Chapter 2

Bayesian basics

In this chapter, we consider the basic definitions and properties of Bayesian inferential and decision-theoretic methods. Naturally the emphasis lies on the posterior distribution, which we derive from the prior based on the subjectivist perspective. However, we also discuss the way prior and posterior should be viewed if one assumes the frequentist point of view. Furthermore, we consider point estimators derived from the posterior, credible sets, testing of hypotheses and Bayesian decision theory. Throughout the chapter, we consider frequentist methods side-by-side with the Bayesian procedures, for comparison and reference.

It should be stressed that the material presented here covers only the most basic Bayesian concepts; further reading is recommended. Various books providing overviews of Bayesian statistics are recommended, depending on the background and interest of the reader: a highly theoretical treatment can be found in Le Cam (1986) [63], which develops a general, mathematical framework for statistics and decision theory, dealing with Bayesian methods as an important area of its application. For a more down-to-earth version of this work, applied only to smooth parametric models, the interested reader is referred to Le Cam and Yang (1990) [64]. The book by Van der Vaart (1998) [83] contains a chapter on Bayesian statistics focusing on the Bernstein-Von Mises theorem (see also section 4.4 in these notes). A general reference of a more decision-theoretic inclination, focusing on Bayesian statistics, is the book by Berger (1985) [8]; a more recent reference of a similar nature is Bernardo and Smith (1993) [13]. Both Berger and Bernardo and Smith devote a great deal of attention to philosophical arguments in favour of the Bayesian approach to statistics, staying rather terse with regard to mathematical detail and focusing almost exclusively on parametric models. Recommendable is also Robert's "The Bayesian choice" (2001) [72], which offers a very useful explanation on computational aspects of Bayesian statistics. Finally, Ripley (1996) [73] discusses Bayesian methods with a very pragmatic focus on pattern classification. The latter reference relates all material with applications in mind but does so based on a firm statistical and decision-theoretic background.

2.1 Bayes' rule, prior and posterior distributions

Formalizing the Bayesian procedure can be done in several ways. We start this section with considerations that are traditionally qualified as being of a "subjectivist" nature, but eventually we revert to the "frequentist" point of view. Concretely this means that we derive an expression for the posterior and prove regularity in the subjectivist framework. In a frequentist setting, this expression is simply used as a definition and properties like regularity and measurability are imposed. Ultimately the philosophical motivation becomes irrelevant from the mathematical point of view, once the posterior and its properties are established.

Perhaps the most elegant (and decidedly subjectivist) Bayesian framework unifies samplespace and model in a product space. Again, the measurement Y is a random variable taking values in a samplespace $\mathscr{Y}$ with σ-algebra $\mathscr{B}$. Contrary to the frequentist case, in the Bayesian approach the model Θ is assumed to be a measurable space as well, with σ-algebra $\mathscr{G}$. The model parameter takes values $\theta \in \Theta$ but is a random variable (denoted ϑ) in this context! We assume that on the product-space $\mathscr{Y} \times \Theta$ (with product σ-algebra $\mathscr{F} = \sigma(\mathscr{B} \times \mathscr{G})$) we have a probability measure

$$\Pi : \sigma(\mathscr{B} \times \mathscr{G}) \to [0,1], \tag{2.1}$$

which is *not* a product measure. The probability measure Π provides a joint probability distribution for (Y, ϑ), where Y is the observation and ϑ (the random variable associated with) the parameter of the model.

Implicitly the choice for the measure Π defines the model in Bayesian context, by the possibility to condition on $\vartheta = \theta$ for some $\theta \in \Theta$. The *conditional distribution* $\Pi_{Y|\vartheta} : \mathscr{B} \times \Theta \to [0,1]$ describes the distribution of the observation Y *given* the parameter ϑ. (For a discussion of conditional probabilities, see appendix A, *e.g.* definition A.6.3 and theorem A.6.1). As such, it defines the elements P_θ of the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$, although the role they play in Bayesian context is slightly different from that in a frequentist setting. The question then arises under which requirements the conditional probability $\Pi_{Y|\vartheta}$ is a so-called regular conditional distribution.

Lemma 2.1.1. *Assume that Θ is a Polish space and that the σ-algebra $\mathscr{G}$ contains the Borel σ-algebra. Let Π be a probability measure, c.f. (2.1). Then the conditional probability $\Pi_{Y|\vartheta}$ is regular.*

Proof The proof is a direct consequence of theorem A.6.1. □

The measures $\Pi_{Y|\vartheta}(\,\cdot\,|\vartheta = \theta)$ form a (Π-almost-sure) version of the elements P_θ of the model $\mathscr{P}$:

$$P_\theta = \Pi_{Y|\vartheta}(\,\cdot\,|\,\vartheta = \theta\,) : \mathscr{B} \to [0,1] \tag{2.2}$$

Consequently, frequentist's notion of a model is only represented up to null-sets of the marginal distribution of ϑ, referred to in Bayesian context as the *prior* for the parameter ϑ.

Definition 2.1.1. *The marginal probability* Π *on* $\mathscr{G}$ *is the prior.*

The prior is interpreted in the subjectivist's philosophy as the "degree of belief" attached to subsets of the model *a priori*, that is, before any observation has been made or incorporated in the calculation. Central in the Bayesian framework is the conditional distribution for ϑ given Y.

Definition 2.1.2. *The conditional distribution*

$$\Pi_{\vartheta|Y} : \mathscr{G} \times \mathscr{Y} \to [0,1], \tag{2.3}$$

is called the posterior distribution.

The posterior is interpreted as a data-amended version of the prior, that is to say, the subjectivist's original "degree of belief", corrected by observation of Y through conditioning, *i.e.* the distribution for ϑ *a posteriori* (that is, after observations have been incorporated).

Assuming that the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure on $\mathscr{Y}$, the above can also be expressed in terms of μ-densities $p_\theta = dP_\theta/d\mu : \mathscr{Y} \to \mathbb{R}$. Using Bayes' rule (*c.f.* lemma A.6.2), we obtain the following expression for the posterior distribution:

$$\Pi(\,\vartheta \in G \,|\, Y\,) = \frac{\displaystyle\int_G p_\theta(Y)\, d\Pi(\theta)}{\displaystyle\int_\Theta p_\theta(Y)\, d\Pi(\theta)}, \tag{2.4}$$

where $G \in \mathscr{G}$ is a measurable subset of the model $\mathscr{P}$. Note that when expressed through (2.4), the posterior distribution can be calculated based on a choice for the model (which specifies p_θ) with a prior Π and the data Y (or a realisation $Y = y$ thereof).

Based on the above definitions, two remarks are in order with regard to the notion of a *model* in Bayesian statistics. First of all, one may choose a large model $\mathscr{P}$, but if for a subset $\mathscr{P}_1 \subset \mathscr{P}$ the prior assigns mass zero, then for all practical purposes $\mathscr{P}_1$ does not play a role, since omission of $\mathscr{P}_1$ from $\mathscr{P}$ does not influence the posterior. As long as the model is parametric, *i.e.* $\Theta \subset \mathbb{R}^d$, we can always use priors that dominate the Lebesgue measure, ensuring that $\mathscr{P}_1$ is a "small" subset of $\mathbb{R}^d$. However, in non-parametric models null-sets of the prior and posterior may be much larger than expected intuitively (for a striking example, see section 4.2, specifically the discussion of Freedman's work).

Example 2.1.1. *Taking the above argument to the extreme, consider a normal location model* $\mathscr{P} = \{N(\theta,1) : \theta \in \mathbb{R}\}$ *with a prior* $\Pi = \delta_{\theta_1}$ *(see example A.2.2), for some* $\theta_1 \in \Theta$*, defined on the Borel* σ*-algebra* $\mathscr{B}$*. Then the posterior takes the form:*

$$\Pi(\vartheta \in A|Y) = \int_A p_\theta(Y)\, d\Pi(\theta) \Big/ \int_\Theta p_\theta(Y)\, d\Pi(\theta) = \frac{p_{\theta_1}(Y)}{p_{\theta_1}(Y)}\Pi(A) = \Pi(A).$$

for any $A \in \mathscr{B}$*. In other words, the posterior* equals *the prior, concentrating all its mass in the point* θ_1*. Even though we started out with a model that suggests estimation of location,* *effectively the model consists of only one point, $\theta_1 \in \Theta$ due to the choice of the prior. In subjectivist terms, the prior belief is fully biased towards θ_1, leaving no room for amendment by the data when we condition to obtain the posterior.*

This example raises the question which part of the model proper $\mathscr{P}$ plays a role. In that respect, it is helpful to make the following definition.

Definition 2.1.3. *In addition to $(\Theta, \mathscr{G}, \Pi)$ being a probability space, let $(\Theta, \mathscr{T})$ be a topological space. Assume that $\mathscr{G}$ contains the Borel σ-algebra $\mathscr{B}$ corresponding to the topology $\mathscr{T}$. The support* supp(Π) *of the prior* Π *is defined as:*

$$\mathrm{supp}(\Pi) = \bigcap \{ G \in \mathscr{G} : G \; closed, \; \Pi(G) = 1 \}.$$

The viability of the above definition is established in the following lemma.

Lemma 2.1.2. *For any topological space Θ with σ-algebra $\mathscr{G}$ that contains the Borel σ-algebra $\mathscr{B}$ and any (prior) probability measure $\Pi : \mathscr{G} \to [0, 1]$, $\mathrm{supp}(\Pi) \in \mathscr{G}$ and $\Pi(\mathrm{supp}(\Pi)) = 1$.*

N ote that supp(Π) is closed, as it is an intersection of closed sets, $\mathrm{supp}(\Pi) \in \mathscr{B} \subset \mathscr{G}$. The proof that the support has measure 1 is left as exercise 2.7. □

In example 2.1.1, the model $\mathscr{P}$ consists of all normal distributions of the form $N(\theta, 1)$, $\theta \in \mathbb{R}$, but the support of the prior supp(Π) equals the singleton $\{N(\theta_1, 1)\} \subset \mathscr{P}$.

Note that the support of the prior is defined based on a topology, the Borel σ-algebra of which must belong to the domain of the prior measure. In parametric models this assumption is rarely problematic, but in non-parametric models finding such a prior may be difficult and the support may be an ill-defined concept. Therefore we may choose to take a less precise but more generally applicable perspective: the model is viewed as the support of the prior Π, but only *up to* Π-*null-sets* (*c.f.* the Π-almost-sure nature of the identification (2.2)). That means that we may add to or remove from the model at will, as long as we make sure that the changes have prior measure equal to zero: the model itself is a Π-almost-sure concept. (Since the Bayesian procedure involves only integration of integrable functions with respect to the prior, adding or removing Π-null-sets to/from the domain of integration will not have unforeseen consequences.)

To many who have been introduced to statistics from the frequentist point of view, including the parameter θ for the model as a random variable ϑ seems somewhat unnatural because the frequentist role of the parameter is entirely different from that of the data. The following example demonstrates that in certain situations the Bayesian point of view is not unnatural at all.

Example 2.1.2. *In the posthumous publication of "An essay towards solving a problem in the doctrine of chances" in 1763 [4], Thomas Bayes included an example of a situation in which the above, subjectivist perspective arises quite naturally. It involves a number of red balls and one white ball placed on a table and has become known in the literature as Bayes' billiard.*

We consider the following experiment: unseen by the statistician, someone places n red balls and one white ball on a billiard table of length 1. Calling the distance between the white ball and the bottom cushion of the table X and the distances between the red balls and the bottom cushion Y_i, $(i = 1, \ldots, n)$, it is known to the statistician that their joint distribution is:

$$(X; Y_1, \ldots, Y_n) \sim U[0,1]^{n+1}, \tag{2.5}$$

i.e. all balls are placed independently with uniform distribution. The statistician will be reported the number K of red balls that is closer to the cushion than the white ball (the data*, denoted Y in the rest of this section) and is asked to give a distribution reflecting his beliefs concerning the position of the white ball X (the* parameter*, denoted ϑ in the rest of this section) based on K. His prior knowledge concerning X (i.e. without knowing the observed value $K = k$) offers little information: the best that can be said is that $X \sim U[0,1]$, the marginal distribution of X, i.e. the prior. The question is how this distribution for X changes when we incorporate the observation $K = k$, that is, when we use the observation to arrive at our* posterior *beliefs based on our* prior *beliefs.*

Since for every i, Y_i and X are independent c.f. (2.5), we have,

$$P(Y_i \leq X | X = x) = P(Y_i \leq x) = x,$$

So for each of the red balls, determining whether it lies closer to the cushion than the white ball amounts to a Bernoulli experiment with parameter x. Since in addition the positions $Y_1, \ldots, Y_n$ are independent, counting the number K of red balls closer to the cushion than the white ball amounts to counting "successes" in a sequence of independent Bernoulli experiments. We conclude that K has a binomial distribution $\mathrm{Bin}(n; x)$*, i.e.*

$$P(K = k | X = x) = \frac{n!}{k!(n-k)!} x^k (1-x)^{n-k}.$$

It is possible to obtain the density for the distribution of X conditional on $K = k$ from the above display using Bayes' rule (c.f. lemma A.6.2):

$$p(x | K = k) = P(K = k | X = x) \frac{p(x)}{P(K = k)}, \tag{2.6}$$

but in order to use it, we need the two marginal densities $p(x)$ and $P(K = k)$ in the fraction. From (2.5) it is known that $p(x) = 1$ and $P(K = k)$ can be obtained by integrating

$$P(K = k) = \int_0^1 P(K = k | X = x)\, p(x)\, dx$$

Substituting in (2.6), we find:

$$p(x | K = k) = \frac{P(K = k | X = x)\, p(x)}{\int_0^1 P(K = k | X = x)\, p(x)\, dx} = B(n,k)\, x^k (1-x)^{n-k}.$$

where $B(n,k)$ is a normalization factor. The x-dependence of the density in the above display reveals that $X | K = k$ is distributed according to a Beta-distribution, $B(k+1, n-k+1)$, so that the normalization factor $B(n,k)$ must equal $B(n,k) = \Gamma(n+2)/\Gamma(k+1)\Gamma(n-k+1)$.

This provides the statistician with distributions reflecting his beliefs concerning the position of the white ball for all possible values k for the observation K. Through conditioning on $K = k$, the prior distribution of X is changed: if a relatively small number of red balls is closer to the cushion than the white ball (i.e. in case k is small compared to n), then the white ball is probably close to the cushion; if k is relatively large, the white ball is probably far from the cushion (see figure 2.1).

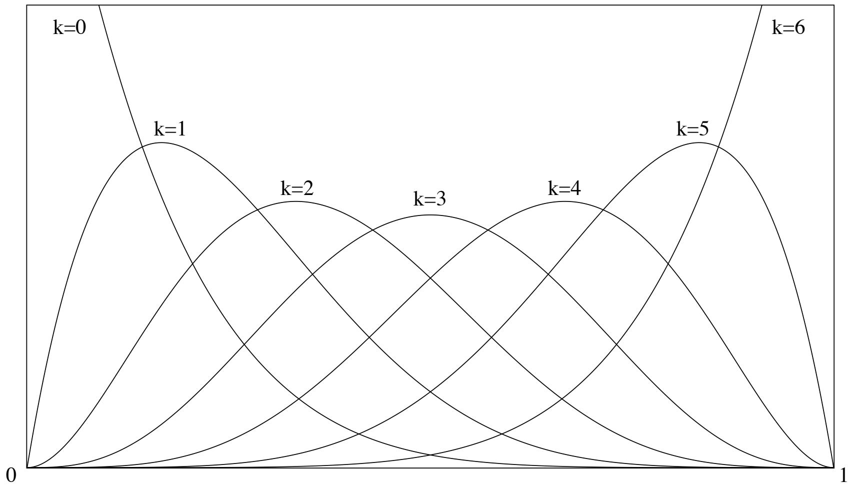

Figure 2.1 Posterior densities for the position X of the white ball, given the number k of red balls closer to the cushion of the billiard (out of a total of $n = 6$ red balls). For the lower values of k, the white ball is close to the cushion with high probability, since otherwise more red balls would probably lie closer to the cushion. This is reflected by the posterior density for $X|K = 1$, for example, by the fact that it concentrates much of its mass close to $x = 0$.

In many experiments or observations, the data consists of a sample of n repeated, stochastically independent measurements of the same quantity. To accommodate this situation formally, we choose $\mathscr{Y}$ equal to the n-fold product of a sample space $\mathscr{X}$ endowed with a σ-algebra $\mathscr{A}$, so that the observation takes the form $Y = (X_1, X_2, \ldots, X_n)$. The additional assumption that the sample is *i.i.d.* (presently a statement concerning the *conditional independence* of the observations given $\vartheta = \theta$) then reads:

$$\Pi_{Y|\vartheta}(X_1 \in A_1, \ldots, X_n \in A_n \,|\, \vartheta = \theta) = \prod_{i=1}^{n} \Pi_{Y|\vartheta}(X_i \in A_i \,|\, \vartheta = \theta) = \prod_{i=1}^{n} P_\theta(X_i \in A_i),$$

for all $(A_1, \ldots, A_n) \in \mathscr{A}^n$. Assuming that the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure μ on $\mathscr{X}$, the above can also be expressed in terms of μ-densities $p_\theta = dP_\theta/d\mu : \mathscr{X} \to \mathbb{R}$. Using Bayes' rule, we obtain the following expression for the posterior distribution:

$$\Pi_n(\,\vartheta \in G \,|\, X_1, X_2, \ldots, X_n\,) = \frac{\displaystyle\int_G \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)}{\displaystyle\int_\Theta \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)}, \tag{2.7}$$

where $G \in \mathscr{G}$ is a measurable subset of the model $\mathscr{P}$.

Remark 2.1.1. *In a dominated model, the Radon-Nikodym derivative (see theorem A.4.2) of the posterior with respect to the prior is the likelihood function, normalized to be a probability density function:*

$$\frac{d\Pi(\,\cdot \,|\, X_1, \ldots, X_n)}{d\Pi}(\theta) = \prod_{i=1}^{n} p_\theta(X_i) \Big/ \int_\Theta \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta), \quad (P_0^n - a.s.). \tag{2.8}$$

The latter fact explains why such strong relations exist (e.g. the Bernstein-Von Mises theorem, theorem 4.4.1) between Bayesian and maximum-likelihood methods. Indeed, the proportionality of the posterior density and the likelihood provides a useful qualitative picture of the posterior as a measure that concentrates on regions in the model where the likelihood is relatively high. This may serve as a direct motivation for the use of Bayesian methods in a frequentist context, c.f. section 1.3. Moreover, this picture gives a qualitative explanation of the asymptotic behaviour of Bayesian methods: under suitable continuity-, differentiability- and tail-conditions, the likelihood remains relatively high in small neighbourhoods of P_0 and drops off steeply outside in the large-sample limit. Hence, if the prior mass in those neighbourhoods is not too small, the posterior concentrates its mass in neighbourhoods of P_0, leading to the asymptotic behaviour described in chapter 4.

Returning to the distributions that play a role in the subjectivist Bayesian formulation, there exists also a marginal for the observation Y.

Definition 2.1.4. *The distribution $P^\Pi : \mathscr{B} \to [0,1]$ defined by*

$$P_n^\Pi(\,X_1 \in A_1, \ldots, X_n \in A_n\,) = \int_\Theta \prod_{i=1}^{n} P_\theta(A_i)\, d\Pi(\theta) \tag{2.9}$$

is called the prior predictive distribution.

Strictly speaking the prior predictive distribution describes a subjectivist's expectations concerning the observations $X_1, X_2, \ldots, X_n$ based only on the prior Π, *i.e.* without involving the data. More readily interpretable is the following definition.

Definition 2.1.5. *For given $n, m \geq 1$, the distribution $P_{n,m}^\Pi$ defined by*

$$P_{n,m}^\Pi(\,X_{n+1} \in A_{n+1}, \ldots, X_{n+m} \in A_{n+m} \,|\, X_1, \ldots, X_n\,) = \int_\Theta \prod_{i=1}^{m} P_\theta(A_{n+i})\, d\Pi(\theta \,|\, X_1, \ldots, X_n\,))$$

is called the posterior predictive distribution.

The prior predictive distribution is subject to correction by observation through substitution of the prior by the posterior: the resulting posterior predictive distribution is interpreted as the Bayesian's expectation concerning the distribution of the observations $X_{n+1}, X_{n+2}, \ldots, X_{n+m}$ given the observations $X_1, X_2, \ldots, X_n$ and the prior Π.

Remark 2.1.2. *The form of the prior predictive distribution is the subject of de Finetti's theorem (see theorem A.2.2), which says that the distribution of a sequence* $(X_1, \ldots, X_n)$ *of random variables is of the form on the r.h.s. of the above display (with uniquely determined prior* Π*) if and only if the sample* $(X_1, \ldots, X_n)$ *is exchangeable, that is, if and only if the joint distribution for* $(X_1, \ldots, X_n)$ *equals that of* $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ *for all permutations* π *of* n *elements.*

Remark 2.1.3. *We conclude the discussion of the distributions that play a role in Bayesian statistics with the following important point: at no stage during the derivation above, was an "underlying distribution of the data" used or needed! For comparison we turn to assumption (1.1), which is fundamental in the frequentist approach. More precisely, the assumption preceding (2.1) (c.f. the subjectivist Bayesian approach) is at odds with (1.1), unless*

$$P_0^n = P_n^\Pi = \int_\Theta P_\theta^n \, d\Pi(\theta),$$

Note, however, that the l.h.s. is a product-measure, whereas on the r.h.s. only exchangeability is guaranteed! (Indeed, the equality in the above display may be used as the starting point for definition of a goodness-of-fit criterion for the model and prior (see section 3.3). The discrepancy in the previous display makes the "pure" (e.g. subjectivist) Bayesian reluctant to assume the existence of a distribution P_0 *for the sample.)*

The distribution P_0 could not play a role in our analysis if we did not choose to adopt assumption (1.1). In many cases we shall assume that Y contains an *i.i.d.* sample of observations $X_1, X_2, \ldots, X_n$ where $X \sim P_0$ (so that $Y \sim P_0^n$). Indeed, if we would not make this assumption, asymptotic considerations like those in chapter 4 would be meaningless. However, adopting (1.1) leaves us with questions concerning the background of the quantities defined in this section because they originate from the subjectivist Bayesian framework.

Remark 2.1.4. (Bayesian/frequentist hybrid approach) *Maintaining the frequentist assumption that* $Y \sim P_0$ *for some* P_0 *requires that we revise our approach slightly: throughout the rest of these lecture notes, we shall assume (1.1) and require the model* $\mathscr{P}$ *to be a probability space* $(\mathscr{P}, \mathscr{G}, \Pi)$ *with a probability measure* Π *referred to as the prior. So the prior is introduced as a measure on the model, rather than emergent as a marginal to a product-space measure. Model and sample space are left in the separate roles they are assigned by the frequentist. We then proceed to* define *the posterior by expression (2.7). Regularity of the posterior is* imposed *(for a more detailed analysis, see Schervish (1995) [75] and Barron, Schervish and Wasserman (1999) [7]). In that way, we combine a frequentist perspective on statistics with Bayesian* *methodology: we make use of Bayesian quantities like prior and posterior, but analyze them from a frequentist perspective.*

Remark 2.1.5. *In places throughout these lecture notes, probability measures P are decomposed into a P_0-absolutely-continuous part $P_{\|}$ and a P_0-singular part $P_{\perp}$. Following Le Cam, we use the convention that if P is not dominated by P_0, the Radon-Nikodym derivative refers to the P_0-absolutely-continuous part only: $dP/dP_0 = dP_{\|}/dP_0$. (See theorem A.4.2.) With this in mind, we write the posterior as follows*

$$\Pi(\,\vartheta \in A \,|\, X_1, X_2, \ldots, X_n\,) = \frac{\displaystyle\int_A \prod_{i=1}^n \frac{dP_\theta}{dP_0}(X_i)\, d\Pi(\theta)}{\displaystyle\int_\Theta \prod_{i=1}^n \frac{dP_\theta}{dP_0}(X_i)\, d\Pi(\theta)}, \quad (P_0^n - a.s.) \tag{2.10}$$

Since the data $X_1, X_2, \ldots$ are i.i.d.-P_0-distributed, the P_0-almost-sure version of the posterior in the above display suffices. Alternatively, any σ-finite measure that dominates P_0 may be used instead of P_0 in (2.10) while keeping the definition P_0^n-almost-sure. Such P_0-almost sure representations are often convenient when deriving proofs.

In cases where the model is not dominated, (2.10) may be used as the definition of the posterior measure but there is no guarantee that (2.10) leads to sensible results!

Example 2.1.3. *Suppose that the samplespace is $\mathbb{R}$ and the model $\mathscr{P}$ consists of all measures of the form (see example A.2.2):*

$$P = \sum_{j=1}^m \alpha_j \delta_{x_j}, \tag{2.11}$$

for some $m \geq 1$, with $\alpha_1, \ldots, \alpha_m$ satisfying $\alpha_j \geq 0$, $\sum_{j=1}^m \alpha_j = 1$ and $x_1, \ldots, x_m \in \mathbb{R}$. A suitable prior for this model exists: distributions drawn from the so-called Dirichlet process prior are of the form (2.11) with (prior) probability one. There is no σ-finite dominating measure for this model and hence the model can not be represented by a family of densities, c.f. definition 1.1.3. In addition, if the true distribution P_0 for the observation is also a convex combination of Dirac measures, distributions in the model are singular with respect to P_0 unless they happen to have support-points in common with P_0. Consequently definition (2.10) does not give sensible results in this case. We have to resort to the subjectivist definition (2.3) in order to make sense of the posterior distribution.

To summarize, the Bayesian procedure consists of the following steps

(i) Based on the background of the data Y, the statistician chooses a model $\mathscr{P}$, usually with some measurable parameterization $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$.

(ii) A prior measure Π on $\mathscr{P}$ is chosen, based either on subjectivist or objectivist criteria. Usually a measure on Θ is defined, inducing a measure on $\mathscr{P}$.

(iii) Based on (2.3), (2.4) or in the case of an *i.i.d.* sample $Y = (X_1, X_2, \ldots, X_n)$, on:

$$d\Pi_n(\,\theta \,|\, X_1, X_2, \ldots, X_n\,) = \frac{\displaystyle\prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)}{\displaystyle\int_\Theta \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)},$$

we calculate the posterior density or posterior as a function of the data.

(iv) We observe a realization of the data $Y = y$ and use it to calculate a realisation of the posterior.

The statistician may then infer properties of the parameter θ from the posterior $\Pi(\,\cdot\,|\,Y = y\,)$, giving them a subjectivist or objectivist interpretation. One important point: when reporting the results of any statistical procedure, one is obliged to also reveal all relevant details concerning the procedure followed and the data. So in addition to inference on θ, the statistician should report on the nature and size of the sample used and, in the Bayesian case, should always report choice of model and prior as well, with a clear motivation.

2.2 Bayesian point estimators

When considering questions of statistical estimation, the outcome of a frequentist procedure is of a different nature than the outcome of a Bayesian procedure: a point-estimator (the frequentist outcome) is a point in the model, whereas the posterior is a distribution on the model. A first question, then, concerns the manner in which to compare the two. The connection between Bayesian procedures and frequentist (point-)estimation methods is provided by point-estimators derived from the posterior, called Bayesian point-estimators. Needless to say, comparison of frequentist and Bayesian point-estimators requires that we assume the "hybrid" point of view presented in remark 2.1.4.

We think of a reasonable Bayesian point-estimators as a point in the model around which posterior mass is accumulated most, a point around which the posterior distribution is concentrated in some way. As such, any reasonable Bayesian point-estimator should represent the *location* of the posterior distribution. But as is well known from probability theory, there is no unique definition of the "location" of a distribution. Accordingly, there are many different ways to define Bayesian point-estimators.

Remark 2.2.1. *Arguably, there are distributions for which even the* existence *of a "location" is questionable. For instance, consider the convex combination of point-masses* $P = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1}$ *on* $(\mathbb{R}, \mathscr{B})$*. Reasonable definitions of location, like the mean and the median of* P*, all assign as the location of* P *the point* $0 \in \mathbb{R}$*. Yet small neighbourhoods of* 0 *do not receive* any *P-mass, so* 0 *can hardly be viewed as a point around which* P *concentrates its mass. The intuitition of a distribution's location can be made concrete without complications of the above* *nature, if we restrict attention to unimodal distributions. However, it is common practice to formulate the notion for all distributions by the same definitions.*

One quantity that is often used to represent a distribution's location is its expectation. This motivates the first definition of a Bayesian point-estimator: the posterior mean.

Definition 2.2.1. *Consider a statistical problem involving data Y taking values in a samplespace $(\mathscr{Y},\mathscr{B})$ and a model $(\mathscr{P},\mathscr{G})$ with prior Π. Assume that the maps $P \mapsto P(B)$, $(B \in \mathscr{B})$ are measurable with respect to $\mathscr{G}$ and that the posterior $\Pi(\,\cdot\,|Y)$ is regular, P_0^n-almost-surely. The posterior mean (or posterior expectation) is a probability measure $\hat{P} : \mathscr{B} \to [0,1]$, defined*

$$\hat{P}(B) = \int_{\mathscr{P}} P(B)\; d\Pi(\,P\,|\,Y\,), \tag{2.12}$$

P_0-almost-surely, for every event $B \in \mathscr{B}$.

Remark 2.2.2. *In order to justify the above definition, we have to show that $\hat{P}$ is a probability measure, P_0-almost-surely. Since the posterior is a regular conditional distribution, the map $B \mapsto \hat{P}(B)$ is defined P_0-almost-surely. Obviously, for all $B \in \mathscr{B}$, $0 \le \hat{P}(B) \le 1$. Let $(B_i)_{i\ge1} \subset \mathscr{B}$ be a sequence of disjoint events. Since $(P,i) \mapsto P(B_i)$ is non-negative and measurable, Fubini's theorem applies in the third equality below:*

$$\begin{aligned}\hat{P}\Bigl(\bigcup_{i\ge1} B_i\Bigr) &= \int_{\mathscr{P}} P\Bigl(\bigcup_{i\ge1} B_i\Bigr)\, d\Pi(\,P\,|\,Y\,) = \int_{\mathscr{P}} \sum_{i\ge1} P(B_i)\, d\Pi(\,P\,|\,Y\,)\\ &= \sum_{i\ge1} \int_{\mathscr{P}} P(B_i)\, d\Pi(\,P\,|\,Y\,) = \sum_{i\ge1} \hat{P}(B_i),\end{aligned}$$

which proves σ-additivity of $\hat{P}$.

Remark 2.2.3. *Note that, unless $\mathscr{P}$ happens to be convex, $\hat{P} \in \mathscr{P}$ is* not *guaranteed! In other words, the posterior mean may lie outside the model!*

In many practical situations, the model $\mathscr{P}$ is parametric with parameterization $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$. In that case a different definition of the posterior mean can be made.

Definition 2.2.2. *Let $\mathscr{P}$ be a model parameterized by a convex subset Θ of $\mathbb{R}^d$. Let Π be a prior defined on Θ. If ϑ is integrable with respect to the posterior, the parametric posterior mean is defined*

$$\hat{\theta}_1(Y) = \int_{\Theta} \theta\; d\Pi(\,\theta\,|\,Y\,) \in \Theta, \tag{2.13}$$

P_0^n-almost-surely.

Remark 2.2.4. *The distinction between the posterior mean and the* parametric *posterior mean, as made above, is non-standard: it is customary in the Bayesian literature to refer to either as "the posterior mean". See, however, example 2.2.1.*

In definition 2.2.2, convexity of Θ is a condition (instead of an afterthought, as with definition 2.2.1): if Θ is not convex there is no guarantee that $\hat{\theta}_1 \in \Theta$, in which case $P_{\hat{\theta}_1}$ is not defined since $\hat{\theta}_1$ does not lie in the domain of the parameterization. Definition 2.2.2 can be extended to non-parametric models, *i.e.* models with an infinite-dimensional Θ. In that case, regularity of the posterior reappears as a condition and the condition of "integrability" of ϑ requires further specification.

It is tempting to assume that there is no difference between the posterior mean and the parametric posterior mean if the model is parametric and priors are brought in correspondence. This is not the case, however, as demonstrated by the following (counter)example.

Example 2.2.1. *Consider a normal location model in two dimensions for an observation Y, where the location $\mu \in \mathbb{R}^2$ lies on the unit circle and the covariance Σ is fixed and known:*

$$\mathscr{P} = \bigl\{ P_\theta = N(\mu(\theta), \Sigma) \,:\, \mu(\theta) = (\cos\theta, \sin\theta),\, \theta \in [0, 2\pi) \bigr\}.$$

This is an identifiable, one-dimensional parametric model with convex parameterizing space $\Theta = [0, 2\pi)$. Assume that Ξ is the uniform distribution on Θ (Ξ plays the role of the posterior; it does not matter what shape the posterior really has, all we need is a counterexample). We define the corresponding measure Ξ' on $\mathscr{P}$ by applying Ξ to the pre-image of the parameterization. By rotational symmetry of Ξ and Fubini's theorem, the expectation of Y under $\hat{P}$ is

$$\int Y \, d\hat{P} = \int_{\mathscr{P}} PY \, d\Xi'(P) = \int_{\Theta} P_\theta Y \, d\Xi(\theta) = \frac{1}{2\pi} \int_0^{2\pi} \mu(\theta) \, d\theta = (0, 0).$$

Note that none of the distributions in $\mathscr{P}$ has the origin as its expectation. We can also calculate the expectation of Y under $P_{\hat{\theta}}$ in this situation:

$$\hat{\theta}_1(Y) = \int_{\Theta} \theta \, d\Xi(\theta) = \frac{1}{2\pi} \int_0^{2\pi} \theta \, d\theta = \pi,$$

which leads to $P_{\hat{\theta}} Y = P_\pi Y = (-1, 0)$. Clearly, the posterior mean does not equal the point in the model corresponding to the parametric posterior mean. In fact, we see from the above that $\hat{P} \notin \mathscr{P}$.

The fact that the expectations of $\hat{P}$ and $P_{\hat{\theta}}$ in example 2.2.1 differ makes it clear that

$$\hat{P} \neq P_{\hat{\theta}},$$

unless special circumstances apply: if we consider a parameterization $\theta \mapsto P_\theta$ from a (closed, convex) parameterizing space Θ with posterior measure $\Pi(d\theta)$ onto a space of probability measures $\mathscr{P}$ (with induced posterior $\Pi(dP)$), it makes a difference whether we consider the posterior mean as defined in (2.12), or calculate $P_{\hat{\theta}}$. The parametric posterior mean $P_{\hat{\theta}}$ lies in the model $\mathscr{P}$; $\hat{P}$ lies in the closed convex hull $\overline{\mathrm{co}}(\mathscr{P})$ of the model, but not necessarily $\hat{P} \in \mathscr{P}$.

Since there are multiple ways of defining the location of a distribution, there are more ways of obtaining point-estimators from the posterior distribution. For example in a one-dimensional parametric model, we can consider the *posterior median* defined by

$$\tilde{\theta}(Y) = \inf\bigl\{s \in \mathbb{R} \,:\, \Pi(\vartheta \leq s|Y) \geq 1/2\bigr\},$$

i.e. the smallest value for θ such that the posterior mass to its left is greater than or equal to 1/2. (Note that this definition simplifies in case the posterior has a continuous, strictly monotone distribution function: in that case the median equals the (unique) point where this distribution function equals 1/2.) More generally, we consider the following class of point-estimators [63].

Definition 2.2.3. *Let $\mathscr{P}$ be a model with metric $d : \mathscr{P} \times \mathscr{P} \to \mathbb{R}$ and a prior Π on $\mathscr{G}$ containing the Borel σ-algebra corresponding to the metric topology on $\mathscr{P}$. Let $\ell : \mathbb{R} \to \mathbb{R}$ be a convex loss-function $\ell : \mathbb{R} \to \mathbb{R}$. The* formal Bayes estimator *is the minimizer of the function:*

$$\mathscr{P} \to \mathbb{R} \,:\, P \mapsto \int_{\mathscr{P}} \ell(d(P,Q))\, d\Pi(\, Q \,|\, Y\,),$$

over the model $\mathscr{P}$ (provided that such a minimizer exists and is unique).

The heuristic idea behind formal Bayes estimators is decision-theoretic (see section 2.4). Ideally, one would like to estimate by a point P in $\mathscr{P}$ such that $\ell(d(P,P_0))$ is minimal; if $P_0 \in \mathscr{P}$, this would lead to $P = P_0$. However, lacking specific knowledge on P_0, we choose to represent it by averaging over $\mathscr{P}$ weighted by the posterior, leading to the notion in definition 2.2.3. Another useful point estimator based on the posterior is defined as follows.

Definition 2.2.4. *Let the data Y with model $\mathscr{P}$, metric d and prior Π be given. Suppose that the σ-algebra on which Π is defined contains the Borel σ-algebra generated by the metric topology. For given $\epsilon > 0$, the small-ball estimator is defined to be the maximizer of the function*

$$P \mapsto \Pi(\, B_d(P,\epsilon) \,|\, Y\,), \tag{2.14}$$

over the model, where $B_d(P,\epsilon)$ is the d-ball in $\mathscr{P}$ of radius ϵ centred on P (provided that such a maximizer exists and is unique).

Remark 2.2.5. *Similarly to definition 2.2.4, for a fixed value p such that $1/2 < p < 1$, we may define a Bayesian point estimator as the centre point of the smallest d-ball with posterior mass greater than or equal to p (if it exists and is unique (see also, exercise 2.6)).*

If the posterior is dominated by a σ-finite measure μ, the posterior density with respect to μ can be used as a basis for defining Bayesian point estimators.

Definition 2.2.5. *Let $\mathscr{P}$ be a model with prior Π. Assume that the posterior is absolutely continuous with respect to a σ-finite measure μ on $\mathscr{P}$. Denote the μ-density of $\Pi(\,\cdot\,|Y)$ by $\theta \mapsto \pi(\theta|Y)$. The maximum-a-posteriori estimator (or MAP-estimator, or posterior mode)*

$\hat{\theta}_2$ for θ is defined as the point in the model where the posterior density takes on its maximal value (provided that such a point exists and is unique):

$$\pi(\hat{\theta}_2|Y) = \sup_{\theta\in\Theta} \pi(\theta|Y). \tag{2.15}$$

Remark 2.2.6. *The MAP-estimator has a serious weak point: a different choice of dominating measure μ leads to a different MAP estimator! A MAP-estimator is therefore unspecified unless we specify also the dominating measure used to obtain a posterior density. It is with respect to this dominating measure that we define our estimator, so a motivation for the dominating measure used is inherently necessary (and often conspicuously lacking). Often the Lebesgue measure is used without further comment, or objective measures (see section 3.2) are used. Another option is to use the prior measure as the dominating measure, in which case the MAP estimator equals the maximum-likelihood estimator (see remark 2.2.7).*

All Bayesian point estimators defined above as maximizers or minimizers over the model suffer from the usual existence and uniqueness issues associated with extrema. However, there are straightforward methods to overcome such issues. We illustrate using the MAP-estimator. Questions concerning the existence and uniqueness of MAP-estimators should be compared to those of the existence and uniqueness of M-estimators in frequentist statistics. Although it is hard to formulate conditions of a general nature to guarantee that the MAP-estimator exists, often one can use the following lemma to guarantee existence.

Lemma 2.2.1. *Consider a parameterized model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ If the Θ is compact*[1] *and the posterior density $\theta \mapsto \pi(\theta|Y)$ is upper-semi-continuous (P_0^n-almost-surely) then the posterior density takes on its maximum in some point in Θ, P_0^n-almost-surely.*

To prove uniqueness, one has to be aware of various possible problems, among which are identifiability of the model (see section 1.1, in particular definition 1.1.5). Considerations like this are closely related to matters of consistency of M-estimators, *e.g.* Wald's consistency conditions for the maximum-likelihood estimator. The crucial property is called *well-separatedness* of the maximum, which says that outside neighbourhoods of the maximum, the posterior density must be uniformly strictly below the maximum. The interested reader is referred to chapter 5 of van der Vaart (1998) [83], *e.g.* theorems 5.7 and 5.8.

Remark 2.2.7. *There is an interesting connection between (Bayesian) MAP-estimation and (frequentist) maximum-likelihood estimation. Referring to formula (2.7) we see that in an i.i.d. experiment with parametric model, the MAP-estimator maximizes:*

$$\Theta \to \mathbb{R} : \theta \mapsto \prod_{i=1}^{n} p_\theta(X_i)\,\pi(\theta),$$

[1] Compactness of the model is a requirement that may be unrealistic or mathematically inconvenient in many statistical problems, especially when the model is non-parametric. However in a Bayesian context Ulam's theorem (see theorem A.2.3) offers a way to relax this condition.

where it is assumed that the model is dominated and that the prior has a density π with respect to the Lebesgue measure μ. If the prior had been uniform, the last factor would have dropped out and maximization of the posterior density is *maximization of the likelihood. Therefore, differences between ML and MAP estimators are entirely due to non-uniformity of the prior. Subjectivist interpretation aside, prior non-uniformity has an interpretation in the frequentist setting as well, through what is called penalized maximum likelihood estimation (see, Van de Geer (2000) [39]): Bayes' rule (see lemma A.6.2) applied to the posterior density $\pi_n(\theta|X_1,\ldots,X_n)$ gives:*

$$\log \pi_n(\theta|X_1,\ldots,X_n) = \log \pi_n(X_1,\ldots,X_n|\theta) + \log \pi(\theta) + D(X_1,\ldots,X_n),$$

where D is a (θ-independent, but stochastic) normalization constant. The first term equals the log-likelihood and the logarithm of the prior plays the role of a penalty term when maximizing over θ. Hence, maximizing the posterior density over the model Θ can be identified with maximization of a penalized likelihood over Θ. So defining a penalized MLE $\hat{\theta}_n$ with the logarithm of the prior density $\theta \mapsto \log \pi(\theta)$ in the role of the penalty, the MAP-estimator coincides with $\hat{\theta}_n$. The above offers a direct connection between Bayesian and frequentist methods of point-estimation. As such, it provides an frequentist interpretation of the prior as a penalty in the ML procedure. The asymptotic behaviour of the MAP-estimator is discussed in chapter 4 (see theorem 4.4.2).

2.3 Credible sets and Bayes factors

Besides point-estimation, frequentist statistics has several other inferential techniques at its disposal. The two most prominent are the analysis of confidence intervals and the testing of statistical hypotheses. Presently, it is assumed that the reader is familiar with these methods, but the essential reasoning is summarized for reference and comparison. The goal of this section is to formulate Bayesian analogs, so-called credible sets and Bayes factors respectively, and to compare them with aforementioned frequentist techniques.

Before we consider the Bayesian definitions, we briefly review the frequentist procedures. We assume that we have data Y and a parameterized model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ such that $Y \sim P_{\theta_0}$ for some $\theta_0 \in \Theta$. For simplicity, we assume that $\Theta \subset \mathbb{R}$ whenever the dimension of Θ is of importance.

We start with the central ideas and definitions that play a role in the Neyman-Pearson approach to statistical hypothesis testing. In this context, the hypotheses are presumptions one can make concerning the distribution of the data. Since the model contains all distributions the statistician is willing to consider as possibilities for P_0, the hypotheses are formulated in terms of a partition of the model (or its parameterization) into two disjoint subsets. One of them corresponds to the so-called null hypothesis and the other to the alternative hypothesis, which do *not* play a symmetric role in the Neyman-Pearson procedure. The goal of

Neyman-Pearson hypothesis testing is to find out whether or not the data contains "enough" evidence to reject the null hypothesis as a likely explanation when compared to alternative explanations. Sufficiency of evidence is formulated in terms of statistical significance.

For simplicity, we consider a so-called simple null hypothesis (*i.e.* a hypothesis consisting of only one point in the model, which is assumed to be identifiable): let a certain point $\theta_1 \in \Theta$ be given and consider the hypotheses:

$$H_0: \quad \theta_0 = \theta_1, \qquad H_1: \quad \theta_0 \neq \theta_1,$$

where H_0 denotes the null-hypothesis and H_1 the alternative. By no means does frequentist hypothesis testing equate to the corresponding *classification* problem, in which one would treat H_0 and H_1 symmetrically and make a choice for one or the other based on the data (for more on frequentist and Bayesian classification, see section 2.4).

To assess both hypotheses using the data, the simplest version of the Neyman-Pearson method of hypothesis testing seeks to find a test-statistic $T(Y) \in \mathbb{R}$ displaying different behaviour depending on whether the data Y is distributed according to (a distribution in) H_0 or in H_1. To make this distinction more precise, we define a so-called *critical region* $K \subset \mathbb{R}$, such that $P_{\theta_1}(T \in K)$ is "small" and $P_\theta(T \notin K)$ is "small" for all $\theta \neq \theta_1$. What we mean by "small" probabilities in this context is a *choice* for the statistician, a so-called significance level α is to be chosen to determine when these probabilities are deemed "small". That way, upon realization $Y = y$, a distribution in the hypothesis H_0 makes an outcome $T(y) \in K$ improbable compared to H_1.

Definition 2.3.1. *Let $\Theta \to \mathscr{P} : \theta \to P_\theta$ be a parameterized model for a sample Y. Formulate two hypotheses H_0 and H_1 by introducing a two-set partition $\{\Theta_0, \Theta_1\}$ of the model Θ:*

$$H_0: \quad \theta_0 \in \Theta_0, \qquad H_1: \quad \theta_0 \in \Theta_1.$$

We say that a test for these hypotheses based on a test-statistic T with critical region K is of level $\alpha \in (0,1)$ if the power function $\pi : \Theta \to [0,1]$, defined by

$$\pi(\theta) = P_\theta\big(T(Y) \in K\big),$$

is uniformly small over Θ_0:

$$\sup_{\theta \in \Theta_0} \pi(\theta) \leq \alpha. \tag{2.16}$$

From the above definition we arrive at the conclusion that if $Y = y$ and $T(y) \in K$, hypothesis H_0 is improbable enough to be rejected, since H_0 forms an "unlikely" explanation of observed data (at said significance level). The degree of "unlikeliness" can be quantified in terms of the so-called *p-value*, which is the lowest significance level at which the realised value of the test statistic $T(y)$ would have led us to reject H_0. Of course there is the possibility that our decision is wrong and H_0 is actually true but $T(y) \in K$ nevertheless, so that our rejection of the null hypothesis is unwarranted. This is called a *type-I error*; a *type-II error* is made when we do *not* reject H_0 while H_0 is not true. The significance level α thus represents a fixed upper-bound for the probability of a type-I error. Having found a collection of tests displaying the chosen significance level, the Neyman-Pearson approach calls for subsequent minimization of the Type-II error probability, *i.e.* of all the pairs (T, K) satisfying (2.16), one prefers a pair that minimizes $P_\theta(T(Y) \notin K)$, ideally uniformly in $\theta \in \Theta_1$. However, generically such uniformly most-powerful tests do not exist due to the possibility that different (T, K) pairs are most powerful over distinct subsets of the alternative. The famed Neyman-Pearson lemma [60] asserts that a most powerful test exists in the case Θ contains only two points and can be extended to obtain uniformly most powerful tests in certain models.

We consider the Neyman-Pearson approach to testing in some more detail in the following example while also extending the argument to the asymptotic regime. Here Y is an *i.i.d.* sample and the test-statistic and critical region are dependent on the size n of this sample. We investigate the behaviour of the procedure in the limit $n \to \infty$.

Example 2.3.1. *Suppose that the data Y forms an i.i.d. sample from a distribution $P_0 = P_{\theta_0}$ and that $P_\theta X = \theta$ for all $\theta \in \Theta$. Moreover, assume that $P_\theta X^2 < \infty$ for all θ. Due to the law of large numbers, the sample-average*

$$T_n(X_1, \ldots, X_n) = \mathbb{P}_n X,$$

converges to θ under P_θ (for all $\theta \in \Theta$) and seems a suitable candidate for the test-statistic, at least in the regime where the sample-size n is large (i.e. asymptotically). The central limit theorem allows us to analyze matters in greater detail, for all $s \in \mathbb{R}$:

$$P_\theta^n\bigl(\mathbb{G}_n X \leq s\sigma(\theta)\bigr) \to \Phi(s), \qquad (n \to \infty). \tag{2.17}$$

where $\sigma(\theta)$ denotes the standard deviation of X under P_θ. For simplicity, we assume that $\theta \mapsto \sigma(\theta)$ is a known quantity in this derivation. The limit (2.17) implies that

$$P_\theta^n\bigl(T_n(X_1, \ldots, X_n) \leq \theta + n^{-1/2}\sigma(\theta)\, s\bigr) \to \Phi(s), \qquad (n \to \infty).$$

Assuming that H_0 holds, i.e. that $\theta_0 = \theta_1$, we then find that, given an asymptotic significance level $\alpha \in (0, 1)$ and with the standard-normal quantiles denoted s_α,

$$P_0^n\bigl(T_n(X_1, \ldots, X_2) \leq \theta_1 + n^{-1/2}\sigma(\theta_1)s_{\alpha/2}\bigr) \to 1 - \tfrac{1}{2}\alpha,$$

For significance levels close to zero, we see that under the null-hypothesis, it is improbable to observe $T_n > \theta_1 + n^{-1/2}\sigma(\theta_1)s_{\alpha/2}$. It is equally improbable to observe $T_n < \theta_1 - n^{-1/2}\sigma s_{\alpha/2}$, which means that we can take as our critical region $K_{n,\alpha}$

$$K_{n,\alpha} = \mathbb{R} \setminus [\theta_1 - n^{-1/2}\sigma(\theta_1)s_{\alpha/2}, \theta_1 + n^{-1/2}\sigma(\theta_1)s_{\alpha/2}],$$

(Note that this choice for the critical region is not unique unless we impose that it be an interval located symmetrically around θ_1.) Then we are in a position to formulate our decision on the null hypothesis, to reject H_0 or not:

(i) if $T_n \in K_{n,\alpha}$, we reject H_0 at significance level α, and,

(ii) if $T_n \notin K_{n,\alpha}$, we do not see enough evidence in the data to reject H_0 at significance level α.

Beware of a very common philosophical pitfall in the last case: even under case (ii), we do not *draw the conclusion that H_0 is accepted. The data does not provide enough evidence to reject the null hypothesis, but that does not imply that we have enough evidence to accept it!*

Note the behaviour of the procedure with varying sample-size: keeping the significance level fixed, the width of the critical regions $K_{n,\alpha}$ is of order $O(n^{-1/2})$, so smaller and smaller critical regions can be used as more information concerning the distribution P_0 (read, data) comes available. Similarly, if instead we keep the critical region fixed, the probability for a Type-I error (sometimes called the p-value if no fixed significance level is used) decreases with growing sample-size.

Strictly speaking the reasoning we follow here is not exact, because in practice n is finite and we are using a criterion based on the limit $n \to \infty$. At any finite n, the distribution of $\mathbb{G}_n X$ may not be close to $N(0,\sigma^2)$. In general we do not know which minimal sample-size n should be used in order for these distributions to be "sufficiently close". Nevertheless, it is common practice to use asymptotic tests like this one, in cases where the sample-size is deemed to be large enough and the test-statistic is expected to assume its asymptotic behaviour behaviour in close approximation.

It is important to stress that our criterion for tests is entirely geared at minimizing the probability of rejecting H_0 when in fact H_0 contains the true distribution. As such, the testing procedure we follow can only lead to one definite conclusion, *rejection* of the null hypothesis. The inverse conclusion, *acceptance* of the null hypothesis, is never the result. Therefore, it is crucial that we choose the null hypothesis to be an assertion that we would like to disprove. In practice, one also tries to find a test such that the probability of *not* rejecting H_0 when it is *not* valid is also small. Before we formalize the latter, we generalize the concepts introduced above in this section somewhat.

Note that the indicators for the events $\{Y \in \mathscr{Y} : T_n(Y) \in K_n\}$ form a (bounded, positive) sequence of random variables, on which we base the decision to reject H_0 or not. The power functions $\pi_n : \Theta \to [0,1]$ are simply the P_θ-expectations of these random variables.

Definition 2.3.2. *Let $\mathscr{P}$ be a model for a sample $X_1, X_2 \ldots$ taking values in $\mathscr{X}$ and assume that the true distribution of the data lies in the model, $(X_1, X_2, \ldots) \sim P_0 \in \mathscr{P}$. Formulate two hypotheses H_0 and H_1 by introducing a two-set partition $\{\mathscr{P}_0, \mathscr{P}_1\}$ of the model $\mathscr{P}$:*

$$H_0: \quad P_0 \in \mathscr{P}_0, \qquad H_1: \quad P_0 \in \mathscr{P}_1.$$

A test sequence $(\phi_n)_{n\geq 1}$ is a sequence of statistics $\phi_n : \mathscr{X}^n \to [0,1]$, (for all $n \geq 1$). An asymptotic test is defined as a criterion for the decision to reject H_0 or not, based on (a realization of) $\phi_n(X_1, \ldots, X_n)$ and is studied in the limit $n \to \infty$.

An example of such a criterion is the procedure given in definition 2.3.1 and example 2.3.1, where test-functions take on the values zero or one depending on the (realized) test-statistic and the critical region. When we replace indicators by test functions as in definition 2.3.2 criteria may vary depending on the nature of the test functions used.

Definition 2.3.3. *Extending definition 2.3.2, we define the power function sequence of the test sequence* (ϕ_n) *as a map* $\pi_n : \mathscr{P} \to [0,1]$ *on the model defined by:*

$$\pi_n(P) = P\phi_n.$$

Like in definition 2.3.1, the quality of the test depends on the behaviour of the power sequence on $\mathscr{P}_0$ and $\mathscr{P}_1$ respectively. If we are interested exclusively in rejection of the null hypothesis, we could reason like in definition 2.3.1 and set a significance level α to select only those test sequences that satisfy

$$\sup_{P\in\mathscr{P}_0} \pi_n(P) \le \alpha.$$

Subsequently, we prefer test sequences that have high power on the alternative. For example, if we have two test sequences (ϕ_n) and (ψ_n) and a point $P \in \mathscr{P}_1$ such that

$$\lim_{n\to\infty} P\phi_n \ge \lim_{n\to\infty} P\psi_n, \tag{2.18}$$

then (ϕ_n) is said to be asymptotically more powerful than (ψ_n) at P. If (2.18) holds for *all* $P \in \mathscr{P}_1$, the test sequence (ϕ_n) is said to be uniformly asymptotically more powerful than (ψ_n). If one can show that this holds for all test sequences (ψ_n), then (ϕ_n) is said to be uniformly asymptotically most powerful. Note, however, that the above ordering of test sequences is not complete: it is quite possible that (ϕ_n) is asymptotically more powerful than (ψ_n) on a subset of $\mathscr{P}_1$, whereas on its complement in $\mathscr{P}_1$, (ψ_n) is asymptotically more powerful. As a result, uniformly most powerful tests do not exist in many problems.

Besides providing a criterion for rejection of a null hypothesis, test sequences may be used to indicate whether the true distribution of the data resides in $\mathscr{P}_0$ or $\mathscr{P}_1$ (where now, $\mathscr{P}_0$ and $\mathscr{P}_1$ are disjoint but may not cover all of $\mathscr{P}$). This requires that we treat H_0 and H_1 on a symmetrical footing, much like in a classification problem. For that purpose, one would like to consider test sequences (ϕ_n) such that the quantity

$$\sup_{P\in\mathscr{P}_0} P\phi_n + \sup_{P\in\mathscr{P}_1} P(1-\phi_n), \tag{2.19}$$

(which is sometimes also referred to as "the power function") is "small" in the limit $n\to\infty$, possibly quantified by introduction of a significance level pertaining to both type-I and type-II errors simultaneously. In many proofs of Bayesian limit theorems (see chapter 4), a test sequence (ϕ_n) is needed such that (2.19) goes to zero, or is bounded by a sequence (a_n) decreasing to zero (typically $a_n = e^{-nD}$ for some $D > 0$). The existence of such test sequences forms the subject of section 4.5.

Closely related to hypothesis tests are confidence intervals. Suppose that pose our inferential problem differently: our interest now lies in using the data $Y \sim P_0$ to find a data-dependent subset $C(Y)$ of the model that contains P_0 with "high" probability. Again, "high" probability requires quantification in terms of a level α, called the confidence level.

Definition 2.3.4. *Let $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ be a parameterized model; let $Y \sim P_{\theta_0}$ for some $\theta_0 \in \Theta$. Choose a confidence level $\alpha \in (0,1)$. Let $C(Y)$ be subset of Θ dependent only on the data Y. Then $C(Y)$ is a confidence region for θ of confidence level α, if*

$$P_\theta\bigl(\theta \in C(Y)\bigr) \geq 1 - \alpha, \tag{2.20}$$

for all $\theta \in \Theta$.

The dependence of C on the data Y is meant to express that $C(Y)$ can be calculated once the data has been observed. The confidence region may also depend on other quantities that are known to the statistician, so $C(Y)$ is a *statistic* (see definition 1.1.9). Note also that the dependence of $C(Y)$ on the data Y makes $C(Y)$ a *random* subset of the model. Compare this to point estimation, in which the data-dependent estimator is a *random* point in the model.

Like hypothesis testing, confidence regions can be considered from an asymptotic point of view, as demonstrated in the following example.

Example 2.3.2. *We consider the experiment of example 2.3.1, i.e. we suppose that the data Y forms an i.i.d. sample from a distribution $P_0 = P_{\theta_0}$ or $\mathbb{R}$ and that $P_\theta X = \theta$ for all $\theta \in \Theta$. Moreover, we assume that for some known constant $S > 0$, $\sigma^2(\theta) = \mathrm{Var}_\theta X \leq S^2$, for all $\theta \in \Theta$. Consider the sample-average $T_n(X_1, \ldots, X_n) = \mathbb{P}_n X$. Choose a confidence level $\alpha \in (0,1)$. The limit (2.17) can be rewritten in the following form:*

$$P_\theta^n\bigl(|T(X_1, \ldots, X_n) - \theta| \leq n^{-1/2}\sigma(\theta) s_{\alpha/2}\bigr) \to 1 - \alpha, \qquad (n \to \infty). \tag{2.21}$$

Define C_n by

$$C_n(X_1, \ldots, X_n) = \bigl[T(X_1, \ldots, X_n) - n^{-1/2} S s_{\alpha/2}, T(X_1, \ldots, X_n) + n^{-1/2} S s_{\alpha/2}\bigr].$$

Then, for all $\theta \in \Theta$,

$$\lim_{n \to \infty} P_\theta^n\bigl(\theta \in C_n(X_1, \ldots, X_n)\bigr) \geq 1 - \alpha.$$

Note that the θ-dependence of $\sigma(\theta)$ would violate the requirement that C_n be a statistic: since the true value θ_0 of θ is unknown, so is $\sigma(\theta)$. Substituting the (known) upper-bound S for $\sigma(\theta)$ enlarges the $\sigma(\theta)$-interval that follows from (2.21) to its maximal extent, eliminating the θ-dependence. In a realistic situation, one would not use S but substitute $\sigma(\theta)$ by an estimator $\hat{\sigma}(Y)$, which amounts to the use of a plug-in version of (2.21). As a result, we would also have to replace the standard-normal quantiles s_α by the quantiles of the Student t-distribution.

Clearly, confidence regions are not unique, but of course small confidence regions are more informative than large ones: if, for some confidence level α, two confidence regions $C(Y)$ and

$D(Y)$ are given, both satisfying (2.20) for all $\theta \in \Theta$, and $C(Y) \subset D(Y)$, P_θ-almost-surely for all θ, then $C(Y)$ is preferred over $D(Y)$.

The Bayesian analogs of tests and confidence regions are called Bayes factors and credible regions, both of which are derived from the posterior distribution. We start by considering credible sets. The rationale behind their definition is exactly the same one that motivated confidence regions: we look for a subset D of the model that is as small as possible, while receiving a certain minimal probability. Presently, however, the word "probability" is in line with the Bayesian notion, *i.e.* probability according to the posterior distribution.

Definition 2.3.5. *Let $(\Theta, \mathscr{G})$ be a measurable space parameterizing a model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ for data Y, with prior $\Pi : \mathscr{G} \to [0,1]$. Choose a level $\alpha \in (0,1)$. Let $D \in \mathscr{G}$ be a subset of Θ. Then D is a level-α credible set for ϑ, if*

$$\Pi\big(\vartheta \in D \bigm| Y\big) \geq 1 - \alpha. \tag{2.22}$$

In a Bayesian setting, one interprets $\Pi(\vartheta \in D|Y)$ as the probability of finding ϑ in D, given the data. Note that credible sets are *random* sets, since they are defined based on the posterior which, in turn, depends on the sample: this data-dependence can be made explicit by writing credible sets as $D(Y)$ instead of D. In practice, one calculates the posterior distribution from the prior and the data and, based on that, proceeds to derive a subset $D(Y)$ such that (2.22) is satisfied. A credible set is sometimes referred to as a credible region, or, if D is an interval in a one-dimensional parametric model, a credible interval.

Remark 2.3.1. *In smooth, parametric models for i.i.d. data there is an close, asymptotic relation between Bayesian credible sets and frequentist confidence intervals centred on the maximum-likelihood estimator: the Bernstein-von Mises theorem (see section 4.4) implies that level-α credible regions coincide with abovementioned level-α confidence intervals asymptotically! In situations where it is hard to calculate the ML estimator or to construct the corresponding confidence interval explicitly, it is sometimes relatively easy to obtain credible regions (based on a simulated sample from the posterior, as obtained from the MCMC procedure (see section 6.1)). In such cases, one can calculate credible regions and conveniently interpret them as confidence intervals centred on the MLE, due to theorem 4.4.1.*

Definition 2.3.5 suffices to capture the concept of a credible set, but offers too much freedom in the choice of D: given a level $\alpha > 0$, many sets will satisfy (2.22), just like confidence regions can be chosen in many different ways. Note that, also here, we prefer smaller sets over large ones: if, for some level α, two different level-α credible sets F and G are given, both satisfying (2.22) and $F \subset G$, then F is preferred over G. If the posterior is dominated with density $\theta \mapsto \pi(\theta|Y)$, we can be more specific. We define, for every $k \geq 0$, the level-set

$$D(k) = \big\{\theta \in \Theta \,:\, \pi(\theta|Y) \geq k\big\}, \tag{2.23}$$

and consider the following.

Definition 2.3.6. *Let $(\Theta,\mathscr{G})$ a measurable space parameterizing a model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ for data $Y \in \mathscr{Y}$, with prior $\Pi : \mathscr{G} \to [0,1]$. Assume that the posterior is dominated by a σ-finite measure μ on $(\Theta,\mathscr{G})$, with density $\pi(\,\cdot\,|Y) : \Theta \to \mathbb{R}$. Choose $\alpha \in (0,1)$. A level-α HPD-credible set (from* highest posterior density*) for ϑ is the subset $D_\alpha = D(k_\alpha)$, where k_α equals:*

$$k_\alpha = \sup\bigl\{k \geq 0 \,:\, \Pi(\vartheta \in D(k)|Y) \geq 1-\alpha\bigr\}.$$

In other words, D_α is the smallest level-set of the posterior density that receives posterior mass greater than or equal to $1-\alpha$. Note that HPD-credible sets depend on the choice of dominating measure: if we had chosen to use a different measure μ, HPD-credible sets would have changed as well! One may wonder what happens if the posterior is dominated by the prior and we use the density of the posterior with respect to the prior to define HPD-credible regions.

Lemma 2.3.1. *Let $(\Theta,\mathscr{G})$ a measurable space parameterizing a model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ for data Y taking values in a measurable space $(\mathscr{Y},\mathscr{B})$. Assume that the model is dominated by some σ-finite measure $\nu : \mathscr{B} \to \mathbb{R}$, with $p_\theta : \mathscr{Y} \to \mathbb{R}$ is the ν-density of P_θ for every $\theta \in \Theta$. Let $\Pi_1, \Pi_2 : \mathscr{G} \to [0,1]$ be two priors, such that $\Pi_1 \ll \Pi_2$ and $\Pi_2 \ll \Pi_1$. Denote the posterior densities with respect to Π_1, Π_2 as $\pi_1(\,\cdot\,|Y), \pi_2(\,\cdot\,|Y) : \mathscr{Y} \to \mathbb{R}$ and corresponding HPD-credible sets as $D_{1,\alpha}$, $D_{2,\alpha}$. Then*

$$D_{1,\alpha} = D_{2,\alpha},$$

for all $\alpha \in (0,1)$.

Proof Under the conditions stated, the densities $\theta \mapsto \pi_1(\theta|Y)$ and $\theta \mapsto \pi_2(\theta|Y)$ are both of the form (2.8). Note that both $\pi_1(\theta|Y)$ and $\pi_1(\theta|Y)$ are almost-sure expressions with respect to their respective priors, but since $\Pi_1 \ll \Pi_2$ and $\Pi_2 \ll \Pi_1$ by assumption, Π_1-almost-sureness and Π_2-almost-sureness are equivalent. From (2.8), we see that

$$\frac{\pi_1(\theta|Y)}{\pi_2(\theta|Y)} = \frac{\displaystyle\int_\Theta p_\theta(Y)\,d\Pi_2(\theta)}{\displaystyle\int_\Theta p_\theta(Y)\,d\Pi_1(\theta)} = K(Y) > 0,$$

almost-surely with respect to both priors (and P_0). So the fraction of posterior densities is a positive constant as a function of θ. Therefore, for all $k \geq 0$,

$$D_1(k) = \bigl\{\theta \in \Theta \,:\, \pi_1(\theta|Y) \geq k\bigr\} = \bigl\{\theta \in \Theta \,:\, \pi_2(\theta|Y)\,K(Y) \geq k\bigr\} = D_2\bigl(K(Y)^{-1}k\bigr).$$

and, hence, for all $\alpha \in (0,1)$,

$$k_{1,\alpha} = \sup\bigl\{k \geq 0 \,:\, \Pi(\vartheta \in D_2(K(Y)^{-1}k)|Y) \geq 1-\alpha\bigr\} = K(Y)\,k_{2,\alpha}.$$

To conclude,

$$D_{1,\alpha} = D_1(k_{1,\alpha}) = D_2\bigl(K(Y)^{-1}k_{1,\alpha}\bigr) = D_2(k_{2,\alpha}) = D_{2,\alpha}.$$

□

The above lemma proves that using the posterior density with respect to the prior leads to HPD-credible sets that are independent of the choice of prior. This may be interpreted further, by saying that *only the data* is of influence on HPD-credible sets based on the posterior density with respect to the prior. Such a perspective is attractive to the objectivist, but rather counterintuitive from a subjectivist point of view: a prior chosen according to subjectivist criteria places high mass in subsets of the model that the statistician attaches "high belief" to. Therefore, the density of the posterior with respect to the prior can be expected to be relatively *small* in those subsets! As a result, those regions may end up in D_α only for relatively high values of α. However, intuition is to be amended by mathematics in this case: when we say above that only the data is of influence, this is due entirely to the likelihood factor in (2.8). Rather than incorporating both prior knowledge and data in HPD credible sets, the above construction emphasizes the *differences* between prior and posterior beliefs, which lie entirely in the data and are represented in the formalism by the likelihood. (We shall reach a similar conclusion when considering the difference between posterior odds and Bayes factors later in this section). To present the same point from a different perspective, HPD credible regions based on the posterior density with respect to the prior coincide with levelsets of the likelihood and centre on the ML estimate if the likelihood is smooth enough and has a well-separated maximum (as a function on the model). We shall see that the coincidence between confidence regions and credible sets becomes more pronounced in the large-sample limit when we study the Bernstein-Von Mises theorem (see chapter 4 for more on large-sample limiting behaviour of the posterior).

Bayesian hypothesis testing is formulated in a far more straightforward fashion than frequentist methods based on the Neyman-Pearson approach. The two hypotheses H_0 and H_1 correspond to a two-set partition $\{\Theta_0, \Theta_1\}$ of the model Θ and for each of the parts, we have both posterior and prior probabilities. Based on the proportions between those, we shall decide which hypothesis is the more likely one. It can therefore be remarked immediately that in the Bayesian approach, the hypotheses are treated on *equal* footing, a situation that is more akin to classification than to Neyman-Pearson hypothesis testing. To introduce Bayesian hypothesis testing, we make the following definitions.

Definition 2.3.7. *Let* $(\Theta, \mathscr{G})$ *a measurable space parameterizing a model* $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ *for data* $Y \in \mathscr{Y}$*, with prior* $\Pi : \mathscr{G} \to [0,1]$*. Let* $\{\Theta_0, \Theta_1\}$ *be a partition of* Θ *such that* $\Pi(\Theta_0) > 0$ *and* $\Pi(\Theta_1) > 0$*. The prior and posterior odds ratios are defined by* $\Pi(\Theta_0)/\Pi(\Theta_1)$ *and* $\Pi(\Theta_0|Y)/\Pi(\Theta_1|Y)$ *respectively. The Bayes factor in favour of* Θ_0 *is defined to be*

$$B = \frac{\Pi(\Theta_0|Y)}{\Pi(\Theta_1|Y)} \frac{\Pi(\Theta_1)}{\Pi(\Theta_0)}.$$

When doing Bayesian hypothesis testing, we have a choice of which ratio to use and that choice will correspond directly with a choice for subjectivist or objectivist philosophies. In the subjectivist's view, the posterior odds ratio has a clear interpretation: if

$$\frac{\Pi(\Theta_0|Y)}{\Pi(\Theta_1|Y)} > 1,$$

then the probability of $\vartheta \in \Theta_0$ is greater than the probability of $\vartheta \in \Theta_0$ and hence, the subjectivist decides to adopt H_0 rather than H_1. If, on the other hand, the above display is smaller than 1, the subjectivist decides to adopt H_1 rather than H_0. The objectivist would object to this, saying that the relative prior weights of Θ_0 and Θ_1 can introduce a heavy bias in favour of one or the other in this approach (upon which the subjectivist would answer that that is exactly what he had in mind). Therefore, the objectivist would prefer to use a criterion that is less dependent on the prior weights of Θ_0 and Θ_1. We look at a very simple example to illustrate the point.

Example 2.3.3. *Let Θ be a dominated model that consists of only two points, θ_0 and θ_1 and let $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \{\theta_1\}$, corresponding to simple null and alternative hypotheses H_0, H_1. Denote the prior by Π and assume that both $\Pi(\{\theta_0\}) > 0$ and $\Pi(\{\theta_1\}) > 0$. By Bayes rule, the posterior weights of Θ_0 and Θ_1 are*

$$\Pi(\vartheta \in \Theta_i|Y) = \frac{p_{\theta_i}(Y)\Pi(\Theta_i)}{p_{\theta_0}(Y)\Pi(\Theta_0) + p_{\theta_1}(Y)\Pi(\Theta_1)},$$

for $i = 0, 1$. Therefore, the posterior odds ratio takes the form:

$$\frac{\Pi(\vartheta \in \Theta_0|Y)}{\Pi(\vartheta \in \Theta_1|Y)} = \frac{p_{\theta_0}(Y)\Pi(\Theta_0)}{p_{\theta_1}(Y)\Pi(\Theta_1)},$$

and the Bayes factor equals the likelihood ratio:

$$B = \frac{p_{\theta_0}(Y)}{p_{\theta_1}(Y)}.$$

We see that the Bayes factor does not depend on the prior weights assigned to Θ_0 and Θ_1 (in this simple example), but the posterior odds ratio does. Indeed, suppose we stack the prior odds heavily in favour of Θ_0, by choosing $\Pi(\Theta_0) = 1-\epsilon$ and $\Pi(\Theta_1) = \epsilon$ (for some small $\epsilon > 0$). Even if the likelihood ratio $p_{\theta_0}(Y)/p_{\theta_1}(Y)$ is much smaller than one (but greater than $\epsilon/1-\epsilon$), the subjectivist's criterion favours H_0. In that case, the data clearly advocates hypothesis H_1 but the prior odds force adoption of H_0. The Bayes factor B equals the likelihood ratio (in this example), so it does not suffer from the bias imposed on the posterior odds.

The objectivist prefers the Bayes factor to make a choice between two hypotheses: if $B > 1$ the objectivist adopts H_0 rather than H_1; if, on the other hand, $B < 1$, then the objectivist adopts H_1 rather than H_0. In example 2.3.3 the Bayes factor is independent of the choice of the prior. In general, the Bayes factor is not completely independent of the prior, but it does not depend on the relative prior weights of Θ_0 and Θ_1. We prove this using the following decomposition of the prior:

$$\Pi(A) = \Pi(A|\Theta_0)\,\Pi(\Theta_0) + \Pi(A|\Theta_1)\,\Pi(\Theta_1), \tag{2.24}$$

for all $A \in \mathscr{G}$ (where it is assumed that $\Pi(\Theta_0) > 0$ and $\Pi(\Theta_1) > 0$). In the above display, $\Pi(\,\cdot\,|\Theta_i)$ can be any probability measure on Θ_i ($i = 0, 1$), and since $\Pi(\Theta_0) + \Pi(\Theta_1) = 1$, Π is decomposed as a convex combination of two probability measures on Θ_0 and Θ_1 respectively. The Bayes factor is then rewritten using Bayes' rule (see lemma A.6.1):

$$B = \frac{\Pi(\Theta_0|Y)}{\Pi(\Theta_1|Y)} \frac{\Pi(\Theta_1)}{\Pi(\Theta_0)} = \frac{\Pi(Y|\Theta_0)}{\Pi(Y|\Theta_1)},$$

where, in a dominated model,

$$\Pi(Y|\Theta_i) = \int_{\Theta_i} p_\theta(Y)\, d\Pi(\theta|\Theta_i),$$

for $i = 0, 1$. In terms of the decomposition (2.24), B depends on $\Pi(\,\cdot\,|\Theta_0)$ and $\Pi(\,\cdot\,|\Theta_1)$, but not on $\Pi(\Theta_0)$ and $\Pi(\Theta_1)$. So using Bayes factors instead of posterior odds exactly eliminates the bias introduced by non-zero prior odds.

Remark 2.3.2. *The condition that both Θ_0 and Θ_1 receive prior mass strictly above zero is important since Bayes factors and odds ratios are based on conditioning of ϑ. Bayesian hypothesis testing is sensible* only *if both Θ_0 and Θ_1 receive non-zero prior mass. This remark plays a role particularly when comparing a* simple *null hypothesis to an alternative, as illustrated in exercise 2.10.*

2.4 Decision theory and classification

Many practical problems require that we make an observation and based on the outcome, make a decision of some kind. For instance when looking for the diagnosis for a patient, a doctor will observe variables like the patients temperature, blood-pressure and appearance, in addition to the results of chemical and physical scans to come to a decision regarding the affliction the patient is probably suffering from. Another example concerns the financial markets, in which past stock- and option-prices are considered by analysts to decide whether to buy or sell stocks and derivatives. In a chemical plant, regulation of a chemical process amounts to a succession of decisions to control and optimize conditions, based on the measurement of thermo-dynamical quantities and concentrations of chemicals involved in the reaction. In this section, we look at problems of this nature, first from a frequentist perspective and then with the Bayesian approach.

Practical problems like those described above usually involve optimality criteria that are prescribed by the context of the problem itself: for example, when a doctor makes the wrong diagnosis for a patient suffering from cancer the consequences can be most serious, whereas the misdiagnosis of a case of influenza is usually no more than unfortunate. In any useful statistical procedure meant to assist in medical diagnosis, such differences should be reflected in the decision-making procedure. That is certainly not the case for the methods that we have discussed thus far. Up to this point, we have used optimality criteria of a more general nature, like the accuracy of an estimation procedure, coverage probabilities for confidence intervals or the probability of Type-I and type-II errors in a testing procedure.

The distinction lies in the nature of the optimality criteria: so far we have practiced what is called statistical inference, in which optimality is formulated entirely in terms of the stochastic description of the data. For that reason, it is sometimes said that statistical inference limits itself to those questions that "summarize the data". By contrast, *statistical decision theory* formalizes the criteria for optimality by adopting the use of a so-called loss-function to quantify the consequences of wrong decisions in a way prescribed by the context of the statistical problem.

In statistical decision theory the nomenclature is slightly different from that introduced earlier. We consider a system that is in an unknown *state* $\theta \in \Theta$, where Θ is called the *state-space*. The observation Y takes its values in the *samplespace* $\mathscr{Y}$, a measurable space with σ-algebra $\mathscr{B}$. The observation is stochastic, its distribution $P_\theta : \mathscr{B} \to [0,1]$ being dependent on the state θ of the system. The observation does not reveal the state of the system completely or with certainty. Based on the outcome $Y = y$ of the observation, we take a *decision* $a \in \mathscr{A}$ (or perform an *action* a, as some prefer to say), where $\mathscr{A}$ is the called the *decision-space*. For each state θ of the system there may be an optimal or prescribed decision, but since observation of Y does not give us the state θ of the system with certainty, the decision is stochastic and may be wrong. The goal of statistical decision theory is to arrive at a rule that decides in the best possible way given only the data Y.

The above does not add anything new to the approach we were already following: aside from the names, the concepts introduced here are those used in the usual problem of statistically estimating $a \in \mathscr{A}$. Decision theory distinguishes itself through its definition of optimality in terms of a so-called loss-function.

Definition 2.4.1. *Any lower-bounded function $L : \Theta \times \mathscr{A} \to \mathbb{R}$ may serve as a loss-function. The utility-function is $-L : \Theta \times \mathscr{A} \to \mathbb{R}$.*

(Although statisticians talk about loss-functions, people in applied fields often prefer to talk of utility-functions, which is why the above definition is given both in a positive and a negative version.) The interpretation of the loss-function is the following: if a particular decision a is taken while the state of the system is θ, then a loss $L(\theta, a)$ is incurred which can be either positive (loss) or negative (profit). To illustrate, in systems where observation of the state is direct (*i.e.* $Y = \theta$) and non-stochastic, the optimal decision $a(\theta)$ given the state θ is the value of a that minimizes the loss $L(\theta, a)$. However, the problem we have set is more complicated because the state θ is unknown and can not be measured directly. All we have is the observation Y.

Definition 2.4.2. *Let $\mathscr{A}$ be a measurable space with σ-algebra $\mathscr{H}$. A measurable $\delta : \mathscr{Y} \to \mathscr{A}$ is called a decision rule.*

A decision-rule is an automated procedure to arrive at a decision $\delta(y)$, given that the observation is $Y = y$. We denote the collection of all decision rules under consideration by Δ. Clearly our goal will be to find decision rules in Δ that "minimize the loss" in an appropriate sense. The above basic ingredients of decision-theoretic problems play a role in both the frequentist and Bayesian analysis. We consider the frequentist approach first and then look at decision theory from a Bayesian perspective.

In frequentist decision theory we assume that $Y \sim P_{\theta_0}$ for some state $\theta_0 \in \Theta$ and we analyze the expectation of the loss.

Definition 2.4.3. *The risk-function $R : \Theta \times \Delta \to \mathbb{R}$ is defined as the expected loss under $Y \sim P_\theta$ when using δ,*

$$R(\theta, \delta) = \int L(\theta, \delta(Y))\, dP_\theta. \tag{2.25}$$

Of interest to the frequentist is only the expected loss under the true distribution $Y \sim P_{\theta_0}$. But since θ_0 is unknown, we are forced to consider *all* values of θ, *i.e.* look at the risk-*function* $\theta \mapsto R(\theta, \delta)$ for each decision rule δ.

Definition 2.4.4. *Let the state-space Θ, states P_θ, ($\theta \in \Theta$), decision space $\mathscr{A}$ and loss L be given. Choose $\delta_1, \delta_2 \in \Delta$. The decision rule δ_1 is R-better than δ_2, if*

$$\forall_{\theta \in \Theta} : \quad R(\theta, \delta_1) < R(\theta, \delta_2). \tag{2.26}$$

A decision rule δ is admissible if there exists no $\delta' \in \Delta$ that is R-better than δ (and inadmissible if such a δ' does exist).

It is clear that the definition of R-better decision-rules is intended to order decision rules: if the risk-function associated with a decision-rule is relatively small, then that decision rule is preferable. Note, however, that the ordering we impose by definition 2.4.4 may be partial rather than complete: pairs δ_1, δ_2 of decision rules may exist such that neither δ_1 nor δ_2 is R-better than the other. This is due to the fact that δ_1 may perform better (in the sense that $R(\theta, \delta_1) \leq R(\theta, \delta_2)$) for values of θ in some $\Theta_1 \subset \Theta$, while δ_2 performs better in $\Theta_2 = \Theta \setminus \Theta_1$, resulting in a situation where (2.26) is true for neither. For that reason, it is important to find a way to compare risks (and thereby decision rules) in a θ-independent way and thus arrive at a complete ordering among decision rules. This motivates the following definition.

Definition 2.4.5. (Minimax decision principle) *Let the state-space Θ, states P_θ, ($\theta \in \Theta$), decision space $\mathscr{A}$ and loss L be given. The function*

$$\Delta \to \mathbb{R} : \delta \mapsto \sup_{\theta \in \Theta} R(\theta, \delta)$$

is called the minimax risk. Let $\delta_1, \delta_2 \in \Delta$ be given. The decision rule δ_1 is minimax-preferred to δ_2, if

$$\sup_{\theta \in \Theta} R(\theta, \delta_1) < \sup_{\theta \in \Theta} R(\theta, \delta_2).$$

If $\delta^M \in \Delta$ *minimizes* $\delta \mapsto \sup_\theta R(\theta, \delta)$*, i.e.*

$$\sup_{\theta\in\Theta} R(\theta, \delta^M) = \inf_{\delta\in\Delta} \sup_{\theta\in\Theta} R(\theta, \delta). \tag{2.27}$$

then δ^M *is called a minimax decision-rule.*

Regarding the existence of minimax decision rules, it is noted that the Minimax theorem (see Strasser (1985) [81]) asserts existence of δ^M and moreover, that

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

under the conditions that R is convex on Δ, concave on Θ and that the topology on Δ is such that Δ is compact, $\delta \mapsto R(\theta, \delta)$ is continuous for all θ. Since many loss-functions used in practice satisfy the convexity requirements, the Minimax theorem has broad applicability in statistical decision theory and many other fields. In some cases, use of the minimax theorem requires that we extend the class Δ to contain more general decision rules. Particularly, it is often necessary to consider the class of all so-called *randomized* decision rules. Randomized decision rules are not only stochastic in the sense that they depend on the data, but also through a further stochastic influence: concretely, this means that after realisation $Y = y$ of the data, uncertainty in the decision remains. To give a formal definition, consider a measurable space $(\Omega, \mathscr{F})$ with data $Y : \Omega \to \mathscr{Y}$ and a decision rule $\delta : \Omega \to \mathscr{A}$. The decision rule δ is a randomized decision rule whenever $\sigma(\delta)$ is not a subset of $\sigma(Y)$, *i.e.* δ is not a function of Y. An example of such a situation is that in which we entertain the possibility of using one of two different non-randomized decision rules $\delta_1, \delta_2 : \mathscr{Y} \to \mathscr{A}$. After the data is realised as $Y = y$, δ_1 and δ_2 give rise to two decisions $\delta_1(y)$, $\delta_2(y)$, which may differ. In that case, we flip a coin with outcome $C \in \{0, 1\}$ to decide which decision to use. The extra stochastic element introduced by the coin-flip has then "randomized" our decision rule. The product space $\mathscr{Y} \times \{0, 1\}$ endowed with the product σ-algebra may serve as the measurable space $(\Omega, \mathscr{F})$ with $\delta : \Omega \to \mathscr{Y}$ defined by,

$$(y, c) \mapsto \delta(y, c) = c\, \delta_1(Y) + (1 - c)\, \delta_2(y),$$

for all $y \in \mathscr{Y}$ and $c \in \{0, 1\}$. Perhaps a bit counterintuitively (but certainly in accordance with the fact that minimization over a larger set produces a lower infimum), in some decision problems the minimax risk associated with such randomized decision rules lies strictly below the minimax risks of both non-randomized decision rules. We return to the Minimax theorem in section 4.3.

Example 2.4.1. (Decision theoretic L_2-estimation) *The decision-theoretic approach can also be used to formulate estimation problems in a generalized way, if we choose the decision space* $\mathscr{A}$ *equal to the state-space* $\Theta = \mathbb{R}$*. Let* $Y \sim N(\theta_0, 1)$ *for some unknown* $\theta_0 \in \Theta$*. Choose* $L : \Theta \times \mathscr{A} \to \mathbb{R}$ *equal to the quadratic difference*

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

a choice referred to as an L_2-loss (or squared-error loss*). Consider the decision-space*

$$\Delta = \{\delta_c : \mathscr{Y} \to \mathscr{A} \,:\, \delta_c(y) = c\,y,\ c \geq 0\}.$$

Note that Δ plays the role of a family of estimators for θ_0 here. The risk-function takes the form:

$$\begin{aligned}
R(\theta,\delta_c) &= \int L(\theta,\delta_c(Y))\,dP_\theta = \int_{\mathbb{R}} (\theta - cy)^2 dN(\theta,1)(y)\\
&= \int_{\mathbb{R}} \bigl(c(\theta-y) + (1-c)\theta\bigr)^2 dN(\theta,1)(y)\\
&= \int_{\mathbb{R}} \Bigl(c^2(y-\theta)^2 + 2c(1-c)\theta(\theta-y) + (1-c)^2\theta^2\Bigr) dN(\theta,1)(y)\\
&= c^2 + (1-c)^2\theta^2.
\end{aligned}$$

It follows that δ_1 is R-better than all δ_c for $c > 1$, so that for all $c > 1$, δ_c is inadmissible. If we had restricted c to be greater than or equal to 1, δ_1 would have been admissible. However, since c may lie in $[0,1)$ as well, admissibility in the uniform sense of (2.26) does not apply to any δ_c. To see this, note that $R(\theta,\delta_1) = 1$ for all θ, whereas for $c < 1$ and some $\theta > c/(1-c)$, $R(0,\delta_c) < 1 < R(\theta,\delta_c)$. Therefore, there is no admissible decision rule in Δ.

The minimax criterion does give rise to a preference. However, in order to guarantee its existence, we need to bound (or rather, compactify) the parameter space: let $M > 0$ be given and assume that $\Theta = [-M,M]$. The minimax risk for δ_c is given by

$$\sup_{\theta\in\Theta} R(\theta,\delta_c) = c^2 + (1-c)^2 M^2,$$

which is minimal iff $c = M^2/(1+M^2)$, i.e. the (unique) minimax decision rule for this problem (or, since we are using decision theory to estimate a parameter in this case, the minimax estimator with respect to L_2-loss) is therefore,

$$\delta^M(Y) = \frac{M^2}{1+M^2}Y.$$

Note that if we let $M \to \infty$, this estimator for θ converges to the MLE for said problem.

As demonstrated in the above example, uniform admissibility of a decision rule (*c.f.* (2.26)) is hard to achieve, but in many such cases a minimax decision rule does exist. One important remark concerning the use the minimax decision principle remains: considering (2.27), we see that the minimax principle chooses the decision rule that minimizes the *maximum* of the risk $R(\,\cdot\,,\delta)$ over Θ. As such, the minimax criterion takes into account *only* the worst-case scenario and prefers decision rules that perform well under those conditions. In practical problems, that means that the minimax principle tends to take a rather pessimistic perspective on decision problems.

Bayesian decision theory presents a more balanced perspective because instead of maximizing the risk function over Θ, the Bayesian has the prior to integrate over Θ. Optimization of the resulting integral takes into account more than just the worst case, so that the resulting decision rule is based on a less pessimistic perspective than the minimax decision rule.

Definition 2.4.6. *Let the state-space Θ, states P_θ, ($\theta \in \Theta$), decision space $\mathscr{A}$ and loss L be given. In addition, assume that Θ is a measurable space with σ-algebra $\mathscr{G}$ and prior $\Pi : \mathscr{G} \to \mathbb{R}$. The function*

$$r(\Pi, \delta) = \int_\Theta R(\theta, \delta)\, d\Pi(\theta), \tag{2.28}$$

is called the Bayesian risk function. Let $\delta_1, \delta_2 \in \Delta$ be given. The decision rule δ_1 is Bayes-preferred to δ_2, if

$$r(\Pi, \delta_1) < r(\Pi, \delta_2).$$

If $\delta^\Pi \in \Delta$ minimizes $\delta \mapsto r(\Pi, \delta)$, i.e.

$$r(\Pi, \delta^\Pi) = \inf_{\delta \in \Delta} r(\Pi, \delta). \tag{2.29}$$

then δ^Π is called a Bayes rule. The quantity $r(\Pi, \delta^\Pi)$ is called the Bayes risk.

Lemma 2.4.1. *Let $Y \in \mathscr{Y}$ denote data in a decision theoretic problem with state space Θ, decision space $\mathscr{A}$ and loss $L : \Theta \times \mathscr{A} \to \mathbb{R}$. For any prior Π and all decision rules $\delta : \mathscr{Y} \to \mathscr{A}$,*

$$r(\Pi, \delta) \leq \sup_{\theta \in \Theta} R(\theta, \delta),$$

i.e. the Bayesian risk is always upper bounded by the minimax risk.

The proof of this lemma follows from the fact that the minimax risk is an upper bound for the integrand in the Bayesian risk function.

Example 2.4.2. (continuation of example 2.4.1) *Let $\Theta = \mathbb{R}$ and $Y \sim N(\theta_0, 1)$ for some unknown $\theta_0 \in \Theta$. Choose the loss-function $L : \Theta \times \mathscr{A} \to \mathbb{R}$ and the decision space Δ as in example 2.4.1. We choose a prior $\Pi = N(0, \tau^2)$ (for some $\tau > 0$) on Θ. Then the Bayesian risk function is give by:*

$$\begin{aligned} r(\Pi, \delta_c) &= \int_\Theta R(\theta, \delta_c)\, d\Pi(\theta) = \int_\mathbb{R} \bigl(c^2 + (1-c)^2\theta^2\bigr) dN(0, \tau^2)(\theta) \\ &= c^2 + (1-c)^2\tau^2, \end{aligned}$$

which is minimal iff $c = \tau^2/(1+\tau^2)$. The (unique) Bayes rule for this problem and corresponding Bayes risk are therefore,

$$\delta^\Pi(Y) = \frac{\tau^2}{1+\tau^2} Y, \qquad r(\Pi, \delta^\Pi) = \frac{\tau^2}{1+\tau^2}.$$

In the Bayesian case, there is no need for a compact parameter space Θ, since we do not maximize but integrate over Θ.

In the above example, we could find the Bayes rule by straightforward optimization of the Bayesian risk function, because the class Δ was rather restricted. If we extend the class Δ to contain *all* non-randomized decision rules, the problem of finding the Bayes rule seems to be far more complicated at first glance. However, as we shall see in theorem 2.4.1, the following definition turns out to be the solution to this question.

Definition 2.4.7. (The conditional Bayes decision principle) *Let the state-space Θ, states P_θ, ($\theta \in \Theta$), decision space $\mathscr{A}$ and loss L be given. In addition, assume that Θ is a measurable space with σ-algebra $\mathscr{G}$ and prior $\Pi : \mathscr{G} \to \mathbb{R}$. We define the decision rule $\delta^* : \mathscr{Y} \to \mathscr{A}$ to be such that for all $y \in \mathscr{Y}$,*

$$\int_\Theta L(\theta, \delta^*(y))\, d\Pi(\theta|Y = y) = \inf_{a \in \mathscr{A}} \int_\Theta L(\theta, a)\, d\Pi(\theta|Y = y), \tag{2.30}$$

i.e. point-wise for every y, the decision rule $\delta^*(y)$ minimizes the posterior *expected loss.*

The above defines the decision rule δ^* implicitly as a point-wise minimizer, which raises the usual questions concerning existence and uniqueness, of which little can be said in any generality. However, if the existence of δ^* is established, it is optimal.

Theorem 2.4.1. *Let the state-space Θ, states P_θ, ($\theta \in \Theta$), decision space $\mathscr{A}$ and loss L be given. In addition, assume that Θ is a measurable space with σ-algebra $\mathscr{G}$ and prior $\Pi : \mathscr{G} \to \mathbb{R}$. Assume that there exists a σ-finite measure $\mu : \mathscr{B} \to \mathbb{R}$ such that $P_\theta \ll \mu$ for all $\theta \in \Theta$. If the decision rule $\delta^* : \mathscr{Y} \to \mathscr{A}$ is well-defined, then δ^* is a Bayes rule.*

Proof Denote the class of all decision rules for this problem by Δ throughout the proof. We start by rewriting the Bayesian risk function for a decision rule $\delta : \mathscr{Y} \to \mathscr{A}$.

$$\begin{aligned} r(\Pi, \delta) &= \int_\Theta R(\theta, \delta)\, d\Pi(\theta) = \int_\Theta \int_{\mathscr{Y}} L(\theta, \delta(y))\, dP_\theta(y)\, d\Pi(\theta) \\ &= \int_{\mathscr{Y}} \int_\Theta L(\theta, \delta(y))\, p_\theta(y)\, d\Pi(\theta)\, d\mu(y) \\ &= \int_{\mathscr{Y}} \Bigl(\int_\Theta p_\theta(y)\, d\Pi(\theta) \Bigr) \int_\Theta L(\theta, \delta(y))\, d\Pi(\theta|Y = y)\, d\mu(y). \end{aligned}$$

where we use definitions (2.28) and (2.25), the Radon-Nikodym theorem (see theorem A.4.2), Fubini's theorem (see theorem A.4.1) and the definition of the posterior, *c.f.* (2.7). Using the prior predictive distribution (2.9), we rewrite the Bayesian risk function further:

$$r(\Pi, \delta) = \int_{\mathscr{Y}} \int_\Theta L(\theta, \delta(y))\, d\Pi(\theta|Y = y)\, dP^\Pi(y). \tag{2.31}$$

By assumption, the conditional Bayes decision rule δ^* exists. Since δ^* satisfies (2.30) point-wise for all $y \in \mathscr{Y}$, we have

$$\int_\Theta L(\theta, \delta^*(y))\, d\Pi(\theta|Y = y) = \inf_{\delta \in \Delta} \int_\Theta L(\theta, \delta(y))\, d\Pi(\theta|Y = y).$$

Substituting this in (2.31), we obtain

$$\begin{aligned}r(\Pi,\delta^*) &= \int_{\mathscr{Y}} \inf_{\delta\in\Delta} \int_\Theta L(\theta,\delta(y))\,d\Pi(\theta|Y=y)\,dP^\Pi(y)\\ &\leq \inf_{\delta\in\Delta} \int_{\mathscr{Y}} \int_\Theta L(\theta,\delta(y))\,d\Pi(\theta|Y=y)\,dP^\Pi(y)\\ &= \inf_{\delta\in\Delta} r(\Pi,\delta).\end{aligned}$$

which proves that δ^* is a Bayes rule. □

To conclude, it is noted that randomization of the decision is not needed when optimizing with respect to the Bayes risk. The conditional Bayes decision rule is non-randomized and optimal.

Example 2.4.3. (Classification and Bayesian classifiers) *Many decision-theoretic questions take the form of a classification problem: under consideration is a population Ω of objects that each belong to one of a finite number of classes $\mathscr{A} = \{1,2,\ldots,L\}$. The class K of the object is the unknown quantity of interest. Observing a vector Y of features of the object, the goal is to* classify *the object, i.e. estimate which class it belongs to. We formalize the problem in decision-theoretic terms: the population is a probability space $(\Omega,\mathscr{F},P)$; both the feature vector and the class of the object are random variables, $Y:\Omega\to\mathscr{Y}$ and $K:\Omega\to\mathscr{A}$ respectively. The state-space in a classification problem equals the decision space $\mathscr{A}$: the class can be viewed as a "state" in the sense that the distribution $P_{Y|K=k}$ of Y* given *the class $K=k$ depends on k. Based on the feature vector Y, we decide to classify in class $\delta(Y)$, i.e. the decision rule (or classifier, as it is usually referred to in the context of classification problems) maps features to classes by means of a map $\delta:\mathscr{Y}\to\mathscr{A}$. A classifier δ can be viewed equivalently as a finite partition of the feature-space $\mathscr{Y}$: for every $k\in\mathscr{A}$, we define*

$$\mathscr{Y}_k = \{y\in\mathscr{Y} : \delta(y)=k\}$$

and note that if $k\neq l$, then $\mathscr{Y}_k\cap\mathscr{Y}_l=\emptyset$ and $\mathscr{Y}_1\cup\mathscr{Y}_2\cup\ldots\cup\mathscr{Y}_L=\mathscr{Y}$. The partition of the feature space is such that if $Y=y\in\mathscr{Y}_k$ for certain $k\in\mathscr{A}$, then we classify the object in class k.

Depending on the context of the classification problem, a loss-function $L:\mathscr{A}\times\mathscr{A}\to\mathbb{R}$ is defined (see the examples in the introduction to this section, e.g. the example on medical diagnosis). Without context, the loss function in a classification problem can be chosen as follows

$$L(k,l) = 1_{\{k\neq l\}}.$$

i.e. we incur a loss equal to one for each misclassification.

Using the minimax decision principle, we look for a classifier $\delta^M:\mathscr{Y}\to\mathscr{A}$ that minimizes:

$$\delta\mapsto \sup_{k\in\mathscr{A}} \int_{\mathscr{Y}} L(k,\delta(y))\,dP(y|K=k) = \sup_{k\in\mathscr{A}} P\big(\delta(Y)\neq k \bigm| K=k\big),$$

i.e. the minimax decision principle prescribes that we minimize the probability of misclassification uniformly over all classes.

In a Bayesian context, we need a prior on the state-space, which equals $\mathscr{A}$ in classification problems. Note that if known (or estimable), the marginal probability distribution for K is to be used as the prior for the state k, in accordance with definition 2.1.1. In practical problems, frequencies of occurrence for the classes $\{1, \ldots, L\}$ in Ω are often available or easily estimable; in the absence of information on the marginal distribution of K equal prior weights can be assigned. Here, we assume that the probabilities $P(K = k)$ are known and use them to define the prior density with respect to the counting measure on the (finite) space $\mathscr{A}$:

$$\pi(k) = P(K = k).$$

The Bayes rule $\delta^* : \mathscr{Y} \to \mathscr{A}$ for this classification problem is defined to as the minimizer of

$$\delta \mapsto \int_{\mathscr{A}} L(k, \delta(y))\, d\Pi(k|Y = y) = \sum_{k=1}^{L} \Pi\big(\delta(y) \neq K \mid Y = y\big)$$

for every $y \in \mathscr{Y}$. According to theorem 2.4.1, the classifier δ^* minimizes the Bayes risk, which in this situation is given by:

$$\begin{aligned} r(\Pi, \delta) &= \int_{\mathscr{A}} R(k, \delta)\, d\Pi(\theta) = \sum_{k \in \mathscr{A}} \int_{\mathscr{Y}} L(k, \delta(y))\, dP(y|K = k)\, \pi(k) \\ &= \sum_{k \in \mathscr{A}} P\big(k \neq \delta(Y) \mid K = k\big)\, P(K = k) = P\big(K \neq \delta(Y)\big). \end{aligned}$$

Summarizing, the Bayes rule δ^* minimizes the overall probability of misclassification, i.e. without referring to the class of the object. (Compare this with the minimax classifier.)

Readers interested in the statistics of classification and its applications are encouraged to read B. Ripley's "Pattern recognition and neural networks" (1996) [73].

To close the chapter, the following remark is in order: when we started our comparison of frequentist and Bayesian methods, we highlighted the conflict in philosophy. However, now that we have seen some of the differences in more detail by considering estimation, testing and decision theory in both schools, we can be far more specific. Statistical problems can be solved in both schools; whether one chooses for a Bayesian or frequentist solution is usually not determined by adamant belief in either philosophy, but by much more practical considerations. Perhaps example 2.4.3 illustrates this point most clearly: if one is concerned about correct classification for objects in the most difficult class, one should opt for the minimax decision rule. If, on the other hand, one wants to minimize the overall misclassification probability (disregarding misclassification per class), one should choose to adopt the conditional Bayes decision rule. In other words, depending on the risk to be minimized (minimax risk and Bayes risk are different!) one arrives at different classifiers. Some formulations are more natural in frequentist context and others belong in the Bayesian realm. Similarly, practicality may form an argument in favour of imposing a (possibly subjective) bias (see example 1.2.1). Bayesian methods are a natural choice in such cases, due to the intrinsic bias priors express. For example, forensic statistics is usually performed using Bayesian methods, in order to leave room for common-sense bias. Another reason to use one or the other may be computational advantages or useful theoretical results that exist for one school but have no analog in the other.

Philosophical preference should not play a role in the choice for a statistical procedure, practicality should (and usually does).

2.5 Exercises

Exercise 2.1. Calibration
A physicist prepares for repreated measurement of a physical quantity Z in his laboratory. To that end, he installs a measurement apparatus that will give him outcomes of the form $Y = Z + e$ where e is a measurement error due to the inaccuracy of the apparatus, assumed to be stochastically independent of Z. Note that if the expectation of e equals zero, long-run sample averages converge to the expectation of Z; if $Pe \neq 0$, on the other hand, averaging does not cancel out the resulting bias.

The manufacturer of the apparatus says that e is normally distributed with known variance $\sigma^2 > 0$. The mean θ of this normal distribution depends on the way the apparatus is installed and thus requires calibration. The following questions pertain to the calibration procedure.

The physicist decides to conduct the following steps to calibrate his measurement: if he makes certain that the apparatus receives no input signal, $Z = 0$. A sample of n independent measurements of Y then amounts to an i.i.d. sample from the distribution of e, which can be used to estimate the unknown mean θ. The physicist expects that Ee lies close to zero.

a. *Explain why, from a subjectivist point of view, the choice $\theta \sim N(0, \tau^2)$ forms a suitable prior in this situation. Explain the role of the parameter $\tau^2 > 0$.*

b. *With the choice of prior as in part* a., *calculate the posterior density for θ.*

c. *Interpret the influence of τ^2 on the posterior, taking into account your answer under part* a. *(Hint: take limits $\tau^2 \downarrow 0$ and $\tau^2 \uparrow \infty$ in the expression you have found under* b.*)*

d. *What is the influence of the samplesize n? Show that the particular choice of the constant τ^2 becomes irrelevant in the large-sample limit $n \to \infty$.*

Exercise 2.2. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(0, \sigma^2)$, with unknown variance $\sigma^2 > 0$. As a prior for σ^2, let $1/\sigma^2 \sim \Gamma(1, 2)$. Calculate the posterior distribution for σ^2 with respect to the Lebesgue measure on $(0, \infty)$.*

Exercise 2.3. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Poisson distribution* $\text{Poisson}(\lambda)$, *with unknown parameter $\lambda > 0$. As a prior for λ, let $\lambda \sim \Gamma(2, 1)$. Calculate the posterior density for λ with respect to the Lebesgue measure on $(0, \infty)$.*

Exercise 2.4. *Let the measurement $Y \sim P_0$ be given. Assume that the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is dominated but possibly misspecified. Let Π denote a prior distribution on Θ. Show that the posterior distribution is P_0-almost-surely equal to the prior distribution iff the likelihood is $\Pi \times P_0$-almost-surely constant (as a function of $(\theta, y) \in \Theta \times \mathscr{Y}$). Explain the result of example 2.1.1 in this context.*

Exercise 2.5. *Consider the following questions in the context of exercise 2.3.*

a. *Calculate the maximum-likelihood estimator and the maximum-a-posteriori estimator for $\lambda \in (0, \infty)$.*

b. *Let $n \to \infty$ both in the MLE and MAP estimator and conclude that the difference vanishes in the limit.*

c. *Following remark 2.2.7, explain the difference between ML and MAP estimators exclusively in terms of the prior.*

d. *Consider and discuss the choice of prior $\lambda \sim \Gamma(2, 1)$ twice, once in a qualitative, subjectivist Bayesian fashion, and once following the frequentist interpretation of the log-prior-density.*

Exercise 2.6. *Let $Y \sim P_0$ denote the data. The following questions pertain to the small-ball estimator defined in remark 2.2.5 for certain, fixed $p \in (1/2, 1)$, which we shall denote by $\hat{P}(Y)$. Assume that the model $\mathscr{P}$ is compact in the topology induced by the metric d.*

a. *Show that for any two measurable model subsets $A, B \subset \mathscr{P}$,*

$$\big|\Pi(\, A \,|\, Y\,) - \Pi(\, B \,|\, Y\,)\big| \leq \Pi(\, A \cup B \,|\, Y\,) - \Pi(\, A \cap B \,|\, Y\,),$$

P_0-almost-surely.

b. *Prove that the map $(\epsilon, P) \mapsto \Pi(\, B_d(P, \epsilon) \,|\, Y\,)$ is continuous, P_0-almost-surely.*

c. *Show that $\hat{P}(Y)$ exists, P_0-almost-surely.*

d. *Suppose that $\epsilon > 0$ denotes the smallest radius for which there exists a ball $B_d(P, \epsilon) \subset \mathscr{P}$ of posterior probability greater than or equal to p. Show that, if both $\hat{P}_1(Y)$ and $\hat{P}_2(Y)$ are centre points of such balls, then $d(\hat{P}_1(Y), \hat{P}_2(Y)) < 2\epsilon$, P_0-almost-surely.*

Exercise 2.7. *Complete the proof of lemma 2.1.2. (Hint: Denote $S = \mathrm{supp}(\Pi)$; assume that $\Pi(S) = \pi < 1$; show that $\Pi(S^c \cap C) = 1 - \pi$ for any closed C such that $\Pi(C) = 1$; then use that intersections of closed sets are closed.*

Exercise 2.8. *Let Y be normally distributed with known variance $\sigma^2 > 0$ and unknown location θ. As a prior for θ, choose $\Pi = N(0, \tau^2)$. Let $\alpha \in (0, 1)$ be given. Using the posterior density with respect to the Lebesgue measure, express the level-α HPD-credible set in terms of Y, σ^2, τ^2 and quantiles of the standard normal distribution. Consider the limit $\tau^2 \to \infty$ and compare with level-α confidence intervals centred on the ML estimate for θ.*

Exercise 2.9. *Let $Y \sim \mathrm{Bin}(n;p)$ for known $n \geq 1$ and unknown $p \in (0,1)$. As a prior for p, choose $\Pi = \mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$. Calculate the posterior distribution for the parameter p. Using the Lebesgue measure on $(0,1)$ to define the posterior density, give the level-α HPD-credible interval for p in terms of Y, n and the quantiles of beta-distributions.*

Exercise 2.10. *Consider a dominated model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ for data Y, where $\Theta \subset \mathbb{R}$ is an interval. For certain $\theta_0 \in \Theta$, consider the simple null-hypothesis and alternative:*

$$H_0 : \quad \theta = \theta_0, \qquad H_1 \quad : \theta \neq \theta_0.$$

Show that if the prior Π is absolutely continuous with respect to the Lebesgue measure on Θ, then the Bayes factor B for the hypotheses H_0 versus H_1 satisfies $B = 0$.
Interpret this fact as follows: calculation of Bayes factors (and posterior/prior odds ratios) makes sense only *if both hypotheses receive non-zero prior mass. Otherwise, the statistical question we ask is rendered invalid* ex ante *by our beliefs concerning θ, as formulated through the choice of the prior.*

Exercise 2.11. Prisoner's dilemma
Two men have been arrested on the suspicion of burglary and are held in separate cells awaiting interrogation. The prisoners have been told that burglary carries a maximum sentence of x years. However, if they confess, their prison terms are reduced to y years (where $0 < y < x$). If one of them confesses and the other does not, the first receives a sentence of y years while the other is sentenced to x years.

Guilty of the crime he is accused of, our prisoner contemplates whether to confess to receive a lower sentence, or to deny involvement in the hope of escaping justice altogether. He cannot confess without implicating the other prisoner. If he keeps his mouth shut and so does his partner in crime, they will both walk away free. If he keeps his mouth shut but his partner talks, he gets the maximum sentence. If he talks, he will always receive a sentence of y years and the other prisoner receives y or x years depending on whether he confessed or not himself. To talk or not to talk, that is the question.

There is no data in this problem, so we set θ equal to 1 or 0, depending on whether the other prisoner talks or not. Our prisoner can decide to talk ($t = 1$) or not ($t = 0$). The loss function $L(\theta, t)$ equals the prison term for our prisoner. In the absence of data, risk and loss are equal.

a. *Calculate the minimax risk for both $t = 0$ and $t = 1$. Argue that the minimax-optimal decision for our prisoner is to confess.*

As argued in section 2.4, the minimax decision can be overly pessimistic. In the above, it assumes that the other prisoner will talk and chooses t accordingly.

The Bayesian perspective balances matters depending on the chance that the other prisoner will confess when interrogated. This chance finds its way into the formalism as a prior for the trustworthiness of the other prisoner. Let $p \in [0,1]$ be the probability that the other prisoner confesses, i.e. $\Pi(\theta = 1) = p$ and $\Pi(\theta = 0) = 1 - p$.

b. Calculate the Bayes risks for $t = 0$ *and* $t = 1$ *in terms of* x*,* y *and* p*. Argue that the Bayes decision rule for our prisoner is as follows: if* $y/x > p$ *then our prisoner does not confess, if* $y/x < p$*, the prisoner confesses. If* $y/x = p$*, the Bayes decision criterion does not have a preference.*

So, depending on the degree to which our prisoner trusts his associate and the ratio of prison terms, the Bayesian draws his conclusion. The latter is certainly more sophisticated and perhaps more realistic, but it requires that our prisoner quantifies his trust in his partner in the form of a prior Bernoulli(p) *distribution.*