2

Normal Models

This was where the work really took place.
—**Ian Rankin**, ***Knots & Crosses***.—

Roadmap

This chapter uses the standard normal $\mathscr{N}(\mu, \sigma^2)$ distribution as an easy entry to generic Bayesian inferential methods. As in every subsequent chapter, we start with a description of the data used as a chapter benchmark for illustrating new methods and for testing assimilation of the techniques. We then propose a corresponding statistical model centered on the normal distribution and consider specific inferential questions to address at this level, namely parameter estimation, model choice, and outlier detection, once set the description of the Bayesian resolution of inferential problems. As befits a first chapter, we also introduce here general computational techniques known as Monte Carlo methods.

J.-M. Marin and C.P. Robert, *Bayesian Essentials with R*, Springer Texts in Statistics, DOI 10.1007/978-1-4614-8687-9_2,
© Springer Science+Business Media New York 2014

2.1 Normal Modeling

The normal (or Gaussian) distribution $\mathscr{N}(\mu, \sigma^2)$, with density on $\mathbb{R}$,

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

is certainly one of the most studied and one of the most used distributions because of its "normality": It appears both as the limit of additive small effects and as a representation of symmetric phenomena without long tails, and it offers many openings in terms of analytical properties and closed-form computations. As such, it is thus the natural opening to a modeling course, even more than discrete and apparently simpler models such as the binomial and Poisson models we will discuss in the following chapters. Note, however, that we do not advocate at this stage the use of the normal distribution as a one-fits-all model: There exist many continuous situations where a normal model is inappropriate for many possible reasons (e.g., skewness, fat tails, dependence, multimodality).

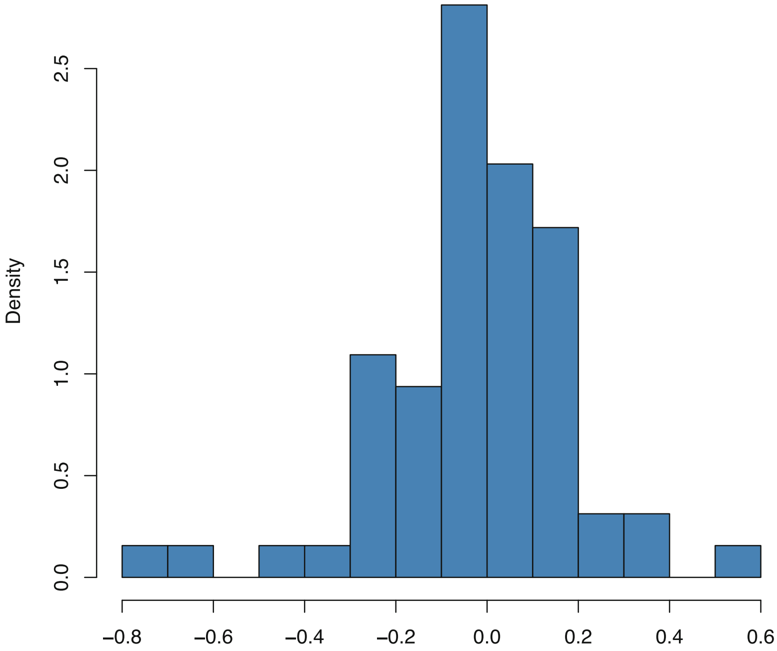

Fig. 2.1. Dataset **normaldata**: Histogram of the observed fringe shifts in Illingworth's 1927 experiment

Our normal dataset, **normaldata**, is linked with the famous Michelson–Morlay experiment that opened the way to Einstein's relativity theory in 1887. The experiment was intended to detect the "æther flow" and hence the existence of æther, this theoretical medium physicists postulated at this epoch was necessary to the transmission of light. Michelson's measuring device consisted in measuring the difference in the speeds of two light beams travelling the same distance in two orthogonal directions. As often in physics, the measurement was done by interferometry and differences in the travelling time inferred from shift in the fringes of the light spectrum. However, the experiment produced very small measurements that were not conclusive for the detection of the æther. Later experiments tried to achieve higher precision, as the one by Illingworth in 1927 used here as **normaldata**, only to obtain smaller and smaller upper bounds on the æther windspeed. While the original dataset is available in R as `morley`, the entries are approximated to the nearest multiple of ten and are therefore difficult to analyze as normal observations.

The 64 data points in **normaldata** are associated with session numbers (first column), corresponding to different times of the day, and the values in the second column represent the averaged fringe displacement due to orientation taken over ten measurements made by Illingworth, who assumed a normal error model. Figure 2.1 produces an histogram of the data by the simple R commands

```
> data(normaldata)
> shift=normaldata[,2]
> hist(shift,nclass=10,col="steelblue",prob=TRUE,main="")
```

This histogram seems compatible with a symmetric unimodal distribution such as the normal distribution. As shown in Fig. 2.2 by a qq-plot obtained by the commands

```
> qqnorm((shift-mean(shift))/sd(shift),pch=19,col="gold2")
> abline(a=0,b=1,lty=2,col="indianred",lwd=2)
```

which compare the empirical cdf with a pluggin normal estimate, The $\mathscr{N}(\mu,\sigma^2)$ fit may not be perfect, though, because of (a) a possible bimodality of the histogram and (b) potential outliers.

As mentioned above, the use of a normal distribution for modeling a given dataset is a convenient device that does not need to correspond to a perfect fit. With some degree of approximation, the normal distribution may agree with the data sufficiently to be used in place of the true distribution (if any). There exist, however, some setups where the normal distribution is thought to be the exact distribution behind the dataset (or where departure from normality has a significance for the theory behind the observations). In Marin and Robert (2007), we introduced a huge dataset related to the astronomical concept of

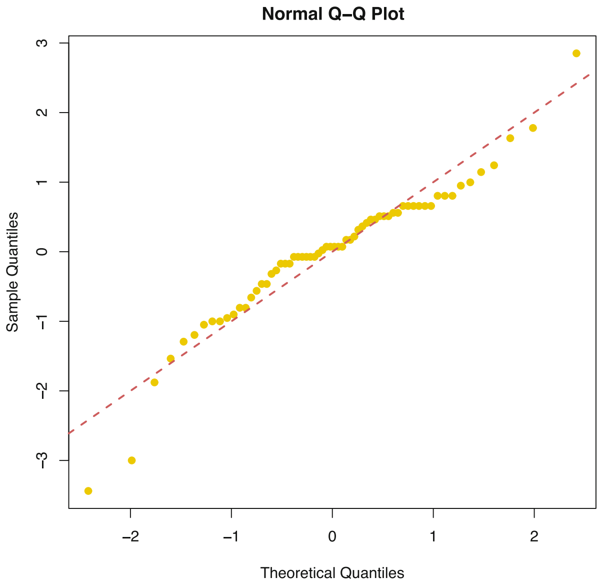

Fig. 2.2. Dataset **normaldata**: qq-plot of the observed fringe shifts against the normal quantiles

the cosmological background noise that illustrated this point, but chose not to reproduce the set in this edition due to the difficulty in handling it.

2.2 The Bayesian Toolkit

2.2.1 Posterior Distribution

Given an independent and identically distributed (later abbreviated as iid) sample $\mathscr{D}_n = (x_1, \ldots, x_n)$ from a density $f(x|\theta)$, depending upon an unknown parameter $\theta \in \Theta$, for instance the mean μ of the benchmark normal distribution, the associated likelihood function is

$$\ell(\theta|\mathscr{D}_n) = \prod_{i=1}^{n} f(x_i|\theta)\,. \tag{2.1}$$

This function of θ is a fundamental entity for the analysis of the information provided about θ by the sample $\mathscr{D}_n$, and Bayesian analysis relies on (2.1) to draw its inference on θ. For instance, when $\mathscr{D}_n$ is a normal $\mathscr{N}(\mu, \sigma^2)$ sample of size n and $\theta = (\mu, \sigma^2)$, we get

$$\ell(\theta|\mathscr{D}_n) = \prod_{i=1}^{n} \exp\{-(x_i - \mu)^2/2\sigma^2\}/\sqrt{2\pi}\sigma$$

$$
\begin{aligned}
&\propto \exp\left\{-\sum_{i=1}(x_i-\mu)^2/2\sigma^2\right\}/\sigma^n \\
&\propto \exp\left\{-\left(n\mu^2-2n\bar{x}\mu+\sum_{i=1} x_i^2\right)/2\sigma^2\right\}/\sigma^n \\
&\propto \exp\left\{-\left[n(\mu-\bar{x})^2+s^2\right]/2\sigma^2\right\}/\sigma^n,
\end{aligned}
$$

where $\bar{x}$ denotes the empirical mean and where s^2 is the sum $\sum_{i=1}^n (x_i-\bar{x})^2$. This shows in particular that $\bar{x}$ and s^2 are sufficient statistics.

↯ In the above display of equations, the sign $\propto$ means *proportional to*. This proportionality is understood for functions of θ, meaning that the discarded constants do not depend on θ but may well depend on the data $\mathscr{D}_n$. This shortcut is both handy in complex Bayesian derivations and fraught with danger when considering several levels of parameters.

The major input of the Bayesian approach, compared with a traditional likelihood approach, is that it modifies the likelihood function into a *posterior* distribution, which is a valid probability distribution on Θ defined by the classical Bayes' formula (or theorem)

$$
\pi(\theta|\mathscr{D}_n)=\frac{\ell(\theta|\mathscr{D}_n)\pi(\theta)}{\int \ell(\theta|\mathscr{D}_n)\pi(\theta)\,\mathrm{d}\theta}\,. \tag{2.2}
$$

The factor $\pi(\theta)$ in (2.2) is called the *prior* and it obviously has to be chosen to start the analysis.

↯ The posterior density is a probability density on the parameter, which does not mean the parameter θ need be a genuine random variable. This density is used as an inferential tool, not as a truthful representation.

A first motivation for this approach is that the prior distribution summarizes the *prior information* on θ; that is, the knowledge that is available on θ *prior* to the observation of the sample $\mathscr{D}_n$. However, the choice of $\pi(\theta)$ is often decided on practical grounds rather than strong subjective beliefs or overwhelming prior information. A second motivation for the Bayesian construct is therefore to provide a fully probabilistic framework for the inferential analysis, with respect to a reference measure $\pi(\theta)$.

As an illustration, consider the simplest case of the normal distribution with known variance, $\mathscr{N}(\mu,\sigma^2)$. If the prior distribution on μ, $\pi(\mu)$, is the normal $\mathscr{N}\left(0,\sigma^2\right)$, the posterior distribution is easily derived via Bayes' theorem

$$
\begin{aligned}
\pi(\mu|\mathscr{D}_n) &\propto \pi(\mu)\,\ell(\theta|\mathscr{D}_n) \\
&\propto \exp\{-\mu^2/2\sigma^2\}\,\exp\left\{-n(\bar{x}-\mu)^2/2\sigma^2\right\}
\end{aligned}
$$

$$\begin{aligned}&\propto \exp\left\{-(n+1)\mu^2/2\sigma^2+2n\mu\bar{x}/2\sigma^2\right\}\\&\propto \exp\left\{-(n+1)[\mu-n\bar{x}/(n+1)]^2/2\sigma^2\right\}\ ,\end{aligned}$$

which means that this posterior distribution in μ is a normal distribution with mean $n\bar{x}/(n+1)$ and variance $\sigma^2/(n+1)$. The mean (and mode) of the posterior is therefore different from the classical estimator $\bar{x}$, which may seem as a paradoxical feature of this Bayesian analysis. The reason for the difference is that the prior information that μ is close enough to zero is taken into account by the posterior distribution, which thus shrinks the original estimate towards zero. If we were given an alternative information that μ was close to ten, the posterior distribution would similarly shrink μ towards ten. The change from a factor n to a factor $(n+1)$ in the (posterior) variance is similarly explained by the prior information, in that accounting for this information reduces the variability of our answer.

For **normaldata**, we can first assume that the value of σ is the variability of the Michelson–Morley apparatus, namely 0.75. In that case, the posterior distribution on the fringe shift average μ is a normal $\mathscr{N}(n\bar{x}/(n+1),\sigma^2/(n+1))$ distribution, hence with mean and variance

```
> n=length(shift)
> mmu=sum(shift)/(n+1); mmu
[1] -0.01461538
> vmu=0.75^2/(n+1); vmu
[1] 0.008653846
```

represented on Fig. 2.3 as a dotted curve.

The case of a normal distribution with a known variance being quite unrealistic, we now consider the general case of an iid sample $\mathscr{D}_n=(x_1,\ldots,x_n)$ from the normal distribution $\mathscr{N}(\mu,\sigma^2)$ and $\theta=(\mu,\sigma^2)$. Keeping the same prior distribution $\mathscr{N}\left(0,\sigma^2\right)$ on μ, which then appears as a conditional distribution of μ given σ^2, *i.e.*, relies on the generic decomposition

$$\pi(\mu,\sigma^2)=\pi(\mu|\sigma^2)\pi(\sigma^2)\,,$$

we have to introduce a further prior distribution on σ^2. To make computations simple at this early stage, we choose an exponential $\mathscr{E}(1)$ distribution on σ^{-2}. This means that the random variable $\omega=\sigma^{-2}$ is distributed from an exponential $\mathscr{E}(1)$ distribution, the distribution on σ^2 being derived by the usual change of variable technique,

$$\pi(\sigma^2)=\exp(-\sigma^{-2})\left|\frac{\mathrm{d}\sigma^{-2}}{\mathrm{d}\sigma^2}\right|=\exp(-\sigma^{-2})\,(\sigma^2)^{-2}\,.$$

(This distribution is a special case of an inverse gamma distribution, namely $\mathcal{IG}(1,1)$.) The corresponding posterior density on θ is then given by

$$
\begin{aligned}
\pi((\mu,\sigma^2)|\mathscr{D}_n) &\propto \pi(\sigma^2)\times\pi(\mu|\sigma^2)\times\ell((\mu,\sigma^2)|\mathscr{D}_n)\\
&\propto (\sigma^{-2})^{1/2+2}\,\exp\left\{-(\mu^2{+}2)/2\sigma^2\right\}\\
&\quad\times(\sigma^{-2})^{n/2}\,\exp\left\{-\left(n(\mu-\overline{x})^2{+}s^2\right)/2\sigma^2\right\}\\
&\propto (\sigma^2)^{-(n+5)/2}\exp\left\{-\left[(n{+}1)(\mu-n\bar{x}/(n{+}1))^2{+}(2{+}s^2)\right]/2\sigma^2\right\}\\
&\propto (\sigma^2)^{-1/2}\exp\left\{-(n{+}1)[\mu-n\bar{x}/(n{+}1)]^2/2\sigma^2\right\}\,.\\
&\quad\times(\sigma^2)^{-(n+2)/2-1}\exp\left\{-(2{+}s^2)/2\sigma^2\right\}\,.
\end{aligned}
$$

Therefore, the posterior on θ can be decomposed as the product of an inverse gamma distribution on σ^2, $\mathscr{IG}((n+2)/2,[2+s^2]/2)$—which is the distribution of the inverse of a gamma $\mathscr{G}((n+2)/2,[2+s^2]/2)$ random variable—and, conditionally on σ^2, a normal distribution on μ, $\mathscr{N}(n\bar{x}/(n+1),\sigma^2/(n+1))$. The interpretation of this posterior is quite similar to the case when σ is known, with the difference that the variability in σ induces more variability in μ, the marginal posterior in μ being then a Student's t distribution[1] (Exercise 2.1)

$$\mu|\mathscr{D}_n \sim \mathscr{T}\left(n+2, n\bar{x}/(n+1), (2+s^2)/(n+1)(n+2)\right)\,,$$

with $n+2$ degrees of freedom, a location parameter proportional to $\bar{x}$ and a scale parameter (almost) proportional to s.

For **normaldata**, an $\mathscr{E}xp(1)$ prior on σ^{-2} being compatible with the value observed on the Michelson–Morley experiment, the parameters of the t distribution on μ are therefore $n=64$,

```
> mtmu=sum(shift)/(n+1);mtmu
[1] -0.01461538
> stmu=(2+(n-1)*var(shift))/((n+2)*(n+1));stmu
[1] 0.0010841496
```

We compare the resulting posterior with the one based on the assumption $\sigma=0.75$ on Fig. 2.3, using the `curve` commands (note that the `mnormt` library may require the preliminary installation of the corresponding package by `install.packages("mnormt")`):

```
> library(mnormt)
> curve(dmt(x,mean=mmu,S=stmu,df=n+2),col="chocolate2",lwd=2,
+ xlab="x",ylab="",xlim=c(-.5,.5))
> curve(dnorm(x,mean=mmu,sd=sqrt(vmu)),col="steelblue2",
+ lwd=2,add=TRUE,lty=2)
```

[1] We will omit the reference to Student in the subsequent uses of this distribution, as is the rule in anglo-saxon textbooks.

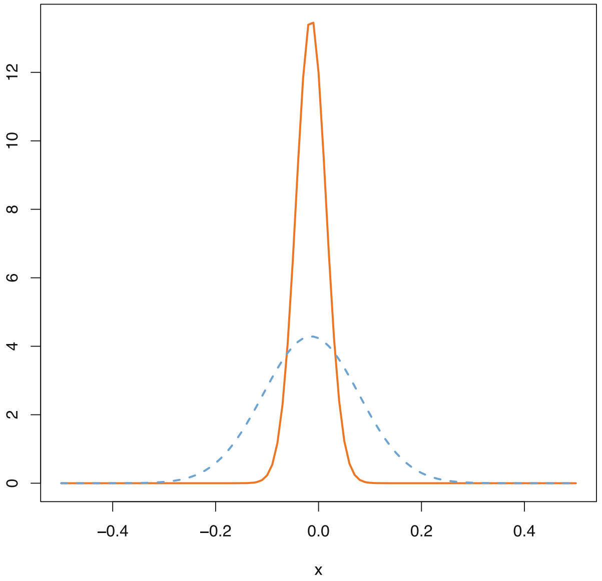

Fig. 2.3. Dataset **normaldata**: Two posterior distributions on μ corresponding to an hypothetical $\sigma = 0.75$ *(dashed lines)* and to an unknown σ^2 under the prior $\sigma^{-2} \sim \mathscr{E}(1)$ *(plain lines)*

Although this may sound counterintuitive, in this very case, estimating the variance produces a reduction in the variability of the posterior distribution on μ. This is because the postulated value of σ^2 is actually inappropriate for Illingworth's experiment, being far too large. Since the posterior distribution on σ^2 is an $\mathscr{IG}(33, 1.82)$ distribution for **normaldata**, the probability that σ is as large as 0.75 can be evaluated as

```
> digmma=function(x,shape,scale){dgamma(1/x,shape,scale)/x^2}
> curve(digmma(x,shape=33,scale=(1+(n+1)*var(shift))/2),
+ xlim=c(0,.2),lwd=2)
> pgamma(1/(.75)^2,shape=33,scale=(1+(n+1)*var(shift))/2)
[1] 8.99453e-39
```

which shows that 0.75 is quite unrealistic, being ten times as large as the mode of the posterior density on σ^2.

The above R command `library(mnormt)` calls the `mnormt` library, which contains useful additional functions related with multivariate normal and t distributions. In particular, `dmt` allows for location and scale parameters in the t distribution. Note also that s^2 is computed as `(n-1)*var(shift)` because R implicitly adopts a classical approach in using the "best unbiased estimator" of σ^2.

2.2.2 Bayesian Estimates

A concept that is at the core of Bayesian analysis is that one should provide an inferential assessment *conditional on the realized value of $\mathscr{D}_n$*. Bayesian analysis gives a proper probabilistic meaning to this conditioning by allocating to θ a probability distribution. Once the prior distribution is selected, Bayesian inference formally is "over"; that is, it is completely determined since the estimation, testing, and evaluation procedures are automatically provided by the prior and the way procedures are compared (or penalized). For instance, if estimations $\hat{\theta}$ of θ are compared via the sum of squared errors,

$$\mathrm{L}(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2 ,$$

the corresponding Bayes optimum is the *expected* value of θ under the posterior distribution,[2]

$$\hat{\theta} = \int \theta\, \pi(\theta|\mathscr{D}_n)\, \mathrm{d}\theta = \frac{\int \theta\, \ell(\theta|\mathscr{D}_n)\, \pi(\theta)\, \mathrm{d}\theta}{\int \ell(\theta|\mathscr{D}_n)\, \pi(\theta)\, \mathrm{d}\theta} , \tag{2.3}$$

for a given sample $\mathscr{D}_n$.

When no specific penalty criterion is available, the estimator (2.3) is often used as a default estimator, although alternatives are also available. For instance, the *maximum a posteriori estimator* (MAP) is defined as

$$\hat{\theta} = \arg\max_\theta \pi(\theta|\mathscr{D}_n) = \arg\max_\theta \pi(\theta)\ell(\theta|\mathscr{D}_n), \tag{2.4}$$

where the function to maximize is usually provided in closed form. However, numerical problems often make the optimization involved in finding the MAP far from trivial. Note also here the similarity of (2.4) with the maximum likelihood estimator (MLE): The influence of the prior distribution $\pi(\theta)$ on the estimate progressively disappears as the number of observations n increases, and the MAP estimator often recovers the asymptotic properties of the MLE.

For **normaldata**, since the posterior distribution on σ^{-2} is a $\mathscr{G}(32, 1.82)$ distribution, the posterior expectation of σ^{-2} given Illingworth's experimental data is $32/1.82 = 17.53$. The posterior expectation of σ^2 requires a supplementary effort in order to derive the mean of an inverse gamma distribution (see Exercise 2.2), namely

$$\mathbb{E}^\pi[\sigma^2|\mathscr{D}_n] = 1.82/(33-1) = 0.057 .$$

[2] Estimators are functions of the data $\mathscr{D}_n$, while estimates are values taken by those functions. In most cases, we will denote them with a "hat" symbol, the dependence on $\mathscr{D}_n$ being implicit.

Similarly, the MAP estimate is given here by

$$\arg\max_\theta \pi(\sigma^2|\mathscr{D}_n) = 1.82/(33+1) = 0.054$$

(see also Exercise 2.2). These values therefore reinforce our observation that the Michelson–Morley precision is not appropriate for the Illingworth experiment, which is much more precise indeed.

2.2.3 Conjugate Prior Distributions

The selection of the prior distribution is an important issue in Bayesian statistics. When prior information is available about the data or the model, it can (and must) be used in building the prior, and we will see some implementations of this recommendation in the following chapters. In many situations, however, the selection of the prior distribution is quite delicate, due to the absence of reliable prior information, and default solutions must be chosen instead. Since the choice of the prior distribution has a considerable influence on the resulting inference, this inferential step must be conducted with the utmost care.

From a computational viewpoint, the most convenient choice of prior distributions is to mimic the likelihood structure within the prior. In the most advantageous cases, priors and posteriors remain within the same parameterized family. Such priors are called *conjugate*. While the foundations of this principle are too advanced to be processed here (see, e.g., Robert, 2007, Chap. 3), such priors exist for most usual families, including the normal distribution. Indeed, as seen in Sect. 2.2.1, when the prior on a normal mean is normal, the corresponding posterior is also normal.

Since conjugate priors are such that the prior and posterior densities belong to the same parametric family, using the observations boils down to an update of the parameters of the prior. To avoid confusion, the parameters involved in the prior distribution on the model parameter are usually called *hyperparameters*. (They can themselves be associated with prior distributions, then called *hyperpriors*.)

For most practical purposes, it is sufficient to consider the conjugate priors described in Table 2.1. The derivation of each row is straightforward if painful and proceeds from the same application of Bayes' formula as for the normal case above (Exercise 2.5). For distributions that are not within this table, a conjugate prior may or may not be available (Exercise 2.6).

An important feature of conjugate priors is that one has a priori to select two hyperparameters, e.g., a mean and a variance in the normal case. On the one hand, this is an advantage when using a conjugate prior, namely that one has to select only a few parameters to determine the prior distribution. On the other hand, this is a drawback in that the information known a priori on μ may be either insufficient to determine both parameters or incompatible with the structure imposed by conjugacy.

Table 2.1. Conjugate priors for the most common statistical families

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathscr{N}(\theta,\sigma^2)$	Normal $\mathscr{N}(\mu,\tau^2)$	$\mathscr{N}(\rho(\sigma^2\mu+\tau^2x),\rho\sigma^2\tau^2)$ $\rho^{-1}=\sigma^2+\tau^2$
Poisson $\mathscr{P}(\theta)$	Gamma $\mathscr{G}(\alpha,\beta)$	$\mathscr{G}(\alpha+x,\beta+1)$
Gamma $\mathscr{G}(\nu,\theta)$	Gamma $\mathscr{G}(\alpha,\beta)$	$\mathscr{G}(\alpha+\nu,\beta+x)$
Binomial $\mathscr{B}(n,\theta)$	Beta $\mathscr{B}e(\alpha,\beta)$	$\mathscr{B}e(\alpha+x,\beta+n-x)$
Negative Binomial $\mathscr{N}eg(m,\theta)$	Beta $\mathscr{B}e(\alpha,\beta)$	$\mathscr{B}e(\alpha+m,\beta+x)$
Multinomial $\mathscr{M}_k(\theta_1,\ldots,\theta_k)$	Dirichlet $\mathscr{D}(\alpha_1,\ldots,\alpha_k)$	$\mathscr{D}(\alpha_1+x_1,\ldots,\alpha_k+x_k)$
Normal $\mathscr{N}(\mu,1/\theta)$	Gamma $\mathscr{G}(\alpha,\beta)$	$\mathscr{G}(\alpha+0.5,\beta+(\mu-x)^2/2)$

2.2.4 Noninformative Priors

There is no compelling reason to choose conjugate priors as our priors, except for their simplicity, but the restrictive aspect of conjugate priors can be attenuated by using *hyperpriors* on the hyperparameters themselves, although we will not deal with this additional level of complexity in the current chapter. The core message is therefore that conjugate priors are nice to work with, but require a hyperparameter determination that may prove awkward in some settings and that may moreover have a lasting impact on the resulting inference.

Instead of using conjugate priors, one can opt for a completely different perspective and rely on so-called *noninformative* priors that aim at attenuating the impact of the prior on the resulting inference. These priors are fundamentally defined as coherent extensions of the uniform distribution. Their purpose is to provide a reference measure that has as little as possible bearing on the inference (relative to the information brought by the likelihood). We first warn the reader that, for unbounded parameter spaces, the densities of noninformative priors actually fail to integrate to a finite number and they are defined instead as positive measures. While this sounds like an invalid extension of the probabilistic framework, it is quite correct to define the corresponding posterior distributions by (2.2), as long as the integral in the denominator is finite (almost surely). A more detailed account is for instance provided in Robert (2007, Sect. 1.5) about this possibility of using σ-finite measures (sometimes called *improper* priors) in settings where true probability prior distributions are too difficult to come by or too subjective to be accepted by all. For instance, *location models*

$$x \sim p(x-\theta)$$

are usually associated with flat priors $\pi(\theta)=1$ (note that these models include the normal $\mathscr{N}(\theta,1)$ as a special case), while *scale models*

$$x \sim \frac{1}{\theta} f\left(\frac{x}{\theta}\right)$$

are usually associated with the log-transform of a flat prior, that is,

$$\pi(\theta)=1/\theta\,.$$

In a more general setting, the (noninformative) prior favored by most Bayesians is the so-called *Jeffreys prior*,[3] which is related to the Fisher information matrix

$$I^F(\theta)=\text{var}_\theta\left(\frac{\partial \log f(X|\theta)}{\partial\theta}\right)$$

by

$$\pi^J(\theta)=\left|I^F(\theta)\right|^{1/2}\,,$$

where $|I|$ denotes the determinant of the matrix I.

Since the mean μ of a normal model is a location parameter, when the variance σ^2 is known, the standard choice of noninformative parameter is an arbitrary constant $\pi(\mu)$ (taken to be 1 by default). Given that this flat prior formally corresponds to the limiting case $\tau=\infty$ in the conjugate normal prior, it is easy to verify that this noninformative prior is associated with the posterior distribution $\mathscr{N}(x,1)$, which happens to be the likelihood function in that case. An interesting consequence of this observation is that the MAP estimator is also the maximum likelihood estimator in that (special) case. For the general case when $\theta=(\mu,\sigma^2)$, the Fisher information matrix leads to the Jeffreys prior $\pi^J(\theta)=1/\sigma^3$ (Exercise 2.4). The corresponding posterior distribution on (μ,σ^2) is then

$$\begin{aligned}\pi((\mu,\sigma^2)|\mathscr{D}_n) &\propto (\sigma^{-2})^{(3+n)/2}\,\exp\left\{-\left(n(\mu-\overline{x})^2+s^2\right)/2\sigma^2\right\}\\ &\propto \sigma^{-1}\exp\left\{-n(\mu-\bar{x})^2/2\sigma^2\right\}\times(\sigma^2)^{-(n+2)/2}\exp\left\{\frac{-s^2}{2\sigma^2}\right\}\,,\end{aligned}$$

that is,

$$\theta\sim\mathscr{N}\left(\bar{x},\sigma^2/n\right)\times\mathscr{IG}\left(n/2,s^2/2\right)\,.$$

a product of a conditional normal on μ by an inverse gamma on σ^2. Therefore the marginal posterior distribution on μ is a t distribution (Exercise 2.1)

$$\mu|\mathscr{D}_n\sim\mathscr{T}\left(n,\bar{x},s^2/n^2)\right)\,.$$

[3] Harold Jeffreys was an English geophysicist who developed and formalized Bayesian methods in the 1930s in order to analyze geophysical data. He ended up writing an influential treatise on Bayesian statistics entitled *Theory of Probability*.

For **normaldata**, the difference in Fig. 2.3 between the noninformative solution and the conjugate posterior is minor, but it expresses that the prior distribution $\mathscr{E}(1)$ on σ^{-2} is not very appropriate for the Illingworth experiment, since it does not put enough prior weight on the region of importance, i.e. near 0.05. As a result, the most concentrated posterior is (seemingly paradoxically) the one associated with the noninformative prior!

↯ A major (and potentially dangerous) difference between proper and improper priors is that the posterior distribution associated with an improper prior is not necessarily defined, that is, it may happen that

$$\int \pi(\theta)\ell(\theta|\mathscr{D}_n)\,\mathsf{d}\theta < \infty \tag{2.5}$$

does not hold. In some cases, this difficulty disappears when the sample size is large enough. In others (see Chap. 6), it may remain whatever the sample size. But the main thing is that, when using improper priors, condition (2.5) must always be checked.

2.2.5 Bayesian Credible Intervals

One point that must be clear from the beginning is that the Bayesian approach is a complete inferential approach. Therefore, it covers confidence evaluation, testing, prediction, model checking, and point estimation. We will progressively cover the different facets of Bayesian analysis in other chapters of this book, but we address here the issue of confidence intervals because it is rather a straightforward step from point estimation.

As with everything else, the derivation of the confidence intervals (or confidence regions in more general settings) is based on the posterior distribution $\pi(\theta|\mathscr{D}_n)$. Since the Bayesian approach processes θ as a random variable, a natural definition of a confidence region on θ is to determine $C(\mathscr{D}_n)$ such that

$$\pi(\theta \in C(\mathscr{D}_n)|\mathscr{D}_n) = 1 - \alpha \tag{2.6}$$

where α is a predetermined level such as 0.05.[4]

The important difference with a traditional perspective in (2.6) is that the integration is done over the parameter space, rather than over the observation space. The quantity $1 - \alpha$ thus corresponds to the probability that a random θ belongs to this set $C(\mathscr{D}_n)$, rather than to the probability that the random set contains the "true" value of θ. Given this drift in the interpretation of a

[4] There is nothing special about 0.05 when compared with, say, 0.87 or 0.12. It is just that the famous 5 % level is accepted by most as an acceptable level of error. If the context of the analysis tells a different story, another value for α (including one that may even depend on the data) should be chosen!

confidence set (rather called a *credible set* by Bayesians), the determination of the best[5] credible set turns out to be easier than in the classical sense: indeed, this set simply corresponds to the values of θ with the highest posterior values,

$$C(\mathscr{D}_n) = \{\theta;\, \pi(\theta|\mathscr{D}_n) \geq k_\alpha\}\ ,$$

where k_α is determined by the coverage constraint (2.6). This region is called the *highest posterior density* (HPD) region.

For **normaldata**, since the marginal posterior distribution on μ associated with the Jeffreys prior is the t distribution, $\mathscr{T}(n, \bar{x}, s^2/n^2)$,

$$\pi(\mu|\mathscr{D}_n) \propto \left[n(\mu - \bar{x})^2 + s^2\right]^{-(n+1)/2}$$

with $n = 64$ degrees of freedom. Therefore, due to the symmetry properties of the t distribution, the 95 % credible interval on μ is centered at $\bar{x}$ and its range is derived from the `0.975` quantile of the t distribution with n degrees of freedom,

```
> qt(.975,df=n)*sqrt((n-1)*var(shift)/n^2)
[1] 0.05082314
```

since the `mnormt` package does not compute quantiles. The resulting confidence interval is therefore given by

```
> qt(.975,df=n)*sqrt((n-1)*var(shift)/n^2)+mean(shift)
[1] 0.03597939
> -qt(.975,df=n)*sqrt((n-1)*var(shift)/n^2)+mean(shift)
[1] -0.06566689
```

i.e. equal to $[-0.066, 0.036]$. In conclusion, the value 0 belongs to this credible interval on μ and this (noninformative) Bayesian analysis of **normaldata** shows that, indeed, the absence of æther wind is not infirmed by Illingworth's experiment.

↯ While the shape of an optimal Bayesian confidence set is easily derived, the computation of either the bound k_α or the set $C(\mathscr{D}_n)$ may be too challenging to allow an analytic construction outside conjugate setups (see Exercise 2.11).

2.3 Bayesian Model Choice

Deciding the validity of some assumptions or restrictions on the parameter θ is a major part of the statistician's job. In classical statistics, this type of

[5] In the sense of producing the smallest possible volume with a given coverage.

problems goes under the name of *hypothesis testing*, following the framework set by Fisher, Neyman and Pearson in the 1930s. Hypothesis testing considers a decision problem where an hypothesis is either true or false and where the answer provided by the statistician is also a statement whether or not the hypothesis is true. However, we deem this approach to be too formalized—even though it can be directly reproduced from a Bayesian perspective, as shown in Robert (2007, Chap. 5)—, we strongly favour a model choice philosophy, namely that two or more models are proposed in parallel and assessed in terms of their respective fits of the data. This view acknowledges the fact that models are at best approximations of reality and it does not aim at finding a "true model", as hypothesis testing may do. In this book, we will thus follow the later approach and take the stand that inference problems expressed as hypothesis testing by the classical statisticians are in fact comparisons of different models. In terms of numerical outcomes, both perspectives—Bayesian hypothesis testing vs. Bayesian model choice—are exchangeable but we already warn the reader that, while the Bayesian solution is formally very close to a likelihood (ratio) statistic, its numerical values often strongly differ from the classical solutions.

2.3.1 The Model Index as a Parameter

The essential novelty when dealing with the comparison of models is that this issue makes the model itself an unknown quantity of interest. Therefore, if we are comparing two or more models with indices $k = 1, 2, \ldots, J$, we introduce a model indicator $\mathfrak{M}$ taking values in $\{1, 2, \ldots, J\}$ and representing the index of the "true" model. If $\mathfrak{M} = k$, then the data $\mathscr{D}_n$ are generated from a statistical model M_k with likelihood $\ell(\theta_k|\mathscr{D}_n)$ and parameter θ_k taking its value in a parameter space Θ_k. An obvious illustration is when opposing two standard parametric families, e.g., a normal family against a t family, in which case $J = 2$, $\Theta_1 = \mathbb{R} \times \mathbb{R}_+^*$—for mean and variance—and $\Theta_2 = \mathbb{R}_+^* \times \mathbb{R} \times \mathbb{R}_+^*$—for degree of freedom, mean and variance—, but this framework also includes soft or hard constraints on the parameters, as for instance imposing that a mean μ is positive.

In this setting, a natural Bayes procedure associated with a prior distribution π is to consider the posterior probability

$$\delta^\pi(\mathscr{D}_n) = \mathbb{P}^\pi(\mathfrak{M} = k|\mathscr{D}_n)\,,$$

i.e., the posterior probability that the model index is k, and select the index of the model with the highest posterior probability as the model preferred by the data $\mathscr{D}_n$. This representation implies that the prior π is defined over the collection of model indices, $\{1, 2, \ldots, J\}$, and, conditionally on the model index $\mathfrak{M}$, on the corresponding parameter space, Θ_k. This construction may sound both artificial and incomplete, as there is no prior on the parameter θ_k unless $\mathfrak{M} = k$, but it nonetheless perfectly translates the problem at hand:

inference on θ_k is meaningless unless this is the parameter of the correct model. Furthermore, the quantity of interest integrates out the parameter, since

$$\mathbb{P}^\pi(\mathfrak{M} = k|\mathscr{D}_n) = \frac{\mathbb{P}^\pi(\mathfrak{M} = k) \int \ell(\theta_k|\mathscr{D}_n)\pi_k(\theta_k)\,\mathrm{d}\theta_k}{\sum_{j=1}^J \mathbb{P}^\pi(\mathfrak{M} = j)\pi_j(\theta_j)\,\mathrm{d}\theta_j}\,.$$

↯ We believe it is worth emphasizing the above point: A parameter θ_k associated with a model does not have a statistical meaning outside this model. This means in particular that the notion of parameters "common to all models" often found in the literature, including the Bayesian literature, is not acceptable within a model choice perspective. Two models must have distinct parameters, if only because the purpose of the analysis is to end up with a single model.

The choice of the prior π is highly dependent on the value of the prior model probabilities $\mathbb{P}^\pi(\mathfrak{M} = k)$. In some cases, there is experimental or subjective evidence about those probabilities, but in others, we are forced to settle for equal weights $\mathbb{P}^\pi(\mathfrak{M} = k) = 1/J$. For instance, given a single observation $x \sim \mathscr{N}(\mu, \sigma^2)$ from a normal model where σ^2 is known, assuming $\mu \sim \mathscr{N}(\xi, \tau^2)$, the posterior distribution $\pi(\mu|x)$ is the normal distribution $\mathscr{N}(\xi(x), \omega^2)$ with

$$\xi(x) = \frac{\sigma^2\xi + \tau^2 x}{\sigma^2 + \tau^2} \quad \text{and} \quad \omega^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

If the question of interest is to decide whether μ is negative or positive, we can directly compute

$$\begin{aligned}\mathbb{P}^\pi(\mu < 0|x) &= \mathbb{P}^\pi\left(\frac{\mu - \xi(x)}{\omega} < \frac{-\xi(x)}{\omega}\right)\\ &= \Phi\left(-\xi(x)/\omega\right)\,, \qquad (2.7)\end{aligned}$$

where Φ is the normal cdf. This computation does not seem to follow from the principles we just stated but it is only a matter of perspective as we can derive the priors on both models from the original prior. Deriving this posterior probability indeed means that, a priori, μ is negative with probability $\mathbb{P}^\pi(\mu < 0) = \Phi(-\xi/\tau)$ and that, in this model, the prior on μ is the truncated normal

$$\pi_1(\mu) = \frac{\exp\{-(\mu - \xi)^2/2\tau^2\}}{\sqrt{2\pi}\tau\Phi(-\xi/\tau)}\,\mathbb{I}_{\mu<0}\,,$$

while μ is positive with probability $\Phi(\xi/\tau)$ and, in this second model, the prior on μ is the truncated normal

$$\pi_2(\mu) = \frac{\exp\{-(\mu - \xi)^2/2\tau^2\}}{\sqrt{2\pi}\tau\Phi(\xi/\tau)}\,\mathbb{I}_{\mu>0}\,.$$

The posterior probability of $\mathbb{P}^\pi(\mathfrak{M} = k|\mathscr{D}_n)$ is the core object in Bayesian model choice and, as indicated above, the default procedure is to select the model with the highest posterior probability. However, in decisional settings where the choice between two models has different consequences depending on the value of k, the boundary in $\mathbb{P}^\pi(\mathfrak{M} = k|\mathscr{D}_n)$ between choosing one model and the other may be far from 0.5. For instance, in a pharmaceutical trial, deciding to start production of a new drug does not have the same financial impact as deciding to run more preliminary tests. Changing the bound away from 0.5 is in fact equivalent to changing the prior probabilities of both models.

2.3.2 The Bayes Factor

A notion central to Bayesian model choice is the *Bayes factor*

$$B^\pi_{21}(\mathscr{D}_n) = \frac{\mathbb{P}^\pi(\mathfrak{M} = 2|\mathscr{D}_n)/\mathbb{P}^\pi(\mathfrak{M} = 1|\mathscr{D}_n)}{\mathbb{P}^\pi(\mathfrak{M} = 2)/\mathbb{P}^\pi(\mathfrak{M} = 1)}\,,$$

which corresponds to the classical odds or likelihood ratio, the difference being that the parameters are integrated rather than maximized under each model. While this quantity is a simple one-to-one transform of the posterior probability, it can be used for Bayesian model choice without first resorting to a determination of the prior weights of both models. Obviously, the Bayes factor depends on prior information through the choice of the model priors π_1 and π_2,

$$B^\pi_{21}(\mathscr{D}_n) = \frac{\int_{\Theta_2} \ell_2(\theta_2|\mathscr{D}_n)\pi_2(\theta_2)\,\mathrm{d}\theta_2}{\int_{\Theta_1} \ell_1(\theta_1|\mathscr{D}_n)\pi_1(\theta_1)\,\mathrm{d}\theta_1} = \frac{m_2(\mathscr{D}_n)}{m_1(\mathscr{D}_n)}\,,$$

and thus it can clearly be perceived as a Bayesian likelihood ratio which replaces the likelihoods with the marginals under both models.

The evidence brought by the data $\mathscr{D}_n$ can be calibrated using for instance Jeffreys' scale of evidence:

- if $\log_{21}(B^\pi_{21})$ is between 0 and 0.5, the evidence against model M_1 is *weak*,
- if it is between 0.5 and 1, it is *substantial*,
- if it is between 1 and 2, it is *strong*, and
- if it is above 2, it is *decisive*.

While this scale is purely arbitrary, it provides a reference for model assessment in a generic setting.

Consider now the special case when we want to assess whether or not a specific value of one of the parameters is appropriate, for instance $\mu = 0$ in the **normaldata** example. While the classical literature presents this problem as *a point null hypothesis*, we simply interpret it as the comparison of two models, $\mathscr{N}(0, \sigma^2)$ and $\mathscr{N}(\mu, \sigma^2)$, for Illingworth's data. In a more general framework, when the sample $\mathscr{D}_n$ is distributed as $\mathscr{D}_n \sim f(\mathscr{D}_n|\theta)$, if we decompose θ as $\theta = (\delta, \omega)$ and if the restricted model corresponds to the fixed value $\delta = \delta_0$, we define $\pi_1(\omega)$ as the prior under the restricted model (labelled M_1) and $\pi_2(\theta)$

as the prior under the unrestricted model (labelled M_2). The corresponding Bayes factor is then

$$B^{\pi}_{21}(\mathscr{D}_n) = \frac{\int_{\Theta} \ell(\theta|\mathscr{D}_n)\pi_2(\theta)\,\mathrm{d}\theta}{\int_{\Omega} \ell((\delta_0,\omega)|\mathscr{D}_n)\pi_1(\omega)\,\mathrm{d}\omega}$$

Note that, as hypotheses, point null problems often are criticized as artificial and impossible to test (in the sense of *how often can one distinguish* $\theta = 0$ *from* $\theta = 0.0001$ *?!*), but, from a model choice perspective, they simply correspond to more parsimonious models whose fit to the data can be checked against the fit produced by an unconstrained model. While the unconstrained model obviously contains values that produce a better fit, averaging over the whole parameter space Θ may still result in a small integrated likelihood $m_2(\mathscr{D}_n)$. The Bayes factor thus contains an automated penalization for complexity, a feature missed by the classical likelihood ratio statistic.

↯ In the very special case when the whole parameter is constrained to a fixed value, $\theta = \theta_0$, the marginal likelihood under model M_1 coincides with the likelihood $\ell(\theta_0|\mathscr{D}_n) = f(\mathscr{D}_n|\theta_0)$ and the Bayes factor simplifies in

$$B^{\pi}_{21}(\mathscr{D}_n) = \frac{\int_{\Theta} f(\mathscr{D}_n|\theta)\pi_2(\theta)\,\mathrm{d}\theta}{f(\mathscr{D}_n|\theta_0)}\,.$$

For $x \sim \mathscr{N}(\mu, \sigma^2)$ and σ^2 known, consider assessing $\mu = 0$ when $\mu \sim \mathscr{N}(0, \tau^2)$ under the alternative model (labelled M_2). The Bayes factor is the ratio

$$\begin{aligned} B^{\pi}_{21}(\mathscr{D}_n) &= \frac{m_2(x)}{f(x|(0,\sigma^2))} \\ &= \frac{\sigma}{\sqrt{\sigma^2+\tau^2}}\,\frac{e^{-x^2/2(\sigma^2+\tau^2)}}{e^{-x^2/2\sigma^2}} \\ &= \sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\,\exp\left\{\frac{\tau^2 x^2}{2\sigma^2(\sigma^2+\tau^2)}\right\}\,. \end{aligned}$$

Table 2.2 gives a sample of the values of the Bayes factor when the normalized quantity x/σ varies. They obviously depend on the choice of the prior variance τ^2 and the dependence is actually quite severe, as we will see below with the *Jeffreys–Lindley paradox.*

For **normaldata**, since we saw that setting σ to the Michelson–Morley value of 0.75 was producing a poor outcome compared with the noninformative solution, the comparison between the constrained and the unconstrained models is not very trustworthy, but as an illustration, it gives the following values:

Table 2.2. Bayes factor $B_{21}(z)$ against the null hypothesis $\mu = 0$ for different values of $z = x/\sigma$ and τ

z	0	0.68	1.28	1.96
$\tau^2 = \sigma^2$	0.707	0.794	1.065	1.847
$\tau^2 = 10\sigma^2$	0.302	0.372	0.635	1.728

```
> BaFa=function(z,rat){
#rat denotes the ratio tau^2/sigma^2
sqrt(1/(1+rat))*exp(z^2/(2*(1+1/rat)))}
> BaFa(mean(shift),1)
[1] 0.7071767
> BaFa(mean(shift),10)
[1] 0.3015650
```

which supports the constraint $\mu = 0$ for those two values of τ, since the Bayes factor is less than 1. (For this dataset, the Bayes factor is always less than one, see Exercise 2.13.)

2.3.3 The Ban on Improper Priors

We introduced noninformative priors in Sect. 2.2.4 as a way to handle situations when the prior information was not sufficient to build proper priors. We also saw that, for **normaldata**, a noninformative prior was able to exhibit conflicts between the prior information (based on the Michelson–Morley experiment) and the data (resulting from Illingworth's experiment). Unfortunately, the use of noninformative priors is very much restricted in model choice settings because the fact that they usually are improper leads to the impossibility of comparing the resulting marginal likelihoods.

Looking at the expression of the Bayes factor,

$$B_{21}^{\pi}(\mathscr{D}_n) = \frac{\int_{\Theta_2} \ell_2(\theta_2|\mathscr{D}_n)\pi_2(\theta_2)\,\mathrm{d}\theta_2}{\int_{\Theta_1} \ell_1(\theta_1|\mathscr{D}_n)\pi_1(\theta_1)\,\mathrm{d}\theta_1}\,,$$

it is clear that, when either π_1 or π_2 are improper, it is impossible to normalize the improper measures in a unique manner. Therefore, the Bayes factor becomes completely arbitrary since it can be multiplied by one or two arbitrary constants.

For instance, when comparing $x \sim \mathscr{N}(\mu, 1)$ (model M_1) with $x \sim \mathscr{N}(0, 1)$ (model M_2), the improper Jeffreys prior on model M_1 is $\pi_1(\mu) = 1$. The Bayes factor corresponding to this choice is

$$B_{12}^{\pi}(x) = \frac{e^{-x^2/2}}{\int_{-\infty}^{+\infty} e^{-(x-\theta)^2/2}\,\mathrm{d}\theta} = \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,.$$

If, instead, we use the prior $\pi_1(\mu) = 100$, the Bayes factor becomes

$$B_{12}^{\pi}(x) = \frac{e^{-x^2/2}}{100 \int_{-\infty}^{+\infty} e^{-(x-\theta)^2/2}\,\mathrm{d}\theta} = \frac{e^{-x^2/2}}{100\sqrt{2\pi}}$$

and is thus one-hundredth of the previous value! Since there is no mathematical way to discriminate between $\pi_1(\mu) = 1$ and $\pi_1(\mu) = 100$, the answer clearly is non-sensical.

Note that, if we are instead comparing model M_1 where $\mu \le 0$ and model M_2 where $\mu > 0$, then the posterior probability of model M_1 under the flat prior is

$$\mathbb{P}^{\pi}(\mu \le 0|x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-(x-\theta)^2/2}\,\mathrm{d}\theta = \Phi(-x)\,,$$

which is uniquely defined.

The difficulty in using an improper prior also relates to what is called the *Jeffreys–Lindley paradox*, a phenomenon that shows that limiting arguments are not valid in testing settings. In contrast with estimation settings, the noninformative prior no longer corresponds to the limit of conjugate inferences. For instance, for the comparison of the normal $x \sim \mathscr{N}(\mu, \sigma^2)$ (model M_1) and of the normal $x \sim \mathscr{N}(\mu, \sigma^2)$ (model M_2) models when σ^2 is known, using a conjugate prior $\mu \sim \mathscr{N}(0, \tau^2)$, the Bayes factor

$$B_{21}^{\pi}(x) = \sqrt{\frac{\sigma^2}{\sigma^2 + \tau^2}} \exp\left[\frac{\tau^2 x^2}{2\sigma^2(\sigma^2 + \tau^2)}\right]$$

converges to 0 when τ goes to $+\infty$, for *every* value of x, again a non-sensical procedure.

Since improper priors are an essential part of the Bayesian approach, there are many proposals found in the literature to overcome this ban. Most of those proposals rely on a device that transforms the improper prior into a proper probability distribution by exploiting a fraction of the data $\mathscr{D}_n$ and then restricts itself to the remaining part of the data to run the test as in a standard situation. The variety of available solutions is due to the many possibilities of removing the dependence on the choice of the portion of the data used in the first step. The resulting procedures are called *pseudo-Bayes factors*, although some may actually correspond to true Bayes factors. See Robert (2007, Chap. 5) for more details, although we do not advocate using those procedures.

There is a major exception to this ban on improper priors that we can exploit. If both models under comparison have parameters that have similar enough meanings to share the same prior distribution, as for instance a measurement error σ^2, then the normalization issue vanishes. Note that we are not assuming that parameters are *common* to both models and thus that we do not contradict the earlier warning about different parameters to different models. An illustration is provided by the above remark on the comparison of $\mu < 0$ with $\mu > 0$. This partial opening in the use of improper priors represents an opportunity but it does not apply to parameters of interest, i.e. to parameters on which restrictions are assessed.

Example 2.1. When comparing two id normal samples, $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, with respective distributions $\mathscr{N}(\mu_x, \sigma^2)$ and $\mathscr{N}(\mu_y, \sigma^2)$, we can examine whether or not the two means are identical, i.e. $\mu_x = \mu_y$ (corresponding to model M_1). To take advantage of the structure of this model, we can assume that σ^2 is a measurement error with a similar meaning under both models and thus that the same prior $\pi_\sigma(\sigma^2)$ can be used under both models. This means that the Bayes factor

$$B_{21}^\pi(\mathscr{D}_n) = \frac{\int \ell_2(\mu_x, \mu_y, \sigma|\mathscr{D}_n)\pi(\mu_x, \mu_y)\pi_\sigma(\sigma^2)\, \mathrm{d}\sigma^2\, \mathrm{d}\mu_x\, \mathrm{d}\mu_y}{\int \ell_1(\mu, \sigma|\mathscr{D}_n)\pi_\mu(\mu)\pi_\sigma(\sigma^2)\, \mathrm{d}\sigma^2\, \mathrm{d}\mu}$$

does not depend on the normalizing constant used for $\pi_\sigma(\sigma^2)$ and thus that we can still use an improper prior such as $\pi_\sigma(\sigma^2) = 1/\sigma^2$ in that case. Furthermore, we can rewrite μ_x and μ_y as $\mu_x = \mu - \xi$ and $\mu_y = \mu + \xi$, respectively, and use a prior of the form $\pi(\mu, \xi) = \pi_\mu(\mu)\pi_\xi(\xi)$ on the new parameterization so that, again, the same prior π_μ can be used under both models. The same cancellation of the normalizing constant occurs for π_μ, which means a Jeffreys prior $\pi_\mu(\mu) = 1$ can be used. However, we need a proper and well-defined prior on ξ, for instance $\xi \sim \mathscr{N}(0, \tau^2)$, which leads to

$$\begin{aligned} B_{21}^\pi(\mathscr{D}_n) &= \frac{\displaystyle\int e^{-n\left[(\mu-\xi-\bar{x})^2+(\mu+\xi-\bar{y})^2+s_{xy}^2\right]/2\sigma^2}\, \sigma^{-2n-2} e^{-\xi^2/2\tau^2}/\tau\sqrt{2\pi}\, \mathrm{d}\sigma^2\, \mathrm{d}\mu\, \mathrm{d}\xi}{\displaystyle\int e^{-n\left[(\mu-\bar{x})^2+(\mu-\bar{y})^2+s_{xy}^2\right]/2\sigma^2}\, \sigma^{-2n-2}\, \mathrm{d}\sigma^2\, \mathrm{d}\mu} \\ &= \frac{\displaystyle\int \left[(\mu-\xi-\bar{x})^2+(\mu+\xi-\bar{y})^2+s_{xy}^2\right]^{-n} e^{-\xi^2/2\tau^2}/\tau\sqrt{2\pi}\, \mathrm{d}\mu\, \mathrm{d}\xi}{\displaystyle\int \left[(\mu-\bar{x})^2+(\mu-\bar{y})^2+s_{xy}^2\right]^{-n} \mathrm{d}\mu}, \end{aligned}$$

where s_{xy}^2 denotes the average

$$s_{xy}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2 .$$

While the denominator can be completely integrated out, the numerator cannot. A numerical approximation to B_{21}^π is thus necessary. (This issue is addressed in Sect. 2.4.) ◀

We conclude this section by a full processing of the assessment of $\mu = 0$ for the single sample normal problem. Comparing models $M_1 :\ \mathscr{N}(0, \sigma^2)$ under the prior $\pi_1(\sigma^2) = 1/\sigma^2$ and $M_2 :\ \mathscr{N}(\mu, \sigma^2)$ under the prior made of $\pi_2(\sigma^2) = 1/\sigma^2$ and $\pi_2(\mu|\sigma^2)$ equal to the normal $\mathscr{N}(0, \sigma^2)$ density, the Bayes factor is

$$
\begin{aligned}
B^\pi_{21}(\mathscr{D}_n) &= \frac{\displaystyle\int e^{-[n(\bar{x}-\mu)^2+s^2]/2\sigma^2}\, e^{-\mu^2/2\sigma^2}\, \sigma^{-n-1-2}\, \frac{\mathrm{d}\mu\mathrm{d}\sigma^2}{\sqrt{2\pi}}}{\displaystyle\int e^{-[n\bar{x}^2+s^2]/2\sigma^2}\, \sigma^{-n-2}\, \mathrm{d}\sigma^2} \\
&= \frac{\displaystyle\int e^{-(n+1)[\mu-n\bar{x}/(n+1)]^2}\, e^{-[n\bar{x}^2/(n+1)+s^2]/2\sigma^2}\, \sigma^{-n-3}\, \frac{\mathrm{d}\mu\mathrm{d}\sigma^2}{\sqrt{2\pi}}}{\displaystyle\left[\frac{n\bar{x}^2+s^2}{2}\right]^{-n/2} \Big/ \Gamma(n/2)} \\
&= \frac{\displaystyle\int (n+1)^{-1/2}\, e^{-[n\bar{x}^2/(n+1)+s^2]/2\sigma^2}\, \sigma^{-n-2}\, \mathrm{d}\sigma^2}{\displaystyle\left[\frac{n\bar{x}^2+s^2}{2}\right]^{-n/2} \Big/ \Gamma(n/2)} \\
&= \frac{\displaystyle(n+1)^{-1/2}\left[\frac{n\bar{x}^2/(n+1)+s^2}{2}\right]^{-n/2} \Big/ \Gamma(n/2)}{\displaystyle\left[\frac{n\bar{x}^2+s^2}{2}\right]^{-n/2} \Big/ \Gamma(n/2)} \\
&= (n+1)^{-1/2}\left[\frac{n\bar{x}^2+s^2}{n\bar{x}^2/(n+1)+s^2}\right]^{n/2},
\end{aligned}
$$

taking once again advantage of the normalizing constant of the gamma distribution (see also Exercise 2.8). It therefore increases to infinity with $\bar{x}^2/s^2$, starting from $1/\sqrt{n+1}$ when $\bar{x}=0$.

The value of this Bayes factor for Illingworth's data is given by

```
> ratio=n*mean(shift)^2/((n-1)*var(shift))
> ((1+ratio)/(1+ratio/(n+1)))^(n/2)/sqrt(n+1)
[1] 0.1466004
```

which confirms the assessment that the model with $\mu=0$ is to be preferred.

2.4 Monte Carlo Methods

While, as seen in Sect. 2.3, the Bayes factor and the posterior probability are the only quantities used in the assessment of models (and hypotheses), the analytical derivation of those objects is not always possible, since they involve integrating the likelihood $\ell(\theta|\mathscr{D}_n)$ both on the sets Θ_1 and Θ_2, under the respective priors π_1 and π_2. Fortunately, there exist special numerical techniques for the computation of Bayes factors, which are, mathematically speaking, simply ratios of integrals. We now detail the techniques used in the approximation of intractable integrals, but refer to Chen et al. (2000) and Robert and Casella (2004, 2009) for book-length presentations.

2.4.1 An Approximation Based on Simulations

The technique that is most commonly used for integral approximations in statistics is called the Monte Carlo method[6] and relies on computer simulations of random variables to produce an approximation technique that converges with the number of simulations. Its justification is thus the *law of large numbers*, that is, if $x_1, \ldots, x_N$ are independent and distributed from g, then the empirical average

$$\hat{\mathfrak{I}}_N = (h(x_1) + \ldots + h(x_N))/N$$

converges (almost surely) to the integral

$$\mathfrak{I} = \int h(x) g(x) \, \mathrm{d}x \, .$$

We will not expand on the foundations of the random number generators in this book, except for an introduction to accept–reject methods in Chap. 5 because of their links with Markov chain Monte Carlo techniques (see, instead, Robert and Casella, 2004). The connections of utmost relevance here are (a) that softwares like R can produce pseudo-random series that are indistinguishable from truly random series with a given distribution, as illustrated in Table 1.1 and (b) that those software packages necessarily cover a limited collection of distributions. Therefore, other methods must be found for simulating distributions outside this collection, while relying on the distributions already available, first and foremost the uniform $\mathscr{U}(0,1)$ distribution.

The implementation of the Monte Carlo method is straightforward, at least on a formal basis, with the following algorithmic representation:

Algorithm 2.1 Basic Monte Carlo Method

```
For $i = 1, \ldots, N$,
    simulate $x_i \sim g(x)$.
Take
    $\hat{\mathfrak{I}}_N = (h(x_1) + \ldots + h(x_N))/N$
to approximate $\mathfrak{I}$.
```

as long as the (computer-generated) pseudo-random generation from g is feasible and the $h(x_i)$ values are computable. When simulation from g is a problem because g is nonstandard and usual techniques such as accept–reject algorithms (see Chap. 5) are difficult to devise, more advanced techniques such as Markov Chain Monte Carlo (MCMC) are required. We will introduce those

[6] This method is named in reference to the central district of Monaco, where the famous Monte-Carlo casino lies.

in both next chapters. When the difficulty is with the intractability of the function h, the solution is often to use an integral representation of h and to expand the random variables x_i in (x_i, y_i), where y_i is an auxiliary variable. The use of such representations will be detailed in Chap. 6.

Example 2.2 (Continuation of Example 2.1). As computed in Example 2.1, the Bayes factor $B_{21}^{\pi}(\mathscr{D}_n)$ can be simplified into

$$B_{21}^{\pi}(\mathscr{D}_n) = \frac{\displaystyle\int \left[(\mu-\xi-\bar{x})^2 + (\mu+\xi-\bar{y})^2 + s_{xy}^2\right]^{-n} e^{-\xi^2/2\tau^2}\,\mathrm{d}\mu\,\mathrm{d}\xi/\tau\sqrt{2\pi}}{\displaystyle\int \left[(\mu-\bar{x})^2 + (\mu-\bar{y})^2 + s_{xy}^2\right]^{-n} \mathrm{d}\mu}$$

$$= \frac{\displaystyle\int \left[(2\xi+\bar{x}-\bar{y})^2 + 2\,s_{xy}^2\right]^{-n+1/2} e^{-\xi^2/2\tau^2}\,\mathrm{d}\xi/\tau\sqrt{2\pi}}{\left[(\bar{x}-\bar{y})^2 + 2\,s_{xy}^2\right]^{-n+1/2}}\,,$$

and we are left with a single integral in the numerator that involves the normal $\mathscr{N}(0,\tau^2)$ density and can thus be represented as an expectation against this distribution. This means that simulating a normal $\mathscr{N}(0,\tau^2)$ sample of ξ_i's $(i = 1, \ldots, N)$ and replacing $B_{21}^{\pi}(\mathscr{D}_n)$ with

$$\hat{B}_{21}^{\pi}(\mathscr{D}_n) = \frac{\frac{1}{N}\sum_{i=1}^{N} \left[(2\xi_i+\bar{x}-\bar{y})^2 + 2\,s_{xy}^2 +^2\right]^{-n+1/2}}{\left[(\bar{x}-\bar{y})^2 + 2\,s_{xy}^2\right]^{-n+1/2}}$$

is an asymptotically valid approximation scheme. ◀

In **normaldata**, if we compare the fifth and the sixth sessions, both with $n = 10$ observations, we obtain

```
> illing=as.matrix(normaldata)
> xsam=illing[illing[,1]==5,2]
> xbar=mean(xsam)
[1] -0.041
> ysam=illing[illing[,1]==6,2]
> ybar=mean(ysam)
[1] -0.025
> Ssquar=9*(var(xsam)+var(ysam))/10
[1] 0.101474
```

Picking $\tau = 0.75$ as earlier, we get the following approximation to the Bayes factor

```
> Nsim=10^4
> tau=0.75
> xis=rnorm(Nsim,sd=tau)
> BaFa=mean(((2*xis+xbar-ybar)^2+2*Ssquar)^(-8.5))/
+ ((xbar-ybar)^2+2*Ssquar)^(-8.5)
[1] 0.0763622
```

This value of $\widehat{B_{21}^{\pi}}(\mathscr{D}_n)$ implies that $\xi = 0$, i.e. $\mu_x = \mu_y$ is much more likely for the data at hand than $\mu_x \neq \mu_y$. Note that, if we use $\tau = 0.1$ instead, the approximated Bayes factor is 0.4985 which slightly reduces the argument in favor of $\mu_x = \mu_y$.

Obviously, this *Monte Carlo estimate* of $\mathfrak{I}$ is not exact, but generating a sufficiently large number of random variables can render this approximation error arbitrarily small in a suitable probabilistic sense. It is also possible to assess the size of this error for a given number of simulations. If

$$\int |h(x)|^2 g(x)\,\mathrm{d}x < \infty\,,$$

the central limit theorem shows that $\sqrt{N}\,[\hat{\mathfrak{I}}_N - \mathfrak{I}]$ is also normally distributed, and this can be used to construct asymptotic confidence regions for $\hat{\mathfrak{I}}_N$, estimating the asymptotic variance from the simulation output.

For the approximation of $B_{21}^{\pi}(\mathscr{D}_n)$ proposed above, its variability is illustrated in Fig. 2.4, based on 500 replications of the simulation of $N = 1000$ normal variables used in the approximation and obtained as follows

```
> xis=matrix(rnorm(500*10^3,sd=tau),nrow=500)
> BF=((2*xis+xbar-ybar)^2+2*Ssquar)^(-8.5)/
+ ((xbar-ybar)^2+2*Ssquar)^(-8.5)
> estims=apply(BF,1,mean)
> hist(estims,nclass=84,prob=T,col="wheat2",
+ main="",xlab="Bayes Factor estimates")
> curve(dnorm(x,mean=mean(estims),sd=sd(estims)),
+ col="steelblue2",add=TRUE)
```

As can be seen on this figure, the value of 0.076 reported in the previous Monte Carlo approximation is in the middle of the range of possible values. More in connection with the above point, the shape of the histogram is clearly compatible with the normal approximation, as shown by the fitted normal density.

2.4.2 Importance Sampling

An important feature of Example 2.2 is that, for the Monte Carlo approximation of $B_{21}^{\pi}(\mathscr{D}_n)$, we exhibited a normal density within the integral and hence derived a representation of this integral as an expectation under this normal distribution. This seems like a very restrictive constraint in the approximation of integrals but this is only an apparent restriction in that we will now show that there is no need to simulate directly from the normal density and furthermore that there is no intrinsic density corresponding to a given integral, but rather an infinity of densities!

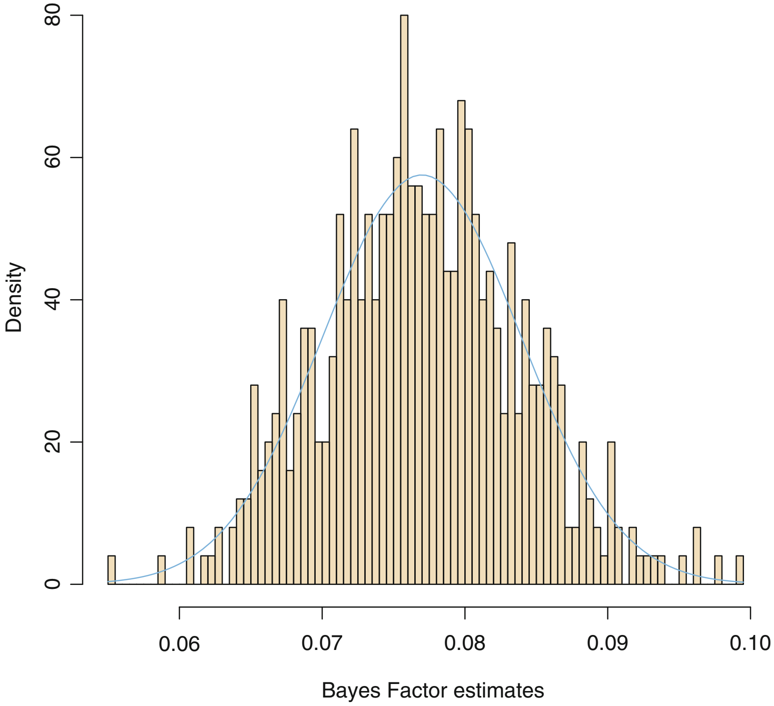

Fig. 2.4. Dataset **normaldata**: Histogram of 500 realizations of the approximation $\widehat{B_{21}(\mathscr{D}_n)}$ based on $N = 1000$ simulations each and normal fit of the sample

Indeed, an arbitrary integral

$$\mathfrak{I} = \int H(x)\,\mathrm{d}x$$

can be represented in infinitely many ways as an expectation, since, for an arbitrary probability density γ, we always have

$$\mathfrak{I} = \int \frac{H(x)}{\gamma(x)}\,\gamma(x)\,\mathrm{d}x\,, \tag{2.8}$$

under the minimal condition that $\gamma(x) > 0$ when $H(x)$. Therefore, the generation of a sample from γ can provide a converging approximation to $\mathfrak{E}$ and the Monte Carlo method applies in a very wide generality. This method is called *importance sampling* when applied to an expectation under a density g,

$$\mathfrak{I} = \int h(x)g(x)\,\mathrm{d}x\,, H(x) = h(x)g(x)$$

since the values x_i simulated from γ are weighted by the importance weights $g(x_i)/\gamma(x_i)$ in the approximation

$$\hat{\mathfrak{I}}_N = \frac{1}{N} \sum_{i=1}^{N} \frac{g(x_i)}{\gamma(x_i)}\, h(x_i)\,.$$

↯ While the representation (2.8) holds for any density γ with a support larger than the support of H, the performance of the empirical average $\hat{\mathfrak{I}}_N$ can deteriorate considerably when the ratio $h(x)g(x)/\gamma(x)$ is not bounded as this raises the possibility for infinite variance in the resulting estimator. When using importance sampling, one must always take heed of a potentially infinite variance of $\hat{\mathfrak{I}}_N$.

An additional incentive in using importance sampling is that this method does not require the density g (or γ) to be known completely. Those densities can be known only up to a normalizing constant, $g(x) \propto \tilde{g}(x)$ and $\gamma(x) \propto \tilde{\gamma}(x)$, since the ratio

$$\sum_{i=1}^{n} h(x_i)\tilde{g}(x_i)/\tilde{\gamma}(x_i) \bigg/ \sum_{i=1}^{n} \tilde{g}(x_i)/\tilde{\gamma}(x_i)$$

also converges to $\mathfrak{I}$ when n goes to infinity and when the x_i's are generated from γ.

The equivalent of Algorithm 2.1 for importance sampling is as follows:

Algorithm 2.2 IMPORTANCE SAMPLING METHOD

```
For $i = 1, \ldots, N$,
        simulate $x_i \sim \gamma(x)$;
        compute $\omega_i = \tilde{g}(x_i)/\gamma(x_i)$ .
Take
        $$\hat{\mathfrak{I}}_N = \sum_{i=1}^{N} \omega_i\, h(x_i) \bigg/ \sum_{i=1}^{N} \omega_i$$
to approximate $\mathfrak{I}$.
```

This algorithm is straightforward to implement. Since it offers a degree of freedom in the selection of γ, simulation from a manageable distribution can be imposed, keeping in mind the constraint that γ should have flatter tails than g. Unfortunately, as the dimension of x increases, differences between the target density g and the importance density γ have a larger and larger impact.

Example 2.3. Consider almost the same setting as in Exercise 2.11: $\mathscr{D}_n = (x_1, \ldots, x_n)$ is an iid sample from $\mathscr{C}(\theta, 1)$ and the prior on θ is a flat prior. We can use a normal importance function from a $\mathscr{N}(\mu, \sigma^2)$ distribution to produce a sample $\theta_1, \ldots, \theta_N$ that approximates the Bayes estimator of θ, i.e. its posterior mean, by

$$\hat{\delta}^{\pi}(\mathscr{D}_n) = \frac{\sum_{t=1}^{N} \theta_t \exp\left\{(\theta_t - \mu)^2/2\right\} \prod_{i=1}^{n} [1 + (x_i - \theta_t)^2]^{-1}}{\sum_{t=1}^{N} \exp\left\{(\theta_t - \mu)^2/2\right\} \prod_{i=1}^{n} [1 + (x_i - \theta_t)^2]^{-1}} \,.$$

But this is a very poor estimation (see Exercise 2.17 for an analytic explanation) and it degrades considerably when μ increases. If we run an R simulation experiment producing a sample of estimates when μ increases, as follows,

```
> Nobs=10
> obs=rcauchy(Nobs)
> Nsim=250
> Nmc=500
> sampl=matrix(rnorm(Nsim*Nmc),nrow=1000) # normal samples
> raga=riga=matrix(0,nrow=50,ncol=2) # ranges
> mu=0
> for (j in 1:50){
+   prod=1/dnorm(sampl-mu) # importance sampling
+   for (i in 1:Nobs)
+     prod=prod*dt(obs[i]-sampl,1)
+   esti=apply(sampl*prod,2,sum)/apply(prod,2,sum)
+   raga[j,]=range(esti)
+   riga[j,]=c(quantile(esti,.025),quantile(esti,.975))
+   sampl=sampl+0.1
+   mu=mu+0.1
+   }
> mus=seq(0,4.9,by=0.1)
> plot(mus,0*mus,col="white",xlab=expression(mu),
+ ylab=expression(hat(theta)),ylim=range(raga))
> polygon(c(mus,rev(mus)),c(raga[,1],rev(raga[,2])),col="grey50")
> polygon(c(mus,rev(mus)),c(riga[,1],rev(riga[,2])),col="pink3")
```

as shown by Fig. 2.5, not only does the range of the approximation increase, but it ends up missing the true value $\theta = 0$ when μ is far enough from 0. ◀

2.4.3 Approximation of Bayes Factors

Bayes factors being ratios of integrals, they can be approximated by regular importance sampling tools. However, given their specificity as ratios of marginal likelihoods, hence of normalizing constants of the posterior distributions, there exist more specialized techniques, including a fairly generic method called *bridge sampling*, developed by Gelman and Meng (1998).

When comparing two models with sampling densities $f_1(\mathscr{D}_n|\theta_1)$ (model M_1) and $f_2(\mathscr{D}_n|\theta_2)$ (model M_2), assume that both models share the same parameter space Θ. This is for instance the case when comparing the fit of two densities with the same number of parameters (modulo a potential reparameterization of one of the models). In this setting, if the corresponding prior

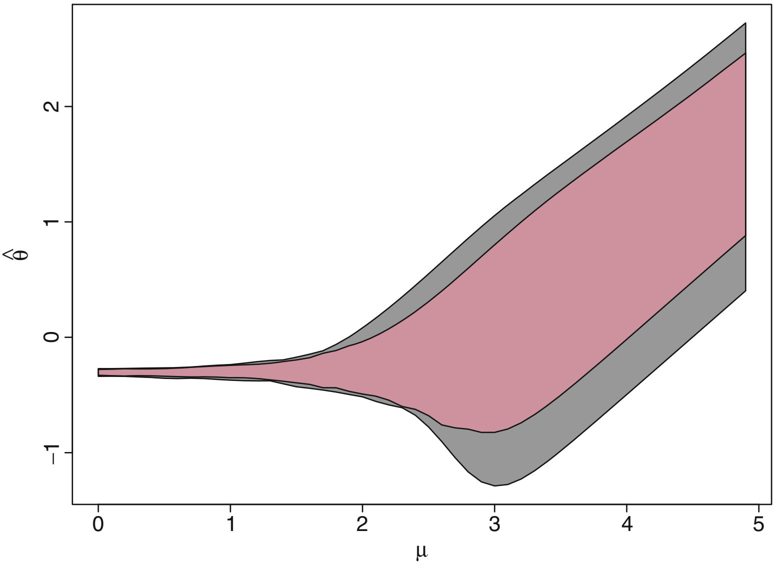

Fig. 2.5. Representation of the whole range (*grey*) and of the 95 % range (*pink*) of variation of the importance sampling approximation to the Bayes estimate for $n = 10$ observations from the $\mathscr{C}(0, 1)$ distribution and $N = 250$ simulations of θ from a $\mathscr{N}(\mu, 1)$ distribution as a function of μ. This range is computed using 500 replications of the importance sampling estimates

densities are $\pi_1(\theta)$ and $\pi_2(\theta)$, we only know the unnormalized posterior densities $\tilde{\pi}_1(\theta|\mathscr{D}_n) = f_1(\mathscr{D}_n|\theta)\pi_1(\theta)$ and $\tilde{\pi}_2(\theta|\mathscr{D}_n) = f_2(\mathscr{D}_n|\theta)\pi_2(\theta)$. In this general setting, for any positive function α such that the integrals below exist, the Bayes factor for comparing the two models satisfies

$$\begin{aligned} B_{12}^{\pi}(\mathscr{D}_n) &= \frac{m_1(x)}{m_2(x)} \\ &= \frac{m_1(x)}{m_2(x)} \frac{\displaystyle\int \tilde{\pi}_1(\theta|\mathscr{D}_n)\alpha(\theta)\tilde{\pi}_2(\theta|\mathscr{D}_n)\mathrm{d}\theta}{\displaystyle\int \tilde{\pi}_2(\theta|\mathscr{D}_n)\alpha(\theta)\tilde{\pi}_1(\theta|\mathscr{D}_n)\mathrm{d}\theta} \\ &= \frac{\displaystyle\int \tilde{\pi}_1(\theta|\mathscr{D}_n)\alpha(\theta)\pi_2(\theta|\mathscr{D}_n)\mathrm{d}\theta}{\displaystyle\int \tilde{\pi}_2(\theta|\mathscr{D}_n)\alpha(\theta)\pi_1(\theta|\mathscr{D}_n)\mathrm{d}\theta}\,. \end{aligned} \tag{2.9}$$

Therefore, the *bridge sampling* approximation

$$\sum_{i=1}^{N} \tilde{\pi}_1(\theta_{2i}|\mathscr{D}_n)\alpha(\theta_{2i}) \Big/ \sum_{i=1}^{N} \tilde{\pi}_2(\theta_{1i}|\mathscr{D}_n)\alpha(\theta_{1i}) \tag{2.10}$$

is a convergent approximation of the Bayes factor $B^{\pi}_{12}(\mathscr{D}_n)$ when $\theta_{ji} \sim \pi_j(\theta|\mathscr{D}_n)$ $(j = 1, 2,\ i = 1, \ldots, N)$. One of the appealing features of the method is that it only requires simulations from the posterior distributions under both models of interest. Another interesting feature is that α is completely arbitrary, which means it can be chosen in the best possible way. Using asymptotic variance arguments, Gelman and Meng (1998) proved that the best choice is

$$\alpha^{\mathrm{O}}(\theta) \propto \frac{1}{\pi_1(\theta|\mathscr{D}_n) + \pi_2(\theta|\mathscr{D}_n)}\,,$$

which bridges both posteriors. This means that the optimal weight of θ_{2i} in (2.10) is

$$\frac{\tilde{\pi}_1(\theta_{2i}|\mathscr{D}_n)}{\pi_1(\theta_{2i}|\mathscr{D}_n) + \pi_2(\theta_{2i}|\mathscr{D}_n)} = \frac{\tilde{\pi}_1(\theta_{2i}|\mathscr{D}_n)}{\tilde{\pi}_1(\theta_{2i}|\mathscr{D}_n) + B^{\pi}_{12}(\mathscr{D}_n)\tilde{\pi}_2(\theta_{2i}|\mathscr{D}_n)}\,,$$

with an appropriate change of indices for the θ_{1i}'s. There is however a caveat with this find in that it cannot be attained because the optimum depends on the very quantity we are trying to approximate! However, the Bayes factor $B^{\pi}_{12}(\mathscr{D}_n)$ can first be approximated on a crude basis and the corresponding construction of α^{O} iterated till the Bayes factor approximation (2.10) stabilizes.

We will now illustrate this derivation in the case of the normal model, with an application to **normaldata**. (We showed in Sect. 2.3.3 that the Bayes factor was available in closed form so this implementation of the bridge sampler is purely for illustrative purposes.) A further implementation is discussed in Chap. 4, Sect. 4.3.2, in connection with the probit model.

When assessing whether or $\mu = 0$ is appropriate for the single sample normal model, the above approximation does not apply directly because there is an extra parameter in the unconstrained model. There are however two easy tricks out of this difficulty. The first one, repeatedly found in the literature, is to add an arbitrary density to make dimensions match. In the normal example, this means introducing an arbitrary (normalized) density $\pi_1^*(\mu|\sigma^2)$ in the constrained model (denoted M_1) and extending the Bayes factor representation (2.9) to

$$B^{\pi}_{12}(\mathscr{D}_n) = \frac{\displaystyle\int \pi_1^*(\mu|\sigma^2)\tilde{\pi}_1(\sigma^2|\mathscr{D}_n)\alpha(\theta)\pi_2(\theta|\mathscr{D}_n)\mathrm{d}\theta}{\displaystyle\int \tilde{\pi}_2(\theta|\mathscr{D}_n)\alpha(\theta)\pi_1(\sigma^2|\mathscr{D}_n)\mathrm{d}\sigma^2\pi_1^*(\mu|\sigma^2)\mathrm{d}\mu}\,.$$

which holds independently of $\pi_1^*(\mu|\sigma^2)$ for the same reason as in (2.9). The choice of the substitute $\pi_1^*(\mu|\sigma^2)$ equal to an approximation of $\pi_2(\mu|\mathscr{D}_n, \sigma^2)$ is suggested by Chen et al. (2000). For instance, we can use as $\pi_1^*(\mu|\sigma^2)$ a normal distribution $\mathscr{N}(\hat{\mu}, \hat{\sigma}^2)$ where $\hat{\mu}$ and $\hat{\sigma}^2$ are computed based on a simulation from $\pi_2(\mu, \sigma|\mathscr{D}_n)$.

The exact value of this Bayes factor $B^{\pi}_{12}(\mathscr{D}_n)$ for Illingworth's data is given by

```
> ((1+ratio)/(1+ratio/(n+1)))^(-n/2)*sqrt(n+1)
[1] 6.821262
```

while the bridge sampling solution is obtained as

```
> n=64
> xbar=mean(shift)
> sqar=(n-1)*var(shift)
> Nmc=10^7
> # Simulation from model M2:
> sigma2=1/rgamma(Nmc,shape=n/2,rate=(n*xbar^2/(n+1)+sqar)/2)
> mu2=rnorm(Nmc,n*xbar/(n+1),sd=sqrt(sigma2/(n+1)))
> # Simulation from model M1:
> sigma1=1/rgamma(Nmc,shape=n/2,rate=(n*xbar^2+sqar)/2)
> muhat=mean(mu2)
> tauat=sd(mu2)
> mu1=rnorm(Nmc,mean=muhat,sd=tauat)
> #tilde functions
> tildepi1=function(sigma,mu){
+   exp(-.5*((n*xbar^2+sqar)/sigma+(n+2)*log(sigma))+
+   dnorm(mu,muhat,tauat,log=T))
+   }
> tildepi2=function(sigma,mu){
+   exp(-.5*((n*(xbar-mu)^2+sqar+mu^2)/sigma+(n+3)*log(sigma)+
+   log(2*pi)))}
> #Bayes Factor loop
> K=diff=1
> rationum=tildepi2(sigma1,mu1)/tildepi1(sigma1,mu1)
> ratioden=tildepi1(sigma2,mu2)/tildepi2(sigma2,mu2)
> while (diff>0.01*K){
+   BF=mean(1/(1+K*rationum))/mean(1/(K+ratioden))
+   diff=abs(K-BF)
+   K=BF}
```

and returns the value

```
> BF
[1] 6.820955
```

which is definitely close to the true value!

The second possible trick to overcome the dimension difficulty while using the bridge sampling strategy is to introduce artificial posterior distributions in each of the parameters spaces and to process each marginal likelihood as an integral ratio in itself. For instance, if $\eta_1(\theta_1)$ is an arbitrary normalized density on θ_1, and α is an arbitrary function, we have

$$m_1(\mathscr{D}_n) = \int \tilde{\pi}_1(\theta_1|\mathscr{D}_n)\,\mathrm{d}\theta_1 = \frac{\displaystyle\int \tilde{\pi}_1(\theta_1|\mathscr{D}_n)\alpha(\theta_1)\eta_1(\theta_1)\,\mathrm{d}\theta_1}{\displaystyle\int \eta_1(\theta_1)\alpha(\theta_1)\pi_1(\theta_1|\mathscr{D}_n)\,\mathrm{d}\theta_1}$$

by application of (2.9). Therefore, the optimal choice of α leads to the approximation

$$\hat{m}_1(\mathscr{D}_n) = \frac{\sum_{i=1}^N \tilde{\pi}_1(\theta_{1i}^\eta|\mathscr{D}_n)\big/\left\{m_1(\mathscr{D}_n)\tilde{\pi}_1(\theta_{1i}^\eta|\mathscr{D}_n) + \eta(\theta_{1i}^\eta)\right\}}{\sum_{i=1}^N \eta(\theta_{1i})\big/\left\{m_1(\mathscr{D}_n)\tilde{\pi}_1(\theta_{1i}|\mathscr{D}_n) + \eta(\theta_{1i})\right\}}$$

when $\theta_{1i} \sim \pi_1(\theta_1|\mathscr{D}_n)$ and $\theta_{1i}^\eta \sim \eta(\theta_1)$. The choice of the density η is obviously fundamental and it should be close to the true posterior $\pi_1(\theta_1|\mathscr{D}_n)$ to guarantee good convergence approximation. Using a normal approximation to the posterior distribution of θ or a non-parametric approximation based on a sample from $\pi_1(\theta_1|\mathscr{D}_n)$, or yet again an average of MCMC proposals (see Chap. 4) are reasonable choices.

The R implementation of this approach can be done as follows

```
> sigma1=1/rgamma(Nmc,shape=n/2,rate=(n*xbar^2+sqar)/2)
> sihat=mean(log(sigma1))
> tahat=sd(log(sigma1))
> sigma1b=exp(rnorm(Nmc,sihat,tahat))
> #tilde function
> tildepi1=function(sigma){
  exp(-.5*((n*xbar^2+sqar)/sigma+(n+2)*log(sigma)))}
> K=diff=1
> rnum=dnorm(log(sigma1b),sihat,tahat)/
+           (sigma1b*tildepi1(sigma1b))
> rden=sigma1*tildepi1(sigma1)/dnorm(log(sigma1),sihat,tahat)
> while (diff>0.01*K){
>   BF=mean(1/(1+K*rnum))/mean(1/(K+rden))
>   diff=abs(K-BF)
>   K=BF}
> m1=BF
```

when using a normal distribution on $\log(\sigma^2)$ as an approximation to $\pi_1(\theta_1|\mathscr{D}_n)$. When considering the unconstrained model, a bivariate normal density can be used, as in

```
> sigma2=1/rgamma(Nmc,shape=n/2,rate=(n*xbar^2/(n+1)+sqar)/2)
> mu2=rnorm(Nmc,n*xbar/(n+1),sd=sqrt(sigma2/(n+1)))
> temean=c(mean(mu2),mean(log(sigma2)))
> tevar=cov.wt(cbind(mu2,log(sigma2)))$cov
> te2b=rmnorm(Nmc,mean=temean,tevar)
> mu2b=te2b[,1]
> sigma2b=exp(te2b[,2])
```

leading to

```
> m1/m2
[1] 6.824417
```

The performances of both extensions are obviously highly dependent on the choice of the completion factors, η_1 and η_2 on the one hand and π_1^* on the other hand. The performances of the first solution, which bridges both models via π_1^*, are bound to deteriorate as the dimension gap between those models increases. The impact of the dimension of the models is less keenly felt for the other solution, as the approximation remains local.

As a simple illustration of the performances of both methods, we produce here a comparison between the completions based on a single pseudo-conditional and on two local approximations to the posteriors, by running repeated approximations for **normaldata** and tracing the resulting boxplot as a measure of the variability of those methods. As shown in Fig. 2.6, the variability is quite comparable for both solutions in this specific case.

Note that there exist many other approaches to the approximative computation of marginal likelihoods and of Bayes factors that we cannot cover here. We want however to point out the dangers of the harmonic mean approximation. This approach proceeds from the interesting identity

$$\begin{aligned}\mathbb{E}^{\pi_1}\left[\left.\frac{\varphi_1(\theta_1)}{\pi_1(\theta_1)\ell_1(\theta_1|\mathscr{D}_n)}\right|\mathscr{D}_n\right] &= \int \frac{\varphi_1(\theta_1)}{\pi_1(\theta_1)\ell_1(\theta_1|\mathscr{D}_n)}\,\frac{\pi_1(\theta_1)\ell_1(\theta_1|\mathscr{D}_n)}{m_1(\mathscr{D}_n)}\,\mathrm{d}\theta_1 \\ &= \frac{1}{m_1(\mathscr{D}_n)}\,,\end{aligned}$$

which holds, no matter what the density $\varphi_1(\theta_1)$ is—provided $\varphi_1(\theta_1)=0$ when $\pi_1(\theta_1)\ell_1(\theta_1|\mathscr{D}_n)=0$—. The most common implementation in approximations of the marginal likelihood uses $\varphi_1(\theta_1)=\pi_1(\theta_1)$, leading to the approximation

$$\hat{m}_1(\mathscr{D}_n) = 1\Big/ N^{-1}\sum_{j=1}^{N}\frac{1}{\ell_1(\theta_{1j}|\mathscr{D}_n)}\,.$$

While very tempting, since it allows for a direct processing of simulations from the posterior distribution, this approximation is unfortunately most often associated with an infinite variance (Exercise 2.19) and, thus, *should not be used*. On the opposite, using φ_1's with supports constrained to the 25 % HPD

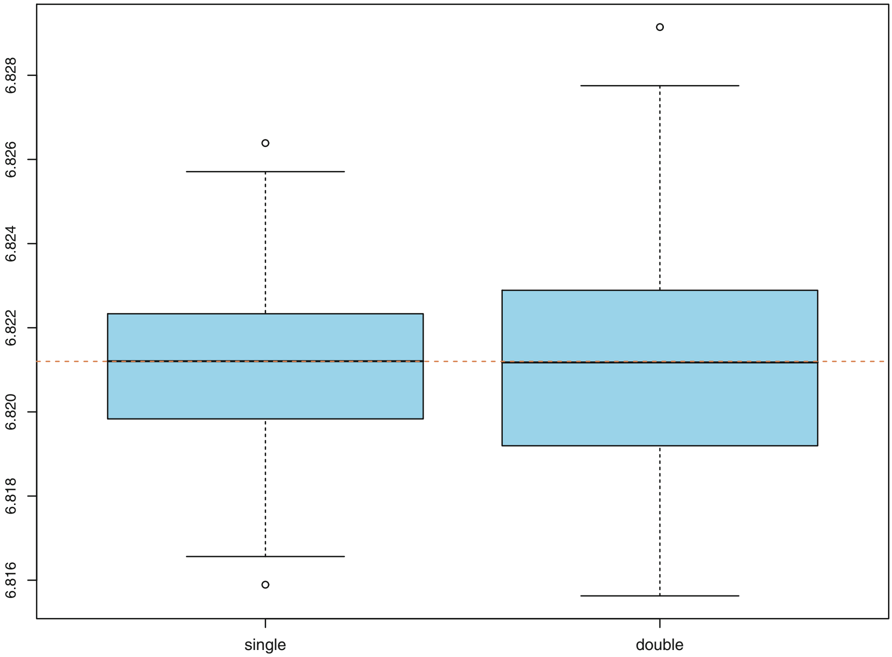

Fig. 2.6. Dataset **normaldata**: Boxplot of the variability of the approximations to the Bayes factor assessing whether or not $\mu = 0$, based on a single and on a double completions. Each approximation is based on 10^5 simulations and the boxplots are based on 250 approximations. The *dotted line* corresponds to the true value of $B_{12}^{\pi}(\mathscr{D}_n)$

regions—approximated by the convex hull of the 10 % or of the 25 % highest simulations—is both completely appropriate and implementable (Marin and Robert, 2010).

2.5 Outlier Detection

The above description of inference in normal models is only an introduction both to Bayesian inference and to normal structures. Needless to say, there exists a much wider range of possible applications. For instance, we will meet the normal model again in Chap. 4 as the original case of the (generalized) linear model. Before that, we conclude this chapter with a simple extension of interest, the detection of outliers.

Since normal modeling is often an approximation to the “real thing,” there may be doubts about its adequacy. As already mentioned above, we will deal later with the problem of checking that the normal distribution is appropriate for the whole dataset. Here, we consider the somehow simpler problem of separately assessing whether or not each point in the dataset is compatible with normality. There are many different ways of dealing with this problem. We choose here to use the *predictive distribution*: If an observation x_i is unlikely under the predictive distribution based on the *other observations*, then we can argue against its distribution being equal to the distribution of the other observations.

If x_{n+1} is a future observation from the same distribution $f(\cdot|\theta)$ as the sample $\mathscr{D}_n$, its *predictive distribution* given the current sample is defined as

$$f^\pi(x_{n+1}|\mathscr{D}_n) = \int f(x_{n+1}|\theta,\mathscr{D}_n)\pi(\theta|\mathscr{D}_n)\,\mathrm{d}\theta = \int f(x_{n+1}|\theta)\pi(\theta|\mathscr{D}_n)\,\mathrm{d}\theta\,.$$

This definition is coherent with the Bayesian approach, which considers x_{n+1} as an extra unknown and then integrates out θ if x_{n+1} is the "parameter" of interest.

For the normal $\mathscr{N}(\mu,\sigma^2)$ setup, using a conjugate prior on (μ,σ^2) of the form

$$(\sigma^2)^{-\lambda_\sigma-3/2}\exp-\left\{\lambda_\mu(\mu-\xi)^2+\alpha\right\}/2\sigma^2\,,$$

the corresponding posterior distribution on (μ,σ^2) given $\mathscr{D}_n$ is

$$\mathscr{N}\left(\frac{\lambda_\mu\xi+n\overline{x}_n}{\lambda_\mu+n},\frac{\sigma^2}{\lambda_\mu+n}\right)\times\mathscr{IG}\left(\lambda_\sigma+n/2,\left[\alpha+s^2+\frac{n\lambda_\mu}{\lambda_\mu+n}(\overline{x}-\xi)^2\right]/2\right)\,,$$

denoted by

$$\mathscr{N}\left(\xi(\mathscr{D}_n),\sigma^2/\lambda_\mu(\mathscr{D}_n)\right)\times\mathscr{IG}\left(\lambda_\sigma(\mathscr{D}_n)/2,\alpha(\mathscr{D}_n)/2\right)\,,$$

and the predictive on x_{n+1} is derived as

$$\begin{aligned}
f^\pi(x_{n+1}|\mathscr{D}_n) \propto & \int (\sigma^2)^{-\lambda_\sigma(\mathscr{D}_n)/2-1-1}\exp-(x_{n+1}-\mu)^2/2\sigma^2\\
&\qquad\times\exp-\left\{\lambda_\mu(\mathscr{D}_n)(\mu-\xi(\mathscr{D}_n))^2+\alpha(\mathscr{D}_n)\right\}/2\sigma^2\,\mathrm{d}(\mu,\sigma^2)\\
\propto & \int (\sigma^2)^{-\lambda_\sigma(\mathscr{D}_n)/2-3/2}\exp-\big\{(\lambda_\mu(\mathscr{D}_n)+1)(x_{n+1}-\xi(\mathscr{D}_n))^2\\
& \quad /\lambda_\mu(\mathscr{D}_n)+\alpha(\mathscr{D}_n)\big\}/2\sigma^2\,\mathrm{d}\sigma^2\\
\propto & \left[\alpha(\mathscr{D}_n)+\frac{\lambda_\mu(\mathscr{D}_n)+1}{\lambda_\mu(\mathscr{D}_n)}(x_{n+1}-\xi(\mathscr{D}_n))^2\right]^{-(\lambda_\sigma(\mathscr{D}_n)+1)/2}\,.
\end{aligned}$$

Therefore, the predictive of x_{n+1} given the sample $\mathscr{D}_n$ is a Student t distribution with mean $\xi(\mathscr{D}_n)$ and $\lambda_\sigma(\mathscr{D}_n)$ degrees of freedom. In the special case of the noninformative prior, corresponding to the limiting values $\lambda_\mu=\lambda_\sigma=\alpha=0$, the predictive is

$$f^\pi(x_{n+1}|\mathscr{D}_n)\propto\left[s^2+\frac{n+1}{n}1(x_{n+1}-\overline{x}_n)^2\right]^{-(n+1)/2}\,. \qquad (2.11)$$

It is therefore a Student's t distribution with n degrees of freedom, a mean equal to $\overline{x}_n$ and a scale factor equal to $(n-1)s^2/n$, which is equivalent to a variance equal to $(n-1)s^2/n^2$ (to compare with the maximum likelihood estimator $\hat{\sigma}_n^2 = s^2/n$).

In the outlier problem, we process each observation $x_i \in \mathscr{D}_n$ as if it was a "future" observation. Namely, we consider $f_i^\pi(x|\mathscr{D}_n^i)$ as being the predictive distribution based on $\mathscr{D}_n^i = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. Considering $f_i^\pi(x_i|\mathscr{D}_n^i)$ or the corresponding cdf $F_i^\pi(x_i|\mathscr{D}_n^i)$ (in dimension one) gives an indication of the level of compatibility of the observation x_i with the sample. To quantify this level, we can, for instance, approximate the distribution of $F_i^\pi(x_i|\mathscr{D}_n^i)$ as being uniform over $[0,1]$ since $F_i^\pi(\cdot|\mathscr{D}_n^i)$ does converge to the true cdf of the model. Simultaneously checking all $F_i^\pi(x_i|\mathscr{D}_n^i)$ over i may signal outliers.

↯ The detection of outliers must pay attention to the *Bonferroni fallacy*, which is that extreme values do occur in large enough samples. This means that, as n increases, we will see smaller and smaller values of $F_i^\pi(x_i|\mathscr{D}_n^i)$ occurring, and this even when the whole sample is from the same distribution. The significance level must therefore be chosen in accordance with this observation, for instance using a bound a on $F_i^\pi(x_i|\mathscr{D}_n^i)$ such that

$$1-(1-a)^n = 1-\alpha\,,$$

where α is the nominal level chosen for outlier detection.

Considering **normaldata**, we can compute the predictive cdf for each of the 64 observations, considering the 63 remaining ones as data.

```
> n=length(shift)
> outl=rep(0,n)
> for (i in 1:n){
+    lomean=-mean(shift[-i])
+    losd=sd(shift[-i])*sqrt((n-2)/n)
+    outl[i]=pt((shift[i]-lomean)/losd,df=n-1)
+    }
```

Figure 2.7 provides the qq-plot of the $F_i^\pi(x_i|\mathscr{D}_n^i)$'s against the uniform quantiles and compares it with a qq-plot based on a dataset truly simulated from the uniform $\mathscr{U}(0,1)$.

```
> plot(c(0,1),c(0,1),lwd=2,ylab="Predictive",xlab="Uniform",
+ type="l")
> points((1:n)/(n+1),sort(outl),pch=19,col="steelblue3")
> points((1:n)/(n+1),sort(runif(n)),pch=19,col="tomato")
```

There is no clear departure from uniformity when looking at this graph, except of course for the multiple values found in **normaldata**.

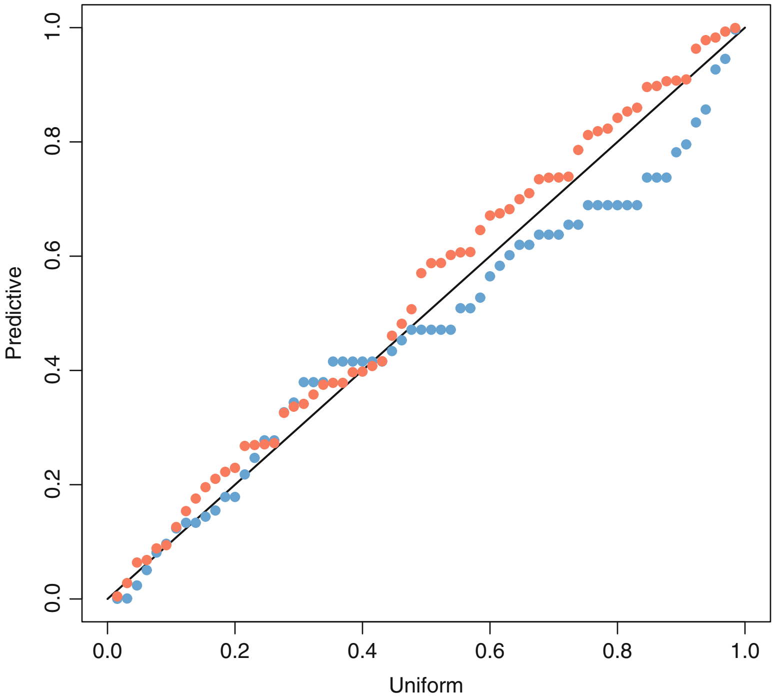

Fig. 2.7. Dataset **normaldata**: qq-plot of the sample of the $F_i^\pi(x_i|\mathscr{D}_n^i)$ for a uniform $\mathscr{U}(0,1)$ distribution (*blue dots*) and comparison with a qq-plot for a uniform $\mathscr{U}(0,1)$ sample (*red dots*)

2.6 Exercises

2.1 Show that, if

$$\mu|\sigma^2 \sim \mathscr{N}(\xi, \sigma^2/\lambda_\mu)\,, \qquad \sigma^2 \sim \mathscr{IG}(\lambda_\sigma/2, \alpha/2)\,,$$

then

$$\mu \sim \mathscr{T}(\lambda_\sigma, \xi, \alpha/\lambda_\mu\lambda_\sigma)$$

a t distribution with λ_σ degrees of freedom, location parameter ξ and scale parameter $\alpha/\lambda_\mu\lambda_\sigma$.

2.2 Show that, if $\sigma^2 \sim \mathscr{IG}(\alpha,\beta)$, then $\mathbb{E}[\sigma^2] = \beta/(\alpha-1)$. Derive from the density of $\mathscr{IG}(\alpha,\beta)$ that the mode is located in $\beta/(\alpha+1)$.

2.3 Show that minimizing (in $\hat\theta(\mathscr{D}_n)$) the posterior expectation $\mathbb{E}^\pi[||\theta-\hat\theta||^2|\mathscr{D}_n]$ produces the posterior expectation as the solution in $\hat\theta$.

2.4 Show that the Fisher information on $\theta = (\mu,\sigma^2)$ for the normal $\mathscr{N}(\mu,\sigma^2)$ distribution is given by

$$I^F(\theta) = \mathbb{E}_\theta\left[\begin{pmatrix} 1/\sigma^2 & 2(x-\mu)/2\sigma^4 \\ 2(x-\mu)/2\sigma^4 & (\mu-x)^2/\sigma^6 - 1/2\sigma^4 \end{pmatrix}\right] = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/2\sigma^4 \end{pmatrix}$$

and deduce that Jeffreys' prior is $\pi^J(\theta) \propto 1/\sigma^3$.

2.5 Derive each line of Table 2.1 by an application of Bayes' formula, $\pi(\theta|x) \propto \pi(\theta)f(x|\theta)$, and the identification of the standard distributions.

2.6 A Weibull distribution $\mathscr{W}(\alpha,\beta,\gamma)$ is defined as the power transform of a gamma $\mathscr{G}(\alpha,\beta)$ distribution: If $x \sim \mathscr{W}(\alpha,\beta,\gamma)$, then $x^\gamma \sim \mathscr{G}(\alpha,\beta)$. Show that, when γ is known, $\mathscr{W}(\alpha,\beta,\gamma)$ allows for a conjugate family, but that it does not an exponential family when γ is unknown.

2.7 Show that, when the prior on $\theta = (\mu,\sigma^2)$ is $\mathscr{N}(\xi,\sigma^2/\lambda_\mu) \times \mathscr{IG}(\lambda_\sigma,\alpha)$, the marginal prior on μ is a Student t distribution $\mathcal{T}(2\lambda_\sigma,\xi,\alpha/\lambda_\mu\lambda_\sigma)$ (see Example 2.18 for the definition of a Student t density). Give the corresponding marginal prior on σ^2. For an iid sample $\mathscr{D}_n = (x_1,\ldots,x_n)$ from $\mathscr{N}(\mu,\sigma^2)$, derive the parameters of the posterior distribution of (μ,σ^2).

2.8 Show that the normalizing constant for a Student $\mathscr{T}(\nu,\mu,\sigma^2)$ distribution is

$$\frac{\Gamma((\nu+1)/2)/\Gamma(\nu/2)}{\sigma\sqrt{\nu\pi}}\,.$$

Deduce that the density of the Student t distribution $\mathscr{T}(\nu,\theta,\sigma^2)$ is

$$f_\nu(x) = \frac{\Gamma((\nu+1)/2)}{\sigma\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1+\frac{(x-\theta)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}\,.$$

2.9 Show that, for location and scale models, the specific noninformative priors are special cases of Jeffreys' generic prior, i.e., that $\pi^J(\theta)=1$ and $\pi^J(\theta)=1/\theta$, respectively.

2.10 Show that, when $\pi(\theta)$ is a probability density, (2.5) necessarily holds for all datasets $\mathscr{D}_n$.

2.11 Consider a dataset $\mathscr{D}_n$ from the Cauchy distribution, $\mathscr{C}(\mu,1)$.

1. Show that the likelihood function is

$$\ell(\mu|\mathscr{D}_n) = \prod_{i=1}^n f_\mu(x_i) = \frac{1}{\pi^n\prod_{i=1}^n(1+(x_i-\mu)^2)}\,.$$

2. Examine whether or not there is a conjugate prior for this problem. (The answer is *no*.)
3. Introducing a normal prior on μ, say $\mathscr{N}(0,10)$, show that the posterior distribution is proportional to

$$\tilde{\pi}(\mu|\mathscr{D}_n) = \frac{\exp(-\mu^2/20)}{\prod_{i=1}^n(1+(x_i-\mu)^2)}\,.$$

4. Propose a numerical solution for solving $\tilde{\pi}(\mu|\mathscr{D}_n)=k$. (*Hint:* A simple trapezoidal integration can be used: based on a discretization size Δ, computing $\tilde{\pi}(\mu|\mathscr{D}_n)$ on a regular grid of width Δ and summing up.)

2.12 Show that the limit of the posterior probability $\mathbb{P}^\pi(\mu < 0|x)$ of (2.7) when τ goes to ∞ is $\Phi(-x/\sigma)$. Show that, when ξ varies in $\mathbb{R}$, the posterior probability can take any value between 0 and 1.

2.13 Define a function `BaRaJ` of the ratio `rat` when `z=mean(shift)/.75` in the function `BaFa`. Deduce from a plot of the function `BaRaJ` that the Bayes factor is always less than one when `rat` varies. (*Note:* It is possible to establish analytically that the Bayes factor is maximal and equal to 1 for $\tau = 0$.)

2.14 In the application part of Example 2.1 to **normaldata**, plot the approximated Bayes factor as a function of τ. (*Hint:* Simulate a single normal $\mathscr{N}(0,1)$ sample and recycle it for all values of τ.)

2.15 In the setup of Example 2.1, show that, when $\xi \sim \mathscr{N}(0,\sigma^2)$, the Bayes factor can be expressed in closed form using the normalizing constant of the t distribution (see Exercise 2.8)

2.16 Discuss what happens to the importance sampling approximation when the support of g is larger than the support of γ.

2.17 Show that, when γ is the normal $\mathscr{N}(0,\nu/(\nu-2))$ density and f_ν is the density of the t distribution with ν degrees of freedom, the ratio

$$\frac{f_\nu^2(x)}{\gamma(x)} \propto \frac{e^{x^2(\nu-2)/2\nu}}{[1+x^2/\nu]^{(\nu+1)}}$$

does not have a finite integral. What does this imply about the variance of the importance weights?

Deduce that the importance weights of Example 2.3 have infinite variance.

2.18 If f_ν denotes the density of the Student t distribution $\mathscr{T}(\nu,0,1)$ (see Exercise 2.8), consider the integral

$$\mathfrak{I} = \int \sqrt{\left|\frac{x}{1-x}\right|}\, f_\nu(x)\,\mathrm{d}x\,.$$

1. Show that $\mathfrak{I}$ is finite but that

$$\int \frac{|x|}{|1-x|} f_\nu(x)\,\mathrm{d}x = \infty\,.$$

2. Discuss the respective merits of the following importance functions γ
   - the density of the Student $\mathscr{T}(\nu,0,1)$ distribution,
   - the density of the Cauchy $\mathscr{C}(0,1)$ distribution,
   - the density of the normal $\mathscr{N}(0,\nu/(\nu-2))$ distribution.

   In particular, show via an R simulation experiment that these different choices all lead to unreliable estimates of $\mathfrak{I}$ and deduce that the three corresponding estimators have infinite variance.
3. Discuss the alternative choice of a gamma distribution folded at 1, that is, the distribution of x symmetric around 1 and such that

$$|x-1| \sim \mathscr{G}a(\alpha,1)\,.$$

Show that

$$h(x)\frac{f^2(x)}{\gamma(x)} \propto \sqrt{x}\, f_\nu^2(x)\, |1-x|^{1-\alpha-1} \exp|1-x|$$

is integrable around $x = 1$ when $\alpha < 1$ but not at infinity. Run a simulation experiment to evaluate the performances of this new proposal.

2.19 Evaluate the harmonic mean approximation

$$\hat{m}_1(\mathscr{D}_n) = 1 \bigg/ N^{-1} \sum_{j=1}^{N} \frac{1}{\ell_1(\theta_{1j}|\mathscr{D}_n)}\,.$$

when applied to the $\mathscr{N}(0,\sigma^2)$ model, **normaldata**, and an $\mathscr{IG}(1,1)$ prior on σ^2.