E}[\mathrm{Var}(h(X) \mid Y)] \geq \mathrm{Var}\, g(Y).$$

This implies that the Mean Squared Error of $\hat{I}$ can never be smaller than that of $\tilde{I}$ and one should prefer $\tilde{I}$ for estimating I (provided that the computational effort of evaluating $h(X)$ and $g(Y)$ is comparable). The Rao-Blackwell theorem asserts that ideally we decompose X such that we condition on a sufficient statistic. Consult page 50 of Young and Smith [2005] to see a nice example of this procedure, which is referred to as **Rao-Blackwellisation**.

6.9 Bayes rules and unbiasedness

In this section we restrict attention to estimation problems where $\mathbb{A} = \Omega \subset \mathbb{R}^k$.

Whereas unbiasedness is an important topic in classical statistics, Bayes rules on the contrary are usually biased. As the parameter is considered random in the Bayesian setup we have to be careful in what we mean by this. We say that $d(X)$ is unbiased for $g(\Theta)$ if $\mathrm{E}[d(X) \mid \Theta] = g(\Theta)$.

Proposition 6.52. Suppose $g : \Omega \to \mathbb{R}$ is measurable. Let d be a Bayes rule for estimating $g(\theta)$ under L_2 loss using the prior μ. If $r(\mu, d) \neq 0$, then d is biased for $g(\theta)$.

Proof. Assume d is unbiased for $g(\theta)$. We show that $r(\mu, d) = 0$. By conditioning on Θ

$$\mathbb{E}[g(\Theta)d(X)] = \mathbb{E}\mathbb{E}[g(\Theta)d(X) \mid \Theta] = \mathbb{E}[g(\Theta)\mathbb{E}[d(X) \mid \Theta]] = \mathbb{E}\left[g(\Theta)^2\right].$$

By conditioning on X

$$\mathbb{E}[g(\Theta)d(X)] = \mathbb{E}\mathbb{E}[g(\Theta)d(X) \mid X] = \mathbb{E}[d(X)\mathbb{E}[g(\Theta) \mid X]] = \mathbb{E}\left[d(X)^2\right],$$

because the Bayes rule is the posterior mean under L_2 loss.

The result follows from (the expectation is over (Θ, X))

$$r(\pi, d) = \mathbb{E}\left[(d(X) - g(\Theta))^2\right] = \mathbb{E}\left[d(X)^2\right] - 2\mathbb{E}[g(\Theta)d(X)] + \mathbb{E}\left[g(\Theta)^2\right]$$

and substituting the previous two displays. □

Bibliography

G. A. Barnard. Statistical inference. *J. Roy. Statist. Soc. Ser. B.*, 11:115–139; discussion, 139–149, 1949. ISSN 0035-9246.

Debabrata Basu. Statistical information and likelihood [with discussion]. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 1–71, 1975.

James Berger. The case for objective Bayesian analysis. *Bayesian Anal.*, 1(3):385–402, 2006. ISSN 1936-0975.

James O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 1985. ISBN 0-387-96098-8. doi: 10.1007/978-1-4757-4286-2. URL http://dx.doi.org/10.1007/978-1-4757-4286-2.

James O. Berger, Jose M. Bernardo, and Dongchu Sun. Overall objective priors. *Bayesian Anal.*, 10(1):189–221, 2015. ISSN 1936-0975. doi: 10.1214/14-BA915. URL http://dx.doi.org/10.1214/14-BA915.

Jose-M. Bernardo and Adrian F. M. Smith. *Bayesian theory*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester, 1994. ISBN 0-471-92416-4. doi: 10.1002/9780470316870. URL http://dx.doi.org/10.1002/9780470316870.

Allan Birnbaum. On the foundations of statistical inference. *J. Amer. Statist. Assoc.*, 57:269–326, 1962. ISSN 0162-1459.

Christopher M Bishop. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

William M. Briggs. *Breaking the Law of Averages. Real-Life Probability and Statistics in Plain English*. 2008. URL http://wmbriggs.com/public/briggs_breaking_law_averages.pdf.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

Aubrey Clayton. Bernoulli's fallacy. In *Bernoulli's Fallacy*. Columbia University Press, 2021.

Jacob Cohen. Tthe earth is round ($p < .05$). *American Psychologist*, 49:997–1003, 1994.

National Research Council. *Frontiers in Massive Data Analysis*. Committee on the Analysis of Massive Data, Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Their Applications & Division on Engineering and Physical Sciences. The National Academies Press, 2013.

M. Dashti and A. M. Stuart. The Bayesian Approach To Inverse Problems. *ArXiv e-prints*, February 2013.

B. Efron. Why isn't everyone a Bayesian? *Amer. Statist.*, 40(1):1–11, 1986. ISSN 0003-1305. doi: 10.2307/2683105. URL http://dx.doi.org/10.2307/2683105. With discussion and a reply by the author.

Bradley Efron. Bayesians, frequentists, and scientists. *J. Amer. Statist. Assoc.*, 100(469):1–5, 2005. ISSN 0162-1459. doi: 10.1198/016214505000000033. URL http://dx.doi.org/10.1198/016214505000000033.

Bradley Efron and David V. Hinkley. Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika*, 65(3):457–487, 1978. ISSN 0006-3444. doi: 10.1093/biomet/65.3.457. URL http://dx.doi.org/10.1093/biomet/65.3.457. With comments by Ole Barndorff-Nielsen, A. T. James, G. K. Robinson and D. A. Sprott and a reply by the authors.

Bradley Efron and Carl Norris. Data analysis using stein's estimator and its generlizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

Ronald A. Fisher. Mathematical probability in the natural sciences. *Technometrics*, 1:21–29, 1959. ISSN 0040-1706.

Greg Gandenberger. A new proof of the likelihood principle. *British J. Philos. Sci.*, 66(3):475–503, 2015. ISSN 0007-0882. doi: 10.1093/bjps/axt039. URL http://dx.doi.org/10.1093/bjps/axt039.

Jayanta K. Ghosh, Mohan Delampady, and Tapas Samanta. *An introduction to Bayesian analysis*. Springer Texts in Statistics. Springer, New York, 2006. ISBN 978-0387-40084-6; 0-387-40084-2. Theory and methods.

Piet Groeneboom and Geurt Jongbloed. *Nonparametric estimation under shape constraints*, volume 38 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, New York, 2014. ISBN 978-0-521-86401-5. doi: 10.1017/CBO9781139020893. URL http://dx.doi.org/10.1017/CBO9781139020893. Estimators, algorithms and asymptotics.

E. T. Jaynes. *Probability theory*. Cambridge University Press, Cambridge, 2003. ISBN 0-521-59271-2. doi: 10.1017/CBO9780511790423. URL http://dx.doi.org/10.1017/CBO9780511790423. The logic of science, Edited and with a foreword by G. Larry Bretthorst.

Robert W. Keener. *Theoretical statistics*. Springer Texts in Statistics. Springer, New York, 2010. ISBN 978-0-387-93838-7. doi: 10.1007/978-0-387-93839-4. URL http://dx.doi.org/10.1007/978-0-387-93839-4. Topics for a core course.

B.J.K. Kleijn. *The frequentist theory of Bayesian statistics*. Springer Verlag, New York, 2020.

Lucien LeCam. On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Publ. Statist.*, 1:277–329, 1953.

D. V. Lindley and L. D. Phillips. Inference for a Bernoulli process (a Bayesian view). *Amer. Statist.*, 30(3):112–119, 1976. ISSN 0003-1305.

Dennis V. Lindley. The 1988 wald memorial lectures: The present position in bayesian statistics. *Statistical Science*, 5(1):44–65, 1990. ISSN 08834237. URL http://www.jstor.org/stable/2245880.

R. McElreath. *Statistical Rethinking: a Bayesian Course with Examples in R and Stan*. Chapman and Hall–CRC, 2015.

Ronald Meester. Waarom p-waardes niet gebruikt mogen worden als statistisch bewijs. *Nieuw archief voor de wiskunde*, 2019.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. Cambridge, MA, 2012.

Giovanni Parmigiani and Lurdes Y. T. Inoue. *Decision theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2009. ISBN 978-0-471-49657-1. doi: 10.1002/9780470746684. URL http://dx.doi.org/10.1002/9780470746684. Principles and approaches, With contributions by Hedibert F. Lopes.

Christian P. Robert. *The Bayesian choice*. Springer Texts in Statistics. Springer, New York, second edition, 2007. ISBN 978-0-387-71598-8. From decision-theoretic foundations to computational implementation.

Christian P. Robert and George Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004. ISBN 0-387-21239-6. doi: 10.1007/978-1-4757-4145-2. URL http://dx.doi.org/10.1007/978-1-4757-4145-2.

Simo Särkkä. *Bayesian filtering and smoothing*. Number 3. Cambridge university press, 2013.

Leonard J. Savage. The foundations of statistics reconsidered. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 575–586, Berkeley, Calif., 1961. University of California Press. URL http://projecteuclid.org/euclid.bsmsp/1200512183.

Mark J. Schervish. *Theory of statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1995. ISBN 0-387-94546-6. doi: 10.1007/978-1-4612-4250-5. URL http://dx.doi.org/10.1007/978-1-4612-4250-5.

Mark J. Schervish. *P* values: what they are and what they are not. *Amer. Statist.*, 50(3):203–206, 1996. ISSN 0003-1305. doi: 10.2307/2684655. URL http://dx.doi.org/10.2307/2684655.

Jun Shao. *Mathematical statistics*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2003. ISBN 0-387-95382-5. doi: 10.1007/b97553. URL http://dx.doi.org/10.1007/b97553.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pages 197–206. University of California Press, Berkeley and Los Angeles, 1956.

Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. In N. Hjort, C. Holmes, P. Muller, and S. Walker, editors, *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, 2010.

Luke Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1):1–9, 1998. ISSN 1050-5164. doi: 10.1214/aoap/1027961031. URL http://dx.doi.org/10.1214/aoap/1027961031.

A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. ISBN 0-521-49603-9; 0-521-78450-6. doi: 10.1017/CBO9780511802256. URL http://dx.doi.org/10.1017/CBO9780511802256.

Robert L. Winkler. *An introduction to Bayesian inference and decision / Robert L. Winkler*. Holt, Rinehart and Winston New York, 1972. ISBN 0030813271.

G. A. Young and R. L. Smith. *Essentials of statistical inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2005. ISBN 978-0-521-83971-6; 0-521-83971-8. doi: 10.1017/CBO9780511755392. URL http://dx.doi.org/10.1017/CBO9780511755392.

Shelemyahu Zacks. *Examples and problems in mathematical statistics*. John Wiley & Sons, Inc., Hoboken, NJ, 2014. ISBN 978-1-118-60550-9.