Bayesian Methods in Epidemiology

31

Leonhard Held

Contents

L. Held
Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Zurich, Switzerland

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_57,

31.1 Introduction

This chapter gives a self-contained introduction to the Bayesian approach to statistical inference. Standard epidemiological problems such as diagnostic tests, the analysis of prevalence, case-control, and cohort data will serve as examples. More advanced topics, such as empirical Bayes methods and Markov chain Monte Carlo techniques, are also covered.

The Bayesian approach is easy to understand, if the reader is able to follow the actual calculations. Only some basic knowledge of the rules of probability and calculus is needed. A reader, willing to dive into these fairly simple technicalities, will be able to fully appreciate the beauty and simplicity of the Bayesian approach. An appendix summarizes the required technical background.

Modern Bayesian statistics is often performed using so-called Monte Carlo methods based on random numbers simulated on a computer. Many quantities of interest can be computed very easily using Monte Carlo. From time to time, I will show very short programming code in `R` to illustrate the simplicity of Monte Carlo methods. Understanding of the code is, however, not necessary for an understanding of this chapter.

To understand Bayesian methods, in particular Bayes' theorem, the most important concept is that of a *conditional probability*. In Sect. 31.2, we will illustrate the notion of conditional probabilities and Bayesian updating in the context of diagnostic testing. Further details on conditional probabilities are listed in Appendix A. The Bayesian approach to parameter estimation is discussed in Sect. 31.3. Appendix B summarizes important properties of the distributions used in this section and their implementation in `R`. After a brief introduction to Bayesian prediction, Sect. 31.4 discusses techniques for prior criticism and Bayesian model selection. Empirical Bayes methods and Markov chain Monte Carlo techniques are described in Sect. 31.5. We close with some discussion in Sect. 31.6.

31.2 Conditional Probabilities and Diagnostic Testing

The use of Bayes' theorem is routine in the context of diagnostic testing.

Example 31.1. Diagnostic testing
Suppose a simple diagnostic test for a specific disease, which produces either a positive or a negative test result, is known to have 90% sensitivity. This means that the probability of a positive test result, if the person being tested has the disease, is 90%. This is a *conditional* probability since we know that the person has the disease and we write $\Pr(T+ \mid D+) = 0.9$, the probability (Pr) of a positive test result $(T+)$ *given* disease $(D+)$ is 0.9. Now, assume that the test also has 90% specificity and write $D-$ if the person being tested is free of the disease. Similarly, let $T-$ denote a negative test result. A 90% specificity simply means that $\Pr(T- \mid D-) = 0.9$.

Conditional probabilities behave just like ordinary probabilities if we always condition on the same event, for example, on $D+$, say. In particular, they must be numbers between 0 and 1 and $\Pr(T- \mid D+)$ must be equal to $1 - \Pr(T+ \mid D+)$, that is, the conditional probability of a negative test result is 1 minus the conditional probability of a positive test result, if both probabilities are conditional on $D+$. The same of course holds if we condition on $D-$ rather than on $D+$.

However, the real power of conditional probabilities emerges if we condition on different events and relate conditional to ordinary (unconditional) probabilities. The most important formula to do this is *Bayes' theorem* (see Appendix A for a derivation). In the diagnostic testing context, we can use Bayes' theorem to compute the conditional probability of disease given a positive test result:

$$\Pr(D+\mid T+) = \frac{\Pr(T+\mid D+)\Pr(D+)}{\Pr(T+)}. \tag{31.1}$$

The prevalence $\Pr(D+)$ is an example of an ordinary (unconditional) probability.

The denominator $\Pr(T+)$ in (31.1), the (unconditional) probability of a positive test result, is unknown, but we know from above that $\Pr(D+\mid T+)$ and

$$\Pr(D-\mid T+) = \frac{\Pr(T+\mid D-)\Pr(D-)}{\Pr(T+)} \tag{31.2}$$

must add to unity, from which we can easily deduce that

$$\Pr(T+) = \Pr(T+\mid D+)\Pr(D+) + \Pr(T+\mid D-)\Pr(D-). \tag{31.3}$$

This equation is sometimes called the *law of total probability*. Thus, we can calculate $\Pr(T+)$ if we know the sensitivity $\Pr(T+\mid D+)$, the prevalence $\Pr(D+)$, and $\Pr(T+\mid D-)$, which is 1 minus the specificity $\Pr(T-\mid D-)$.

Equation 31.1 exemplifies the process of *Bayesian updating*: We update the prior risk $\Pr(D+)$ in the light of a positive test result $T+$ to obtain the posterior risk $\Pr(D+\mid T+)$, the conditional probability of disease given a positive test result, also known as the *positive predictive value*.

Example 31.1. (Continued)
In the following we assume that the prevalence $\Pr(D+) = 1\%$ for the disease considered above. Then

$$\Pr(T+) = 0.9 \cdot 0.01 + 0.1 \cdot 0.99 = 0.108$$

and hence

$$\Pr(D+\mid T+) = \frac{0.9 \cdot 0.01}{0.108} \approx 0.083,$$

i.e. the disease risk increases from 1% to 8.3% after a positive test result. It is up to the reader to write down the corresponding formula to (31.1) to compute the *negative predictive value* $\Pr(D-\mid T-)$, which turns out to be approximately 0.999. Thus, the disease risk decreases from 1% to $\Pr(D+\mid T-)$=$1 - \Pr(D-\mid T-) = 100\% - 99.9\% = 0.1\%$ if the test was negative. The disease risk changes in the expected direction depending on the test result.

Equation 31.1, with the denominator $\Pr(T+)$ replaced by (31.3), is often referred to as Bayes' theorem in probability textbooks. However, the resulting formula is somewhat complex and not particularly intuitive. A simpler version of

Bayes' theorem can be obtained if we switch from probabilities to *odds*. In general, whenever we take a probability, and divide it by 1 minus that probability, the resulting ratio is referred to as the corresponding odds. Of course, every probability refers to a particular event happening and 1 minus that probability is the probability that the event is *not* happening. Odds are hence nothing more than a ratio of two probabilities: the probability of an event happening divided by the probability that the event is not happening. For example, a probability of 10% corresponds to odds of 1/9, often described as 1 to 9. Conversely, 3 to 1 odds, say, correspond to a probability of $3/4 = 0.75$.[1]

We can now derive a simple version of Bayes' theorem in terms of odds, if we divide (31.1)–(31.2):

$$\frac{\Pr(D+\mid T+)}{\Pr(D-\mid T+)} = \frac{\Pr(T+\mid D+)}{\Pr(T+\mid D-)} \times \frac{\Pr(D+)}{\Pr(D-)}. \tag{31.4}$$

We will refer to this equation as Bayes' theorem in *odds form*. Here $\Pr(D+)/\Pr(D-)$ are the *prior odds*, $\Pr(D+\mid T+)/\Pr(D-\mid T+)$ are the *posterior odds* and $\Pr(T+\mid D+)/\Pr(T+\mid D-)$ is the so-called *likelihood ratio* for a positive test result, which we can easily identify as the sensitivity divided by 1 minus the specificity. Bayesian updating is thus just one simple mathematical operation: Multiply the prior odds with the likelihood ratio to obtain the posterior odds.

Example 31.1. (Continued)
The prior odds are 1/99 and the likelihood ratio (for a positive test result) is $0.9/0.1 = 9$. The posterior odds are therefore $9 \cdot 1/99 = 1/11 \approx 9.1\%$. So the prior odds of 1 to 99 change to posterior odds of 1 to 11 in the light of a positive test result. If the test result was negative, then the prior odds need to be multiplied with the likelihood ratio for a negative test result, which is $\Pr(T-\mid D+)/\Pr(T-\mid D-) = 0.1/0.9 = 1/9$. (Note that the likelihood ratio for a negative test result is 1 minus the sensitivity divided by the specificity.) This leads to posterior odds of $1/9 \cdot 1/99 = 1/891$. We leave it up to the reader to check that these posterior odds correspond to the positive and the negative predictive value, respectively, calculated earlier. Figure 31.1 illustrates Bayesian learning using odds in this example.

We now discuss an important formal aspect. Formula (31.1) is specified for a positive test result $T+$ and a diseased person $D+$ but is equally valid if we replace a positive test result $T+$ by a negative one, that is, $T-$, or $D+$ by $D-$, or both. In fact, we have already replaced $D+$ by $D-$ to write down (31.2).

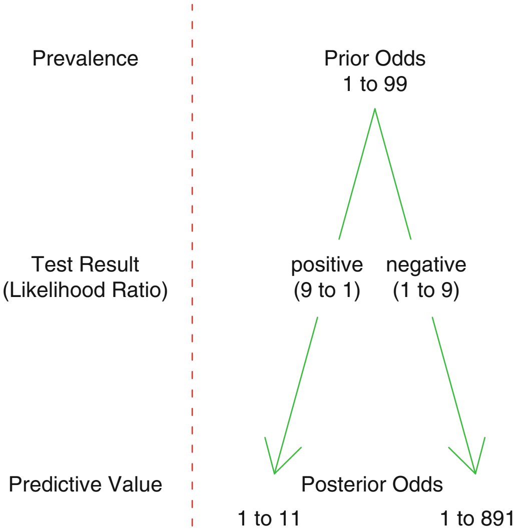
A more general description of Bayes' theorem is given by

$$\mathrm{p}(D = d \mid T = t) = \frac{\mathrm{p}(T = t \mid D = d) \times \mathrm{p}(D = d)}{\mathrm{p}(T = t)}, \tag{31.5}$$

where D and T are binary *random variables* which take the values d and t, respectively. In the diagnostic setting outlined above, d and t can be either "+"

[1] Odds $\omega = \pi/(1-\pi)$ can be back-transformed to probabilities π using $\pi = \omega/(1+\omega)$.

Fig. 31.1 Schematic representation of Bayesian learning in a diagnostic test example



or "–". Note that we have switched notation from Pr(.) to p(.) to emphasize that (31.5) relates to general *probability functions* of the random variables D and T, and not only to probabilities of the events $D+$ and $T+$, say.

An even more compact version of (31.5) is

$$p(d \mid t) = \frac{p(t \mid d) \times p(d)}{p(t)}. \tag{31.6}$$

Note that this equation also holds if the random variables D or T can take more than two possible values. The formula is also correct if it involves continuous random variables, in which case $p(\cdot)$ denotes a density function.[2]

In reality, information on prevalence is typically estimated from a prevalence study while sensitivity and specificity are derived from a diagnostic study. However, the uncertainty associated with these estimates has been ignored in the above calculations. In the following, we will describe the Bayesian approach to quantify the uncertainty associated with these estimates. This can subsequently be used to assess the uncertainty of the positive and negative predictive values.

31.3 Bayesian Parameter Estimation

Conceptually, the Bayesian approach to parameter estimation treats all unknown quantities as random variables with appropriate prior distributions. Some knowledge of important elementary probability distributions is therefore required. Appendix B summarizes properties of the distributions used in this chapter.

[2] Probability statements for continuous random variables X can be obtained through integration of the density function, for example, $\Pr(a \leq X \leq b) = \int_a^b p(x)dx$.

A *prior distribution* $\mathrm{p}(\theta)$ represents our knowledge about an unknown parameter θ, which we would like to update in the light of observed data x, whose probability of occurrence depends on θ. For example, x might be the results from an epidemiological study. The conditional probability function $\mathrm{p}(x \mid \theta)$ of x given θ is called the *likelihood function*. Combined with the prior distribution $\mathrm{p}(\theta)$, we can calculate the *posterior distribution* $\mathrm{p}(\theta \mid x)$ using Bayes' theorem:

$$\mathrm{p}(\theta \mid x) = \frac{\mathrm{p}(x \mid \theta) \times \mathrm{p}(\theta)}{\mathrm{p}(x)}. \tag{31.7}$$

This formula is of course just Eq. 31.6 with d replaced by θ and t replaced by x.

Note that the denominator $\mathrm{p}(x)$ does not depend on θ, its particular value is therefore not of primary interest. To compute the posterior distribution $\mathrm{p}(\theta \mid x)$ (up to a multiplicative constant), a simplified version of Bayes' theorem

$$\mathrm{p}(\theta \mid x) \propto \mathrm{p}(x \mid \theta) \times \mathrm{p}(\theta)$$

is sufficient.[3] In words, this formula says that the posterior distribution is proportional to the product of the likelihood function $\mathrm{p}(x \mid \theta)$ and the prior distribution $\mathrm{p}(\theta)$. Note that $\mathrm{p}(x \mid \theta)$, originally the probability or density function of the data x given the (unknown) parameter θ, is used as a function of θ for fixed x. It is convenient to write $\mathrm{L}_x(\theta)$ for $\mathrm{p}(x \mid \theta)$ to emphasize this fact:

$$\mathrm{p}(\theta \mid x) \propto \mathrm{L}_x(\theta) \times \mathrm{p}(\theta). \tag{31.8}$$

Note also that we need to know $\mathrm{L}_x(\theta)$ and $\mathrm{p}(\theta)$ only "up to scale," that is, we can ignore any multiplicative factors which do not depend on θ. This will often make the computations simpler.

A likelihood approach to statistical inference, see, for example, Pawitan (2001), uses only the likelihood $\mathrm{L}_x(\theta)$ and calculates the *Maximum Likelihood estimate* (MLE) $\hat{\theta}_{ML}$ defined as that particular value of θ which maximizes $\mathrm{L}_x(\theta)$. The likelihood function can also be used to compute frequentist *confidence intervals* based on the likelihood ratio test statistic. Alternatively, a Wald confidence interval can be calculated based on the *standard error* $\mathrm{se}(\hat{\theta}_{ML})$ of $\hat{\theta}_{ML}$, an estimate of the standard deviation of the MLE in (fictive) repeated experiments under identical conditions. The standard error can be calculated based on the curvature of the logarithm of the likelihood function (the so-called *log likelihood*) at the MLE.

In contrast, Bayes' theorem formalizes the fundamental principle of Bayesian inference in that the prior assumptions $\mathrm{p}(\theta)$ about θ are updated using the likelihood $\mathrm{p}(x \mid \theta)$ to obtain the posterior distribution $\mathrm{p}(\theta \mid x)$. The posterior distribution provides all information about the quantity of interest, but usually, we want to summarize it using point and interval estimates. The *posterior mean*, the mean of the posterior distribution, is the most commonly used point estimate, alternatively the

[3] The mathematical symbol $\propto$ stands for "is proportional to."

posterior median and the *posterior mode* may also be used. For interval estimation, any interval $[\theta_l, \theta_u]$ with $\Pr(\theta_l \leq \theta \leq \theta_u \,|\, x) = 1 - \alpha$ serves as a $(1 - \alpha) \cdot 100\%$ *credible interval*. A common choice is $\alpha = 5\%$, in which case we obtain 95% credible intervals. Often, *equi-tailed* credible intervals are used, where the same amount $(\alpha/2)$ of probability mass is cut-off from the left and right tail of the posterior distribution, that is,

$$\Pr(\theta < \theta_l \,|\, x) = \Pr(\theta > \theta_u \,|\, x) = \alpha/2.$$

So-called *highest posterior density* (HPD) intervals are also commonplace. They have the defining feature that the posterior density at any value of θ inside the credible interval must be larger than anywhere outside the credible interval. It can be shown that HPD credible intervals have the smallest width among all $(1-\alpha)$ credible intervals. If the posterior distribution is symmetric, for example, normal, then posterior mean, mode, and median coincide and HPD credible intervals are also equi-tailed. For non-symmetric posterior distributions, these quantities typically differ.

31.3.1 Choice of Prior Distribution

Compared with a classical approach to inference, the prior distribution has to be chosen appropriately, which often causes concerns to practitioners. In particular, a Bayesian analysis is often feared to introduce more unrealistic assumptions than a standard frequentist analysis. However, this is viewed as a misconception by many authors who consider the possibility to specify a prior distribution as something useful (Spiegelhalter et al. 2004; Greenland 2006, 2007).

The prior distribution should reflect the knowledge about the parameter of interest (e.g., a relative risk parameter in an epidemiological study). Ideally, this prior distribution should be elicited from experts (Spiegelhalter et al. 2004; O'Hagan et al. 2006). In the absence of expert opinions, simple *informative* prior distributions (e.g., that the relative risk parameter is with prior probability 95% between 0.5 and 2) may still be a reasonable choice. A sensitivity analysis with different prior distributions will help to examine how the conclusions depend on the choice of prior.

However, there have been various attempts to specify non-informative or reference priors to lessen the influence of the prior distribution. Reference priors used in such an "objective Bayes" approach typically correspond to rather unrealistic prior beliefs. However, "non-informative" priors provide a reference posterior where the impact of the prior distribution on the posterior distribution is minimized. Quite interestingly, such reference analyses may have equally good or even better frequentist properties than truly frequentist procedures (Bayarri and Berger 2004).

The most commonly used reference prior is *Jeffreys' prior*, named after the British physicist Harold Jeffreys (1891–1989). He proposed a general device to derive a non-informative prior distribution for a given likelihood function. It is interesting that the resulting non-informative reference prior is not necessarily a uniform prior. In many cases it is *improper*, that is, it does not sum or integrate to unity. For example, if the parameter θ can be any non-negative integer (without an upper limit) and we assume the same prior probability for each possible value of θ, then, this constitutes an improper prior distribution. In contrast, a *proper* prior will be a proper distribution in the mathematical sense. A proper prior can be easily achieved in this example by fixing an upper limit, that is, setting the prior probability of all values above that upper limit to zero. Operationally, improper priors are not a problem for parameter estimation, but they do cause problems in model selection, as we will see in Sect. 31.4. We will see some examples of Jeffreys' prior in the following.

31.3.2 Bayesian Analysis of Prevalence Data

The prevalence π is defined as the proportion of people in a population that has a specific disease. A simple prevalence study selects a random sample of n individuals from that population and counts the number x of diseased individuals. If the number of people in the population is large, then a binomial model $X \mid \pi \sim Bin(n, \pi)$[4] is appropriate to describe the statistical variability occurring in such a study design, see Appendix B for properties of the binomial distribution. Note that the MLE of π is $\hat{\pi}_{ML} = x/n$.

It is commonplace to select a beta distribution as prior distribution for π, because the beta distribution can only take values within the unit interval, that is, within the range of possible values of π. So assume that $\pi \sim Be(\alpha, \beta)$ a priori with $\alpha, \beta > 0$. Properties of the beta distribution are listed in Appendix B. Multiplying the binomial likelihood

$$\mathrm{L}_x(\pi) \propto \pi^x (1-\pi)^{n-x}$$

with the beta prior density

$$\mathrm{p}(\pi) \propto \pi^{\alpha-1}(1-\pi)^{\beta-1},$$

one easily obtains the posterior density

$$\begin{aligned}\mathrm{p}(\pi \mid x) &\propto \mathrm{L}_x(\pi) \times \mathrm{p}(\pi)\\ &\propto \pi^{\alpha+x-1}(1-\pi)^{\beta+n-x-1},\end{aligned}$$

compare Eq. 31.8. This can easily be identified as yet another beta distribution with parameters $\alpha + x$ and $\beta + n - x$:

$$\pi \mid x \sim Be(\alpha + x, \beta + n - x). \tag{31.9}$$

[4] The mathematical symbol $\sim$ stands for "is distributed as."

Compared with the prior distribution $\pi \sim Be(\alpha, \beta)$, the number of successes x is added to the first parameter while the number of failures $n - x$ is added to the second parameter. Note that the beta distribution is called *conjugate* to the binomial likelihood since the posterior is also beta distributed. Both prior and posterior density function are displayed in Fig. 31.2 for a simple example.

It is convenient to think of the $Be(\alpha, \beta)$ prior distribution as that which would have arisen had we started with an improper $Be(0, 0)$ prior and then observed α successes in $\alpha + \beta$ trials. Thus, $n_0 = \alpha + \beta$ can be viewed as a *prior sample size* and $\alpha/(\alpha + \beta)$ is the *prior mean*. This interpretation of the prior parameters is useful in order to intuitively assess the weight attached to the prior distribution, as we will see soon. It also stresses an important feature of Bayesian inference, the consistent processing of sequentially arising data. Indeed, suppose new independent data x_2 from the same $Bin(n, \pi)$ distribution arrives, then the posterior distribution following the original observation (with x now renamed to x_1) becomes the prior for the next observation x_2:

$$\mathrm{p}(\pi \mid x_1, x_2) \propto \mathrm{p}(x_2 \mid \pi) \times \mathrm{p}(\pi \mid x_1).$$

Here, we have been able to replace $\mathrm{p}(x_2 \mid \pi, x_1)$ by $\mathrm{p}(x_2 \mid \pi)$ due to the conditional independence of x_1 and x_2, given π. Now, $\mathrm{p}(\pi \mid x_1)$ is of course proportional to $\mathrm{p}(x_1 \mid \pi) \times \mathrm{p}(\pi)$, so an alternative formula is

$$\begin{aligned}\mathrm{p}(\pi \mid x_1, x_2) &\propto \mathrm{p}(x_2 \mid \pi) \times \mathrm{p}(x_1 \mid \pi) \times \mathrm{p}(\pi)\\ &= \mathrm{p}(x_1, x_2 \mid \pi) \times \mathrm{p}(\pi).\end{aligned}$$

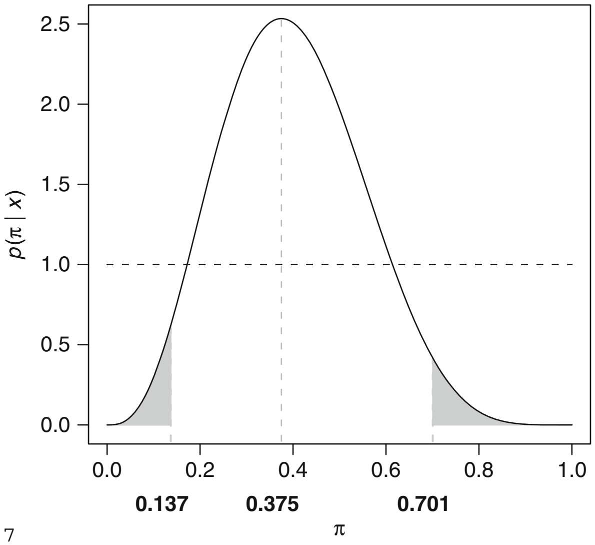


Fig. 31.2 A $Be(4, 6)$ posterior density $\mathrm{p}(\pi \mid x)$ (*solid line*) obtained from combining a $Be(1, 1)$ prior density (*dashed line*) with an observation $x = 3$ in a binomial experiment with $n = 8$ trials. The posterior mean is 0.4 and the posterior mode is 0.375. The equi-tailed 95% credible interval with limits $\theta_l = 0.137$ and $\theta_u = 0.701$ is also shown. The limits are calculated using the `R` function qbeta, that is, `qbeta(0.025,4,6) = 0.137`

In other words, $p(\pi \mid x_1, x_2)$ is the same whether or not the data are processed sequentially. Cornfield (1966, 1976) discusses this issue extensively in the context of clinical trials, see also Spiegelhalter et al. (2004, Sect. 4.3.2). Viewed from that perspective, every prior distribution is a posterior distribution based on the information available prior to the processing of the current data, and it makes sense to speak of a prior sample size. Bayesian inference is therefore sometimes described as Bayesian *learning*, which emphasizes the sequential nature inherent in the approach.

We now return to the posterior distribution in the binomial experiment. There are particularly simple explicit formulae for the mean and mode of a $Be(\alpha, \beta)$-distribution, see Appendix B for details. For example, the mean is simply $\alpha/(\alpha + \beta)$. Therefore, the posterior mean of $\pi \mid x \sim Be(\alpha + x, \beta + n - x)$ is

$$\frac{\alpha + x}{\alpha + \beta + n}.$$

Rewriting this as

$$\frac{\alpha + x}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \cdot \frac{x}{n}$$

shows that the posterior mean is a weighted average of the prior mean $\alpha/(\alpha + \beta)$ and the MLE $\bar{x} = x/n$ with weights proportional to the prior sample size $n_0 = \alpha + \beta$ and the data sample size n, respectively. This further supports the interpretation of n_0 as a prior sample size. The *relative prior sample size* $n_0/(n_0 + n)$ quantifies the weight of the prior mean in the posterior mean. Note that the relative prior sample size decreases with increasing data sample size n.

The case $\alpha = \beta = 1$ is of particular interest, as it corresponds to a uniform prior distribution on the interval $(0, 1)$, a natural "non-informative" choice. The prior sample size n_0 is now 2, one success and one failure. This is in fact exactly the prior used by Thomas Bayes (1702–1761) in his famous essay (Bayes 1763). The posterior mean is now $(x + 1)/(n + 2)$ and the posterior mode equals the MLE $\bar{x}$.

Incidentally, we note that Jeffreys' reference prior is not the uniform prior for π, but a beta distribution with both parameters α and β equal to 1/2, that is, $p(\pi) \propto \pi^{-0.5}(1 - \pi)^{-0.5}$ (compare Appendix B). This prior is proper and favors extreme values of π, that is, those which are close to either zero or one. From the above, we observe that Jeffreys' prior sample size n_0 is 1, half a success and half a failure.

Example 31.1. (Continued)
We now revisit the diagnostic test example discussed in Sect. 31.2 under the more realistic scenario that the disease prevalence $\pi = \Pr(D+)$ is not known, but only estimated from a prevalence study. For example, suppose there was $x = 1$ diseased individual in a random sample of size $n = 100$. Using a uniform $Be(1, 1)$ prior, the posterior distribution of π is $Be(2{,}100)$, with posterior mean $1/51$ and posterior mode $1/100$.

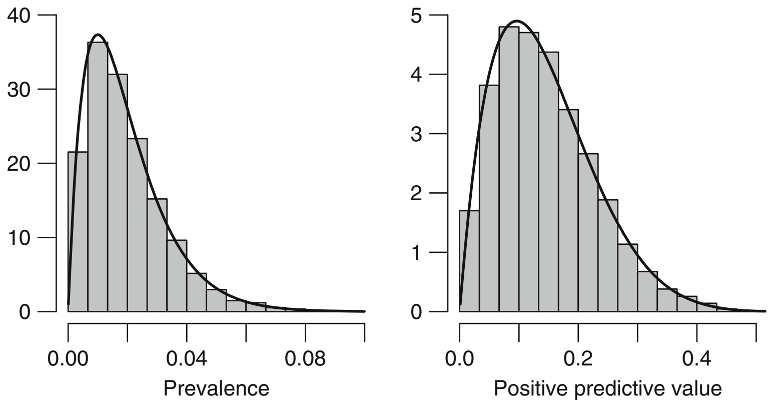


Fig. 31.3 *Left*: Posterior distribution of the prevalence. *Right*: Posterior distribution of the positive predictive value. Shown are histograms based on samples from these distributions. The *solid bold line* is the exact posterior density

It is tempting to replace the fixed prevalence $\Pr(D+) = 1/100$ in (31.1) with this $Be(2, 100)$ distribution to acknowledge the uncertainty involved in the prevalence estimation. The positive predictive value then follows a particular distribution, which can be computed analytically. However, it is much simpler to generate a random sample from the distribution of the positive predictive value using samples from the posterior distribution of π (Mossman and Berger 2001; Bayarri and Berger 2004). The following R-code illustrates this. Histograms of $n = 10{,}000$ samples from the prevalence and the associated positive predictive value are shown in Fig. 31.3.

```
> nsamples = 10000
> prev = rbeta(nsamples, 2, 100)
> sens = 0.9
> spec = 0.9
> ppv = sens * prev/(sens * prev + (1 - spec) * (1 - prev))
```

It is interesting that there is quite large uncertainty about the positive predictive value with 95% equi-tailed credible interval [0.02,0.34], which can be calculated from the corresponding quantiles of the sample. The 95% HPD credible interval is [0.009,0.31], so shifted to the left and slightly narrower, as expected. Note that the posterior mean is 0.145 (14.5%) and the posterior mode is 0.096 (9.6%). Both are larger than the positive predictive value 8.3% obtained for a fixed prevalence $\Pr(D+) = 0.01$ (see Sect. 31.2).

Mossman and Berger (2001) have considered a more general scenario where the characteristics of the diagnostic test are also not known exactly but based on estimates from a diagnostic study. Then, sensitivity `sens` and specificity `spec` in the above R-code needs to be replaced by corresponding samples from suitable beta (posterior) distributions, and the positive predictive value will have even more variation.

31.3.3 Bayesian Analysis of Incidence Rate Data

Incidence rate data typically consist of the number of cases x observed over the total person-time e. Alternatively, e may represent the expected number of cases under a specific assumption. A common approach, see, for example, Rothman (2002), is to assume that $X \sim Po(e\lambda)$, that is, the number of cases X is Poisson distributed with mean $e\lambda$ where e is a known constant and $\lambda > 0$ is an unknown parameter. If e is person-time, then λ represents the unknown incidence rate and if e is the number of expected cases, then λ is the unknown *rate ratio*, also called the *relative risk*.

It is commonplace to select a gamma distribution $Ga(\alpha, \beta)$ as prior distribution for λ, because it is conjugate to the Poisson likelihood, see Appendix B for details on the gamma distribution. The likelihood function of a Poisson observation with mean $e\lambda$ is

$$\mathrm{L}_x(\lambda) \propto \lambda^x \exp(-e\lambda).$$

It is easy to show that the MLE of λ is $\hat{\lambda}_{ML} = x/e$. Combining $\mathrm{L}_x(\lambda)$ with the density

$$\mathrm{p}(\lambda) \propto \lambda^{\alpha-1} \exp(-\beta\lambda)$$

of the gamma $Ga(\alpha, \beta)$ prior distribution, one obtains the posterior distribution of λ:

$$\begin{aligned}\mathrm{p}(\lambda \mid x) &\propto \mathrm{L}_x(\lambda) \times \mathrm{p}(\lambda)\\ &\propto \lambda^{\alpha+x-1} \exp(-(\beta+e)\lambda).\end{aligned}$$

This can be identified as another gamma distribution with parameters $\alpha + x$ and $\beta + e$:

$$\lambda \mid x \sim Ga(\alpha + x, \beta + e). \tag{31.10}$$

Compared with the prior distribution $Ga(\alpha, \beta)$, the number of observed counts x are added to the first parameter and the number of expected counts e are added to the second parameter.

The mean of a $Ga(\alpha, \beta)$ distribution is α/β, so the posterior mean is

$$\frac{\alpha + x}{\beta + e} = \frac{\beta}{\beta + e} \cdot \frac{\alpha}{\beta} + \frac{e}{\beta + e} \cdot \frac{x}{e}.$$

This equation illustrates that the posterior mean can be written as a weighted average of the prior mean α/β and the Maximum Likelihood estimate x/e with weights proportional to β and e, respectively. Hence, for Poisson data, there is a similar decomposition of the posterior mean as in the binomial case, see Sect. 31.3.2. Note that now β can be interpreted as prior sample size n_0 while e represents the data sample size.

Example 31.2. Breast cancer after fluroscopic examinations of the chest
For illustration, consider an example taken from Boice and Monson (1977), see also Greenland and Rothman (2008). A total of $x = 41$ breast cancer cases have been reported in a cohort of women treated for tuberculosis with x-ray fluoroscopy. Only $e = 23.3$ cases

were expected based on age-specific rates among women in Connecticut. We are interested in the posterior distribution of the rate ratio λ.

As prior distribution for the rate ratio λ, we may assume a gamma distribution with $\alpha = \beta$, and hence a prior mean of 1.0, that is, a prior expectation of a breast cancer rate after exposure to x-ray fluoroscopy equal to the overall rate in Connecticut. With a specific choice of α, we specify a range of plausible values around 1.0 which we consider believable a priori. For example, for $\alpha{=}\beta{=}8.78$, we believe that λ is in the range $[0.5, 2]$ with approximately 95% probability. Using this prior and Eq. 31.10, we obtain the posterior distribution $\lambda \mid x \sim Ga(8.78+41, 8.78+23.3) = Ga(49.78, 32.08)$. Note that the relative prior sample size is $8.78/32.08 \approx 27\%$, so the selected prior does have some weight in the posterior distribution.

The posterior mean of the relative risk λ is $49.78/32.08 = 1.55$. The equi-tailed 95% posterior credible interval for λ is $[1.15, 2.01]$. Thus, there is some evidence of an excess of breast cancers among x-rayed women relative to the reference group, but with quite large uncertainty about the actual size of the effect.

The above prior may be criticized for placing too much prior weight on relative risk values below unity with $\Pr(\lambda < 1) = 0.54$ and $\Pr(\lambda < 0.5) = 0.04$, but only $\Pr(\lambda > 2) = 0.01$. As a possible remedy, one may still pick a gamma prior with probability mass of 95% in the interval $[0.5, 2]$; however, one might want achieve symmetry by choosing a gamma prior which fulfills $\Pr(\lambda < 0.5) = \Pr(\lambda > 2) = 0.025$. This leads to the prior parameters $\alpha = 8.50$ and $\beta = 7.50$ and the posterior distribution $\lambda \mid x \sim Ga(49.50, 30.80)$. The posterior mean of the relative risk λ is now 1.61. The equi-tailed 95% posterior credible interval for λ is $[1.19, 2.08]$. The new prior gives very similar results compared to the original one.

Jeffreys' prior for the Poisson likelihood is a gamma distribution with parameters $\alpha = 1/2$ and $\beta = 0$. Since $\beta = 0$, it is an improper distribution but the associated posterior will be proper as long as $e > 0$. In the above example, the posterior mean 1.78 under Jeffreys' prior is larger as well as the limits of the 95% credible interval, which are 1.28 and 2.36.

31.3.4 Bayesian Analysis of Case-Control Data

We now turn to the Bayesian analysis of counts in a 2 × 2 table, with particular focus on the analysis of case-control studies with dichotomous exposure. Let E denote exposure and D disease, so let $\pi_1 = \Pr(E+ \mid D+)$ denote the probability that a case was exposed and $\pi_0 = \Pr(E+ \mid D-)$ the corresponding probability for a control. Assuming independent cases and controls and suitable (independent) priors for π_1 and π_0, we can easily derive the corresponding posterior distributions of π_1 and π_0, which are still independent. We may now proceed to infer the posterior distribution of the *odds ratio* $[\pi_1/(1-\pi_1)]/[\pi_0(1-\pi_0)]$. Conceptually, this is a simple mathematical problem; however, analytical calculation of the posterior density can be quite tedious (Nurminen and Mutanen 1987), so we use a simple Monte Carlo approach instead. The method gives independent samples from the posterior distribution of the odds ratio, as illustrated in the following example.

Example 31.3. Childhood leukemia and residential magnetic fields
Consider case-control data from Savitz et al. (1988), as reported in Greenland (2008), investigating a possible association between residential magnetic fields and childhood leukemia. For simplicity, the exposure variable was dichotomized based on a threshold of 3 milligauss (mG) exposure. The data are shown in Table 31.1. The entries of the 2 × 2 table are denoted by x_1 and y_1 for the exposed and unexposed cases, respectively, and x_0 and y_0

Table 31.1 Case-control data on residential magnetic field exposure and childhood leukemia

		Exposure	
		Exposed	Unexposed
Disease status	Cases	$x_1 = 3$	$y_1 = 5$
	Controls	$x_0 = 33$	$y_0 = 193$

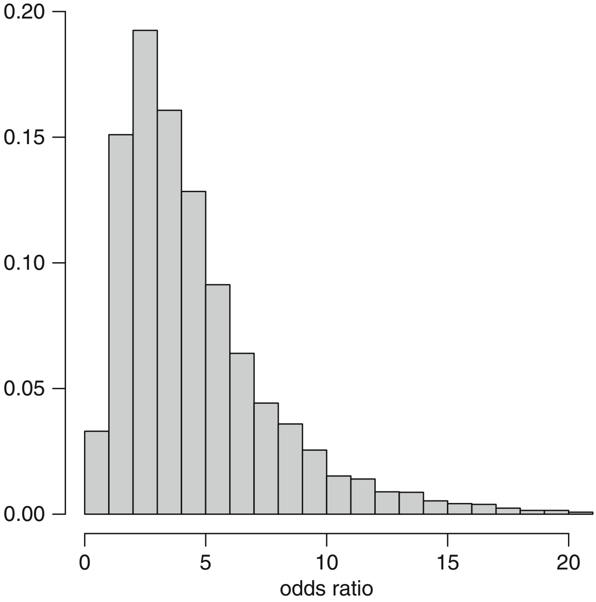


Fig. 31.4 A histogram based on 10,000 samples from the posterior odds ratio

for the corresponding controls. For simplicity, we continue to use the generic symbol x to denote all the available data for both cases and controls.

We assume independent uniform prior distributions for π_1 and π_0. It then follows that π_1 and π_0 are also a posteriori independent: $\pi_1 \,|\, x \sim Be(x_1 + 1 = 4, y_1 + 1 = 6)$ and $\pi_0 \,|\, x \sim Be(x_0 + 1 = 34, y_0 + 1 = 194)$, compare Sect. 31.3.2. In fact, Fig. 31.2 has shown the posterior of π_1. The following R-code illustrates a Monte Carlo approach to generate random samples from the posterior distribution of the odds ratio.

```
> nsamples = 10000
> pi1 = rbeta(nsamples, 4, 6)
> pi0 = rbeta(nsamples, 34, 194)
> or = (pi1/(1 - pi1))/(pi0/(1 - pi0))
```

Figure 31.4 gives a histogram of the posterior samples from the odds ratio `or`. The resulting posterior mean of the odds ratio is 4.7 with equi-tailed 95% credible interval [0.9,14.3]. Thus, there is large uncertainty about the odds ratio with values around unity not completely unrealistic.

The odds ratio $[\pi_1/(1-\pi_1)]/[\pi_0/(1-\pi_0)]$ is the odds $\pi_1/(1-\pi_1)$ to be exposed for a case divided by the odds $\pi_0/(1-\pi_0)$ to be exposed for a control, the so-called *exposure odds ratio*. Of more practical interest is the *disease odds ratio*, that is, the odds to be a case if exposed divided by the odds to be a case if not exposed. As Jerome Cornfield (1912–1979) has shown (Cornfield 1951) through yet another application of Bayes' theorem, the exposure odds ratio is in fact equal to the disease odds ratio. Cornfield's proof is quite simple: Bayes theorem (31.4) in odds form gives

$$\frac{\Pr(D+\mid E+)}{\Pr(D-\mid E+)} = \frac{\Pr(E+\mid D+)}{\Pr(E+\mid D-)} \times \frac{\Pr(D+)}{\Pr(D-)}.$$

and likewise

$$\frac{\Pr(D+\mid E-)}{\Pr(D-\mid E-)} = \frac{\Pr(E-\mid D+)}{\Pr(E-\mid D-)} \times \frac{\Pr(D+)}{\Pr(D-)}.$$

Dividing the first through the second equation gives after some rearrangement

$$\frac{\Pr(D+\mid E+)/\Pr(D-\mid E+)}{\Pr(D+\mid E-)/\Pr(D-\mid E-)} = \frac{\Pr(E+\mid D+)/\Pr(E-\mid D+)}{\Pr(E+\mid D-)/\Pr(E-\mid D-)}.$$

The left side of this equation is the disease odds ratio and the right side is the exposure odds ratio. For more details on statistical issues of the case-control study, see, for example, Breslow (1996) or chapter ▶Case-Control Studies of this handbook.

31.3.5 Approximate Bayesian Analysis

The shapes of many log likelihood functions $\log \mathrm{L}_x(\theta)$ are approximately quadratic, see, for example, Clayton and Hills (1993, Chap. 9). The log likelihood function of the normal distribution is exactly quadratic, and this fact can be used to apply techniques based on the normal distribution for approximate Bayesian inference (Greenland 2008). Methods based on approximate likelihoods are particularly important because the quadratic approximation becomes closer to the true likelihood as the sample size increases. Figure 31.5 illustrates this for the log likelihood of a Poisson observation $x = 41$ with mean $e \cdot \lambda$ where $e = 23.3$ (Example 31.2). Note that the log likelihood is shown not with respect to λ, but in terms of the log relative risk $\theta = \log(\lambda)$. The normal approximation is typically better if the parameter of interest is unrestricted, so it is better to approximate the log relative risk rather than the relative risk, which can take only positive values.

It is often appropriate to approximate a likelihood function of an unknown parameter θ by viewing the MLE as the actually observed (normal) data x. The associated standard error serves as (known) standard deviation $\sigma = \mathrm{se}(\hat{\theta}_{ML})$ of that normal distribution: $\hat{\theta}_{ML} \sim N(\theta, \sigma^2)$. The original likelihood function is hence replaced with its quadratic approximation, a likelihood function of one single normal observation x (the MLE) with known variance σ^2 (the squared standard

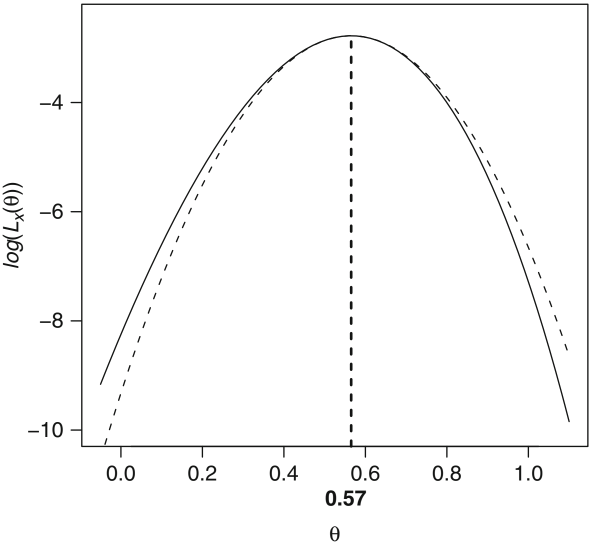


Fig. 31.5 Log likelihood function $\log \mathrm{L}_x(\theta)$ of a Poisson observation $x = 41$ with mean $e \cdot \exp(\theta) = 23.3 \cdot \exp(\theta)$ (*solid line*). Also shown is the MLE $\hat{\theta}_{ML} = 0.57$ and the quadratic approximation to the log likelihood (*dashed line*)

error). The unknown parameter θ is the mean of that normal distribution. Such an approach makes approximate Bayesian inference particularly simple, as we will see in the following.

So let X denote a sample from a normal $N(\theta, \sigma^2)$ distribution with mean θ and known variance σ^2. The corresponding likelihood function is

$$\mathrm{L}_x(\theta) \propto \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}.$$

Combined with a normal prior distribution for the unknown mean $\theta \sim N(\nu, \tau^2)$ with mean ν and variance τ^2, that is,

$$\mathrm{p}(\theta) \propto \exp\left\{-\frac{(\theta-\nu)^2}{2\tau^2}\right\}.$$

the posterior distribution is given by

$$\begin{aligned}\mathrm{p}(\theta \mid x) &\propto \mathrm{L}_x(\theta) \times \mathrm{p}(\theta) \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{(x-\theta)^2}{\sigma^2} + \frac{(\theta-\nu)^2}{\tau^2}\right)\right\}.\end{aligned}$$

It can be shown that this is the density function of yet another normal distribution with variance $\tilde{\tau}^2 = 1/(1/\sigma^2 + 1/\tau^2)$ and mean $\tilde{\nu} = \tilde{\tau}^2(x/\sigma^2 + \nu/\tau^2)$:

$$\theta \mid x \sim N(\tilde{\nu}, \tilde{\tau}^2). \tag{31.11}$$

Table 31.2 A comparison of posterior characteristics for various prior distributions in the breast cancer study. Shown is the posterior mean $\hat{\lambda}$, the limits λ_l and λ_u of the equi-tailed 95% credible interval, and the tail probability $\Pr(\lambda < 1|x)$

Prior distribution	$\hat{\lambda}$	λ_l	λ_u	$\Pr(\lambda < 1\|x)$
$Ga(8.78, 8.78)$	1.55	1.15	2.01	0.00226
$Ga(8.5, 7.5)$	1.61	1.19	2.08	0.00116
$Ga(0.5, 0.0)$ (Jeffreys' prior)	1.78	1.28	2.36	0.00042
$LN(0, 0.125)$ (approximate)	1.62	1.21	2.12	0.00047
$LN(0, 0.125)$ (exact)	1.60	1.17	2.10	0.00170
$LN(0, \infty)$ (approximate)	1.78	1.30	2.39	0.00015

As for binomial samples, the posterior mean is a weighted mean of the prior mean ν and the data x with weights proportional to $1/\tau^2$ and $1/\sigma^2$, respectively. The relative prior sample size is thus $\tilde{\tau}^2/\tau^2$.

Example 31.2. (Continued)

It is well-known that the MLE $\hat{\theta}_{ML} = \log(x/e)$ of the log relative risk $\theta = \log(\lambda)$ is approximately normally distributed with mean equal to the true log relative risk θ and standard error $\sigma = 1/\sqrt{x}$ (Clayton and Hills 1993, Chap. 9). Using the data on breast cancer incidence after fluroscopic examinations of the chest from Sect. 31.3.3 where $x = 41$ and $e = 23.3$, the MLE of θ is hence 0.57 with standard error 0.16.

The MLE $\hat{\theta}_{ML}$ serves now as a summary of the information in the data to update our prior beliefs about θ. As prior distribution for θ, we select a mean-zero normal distribution such that the relative risk $\lambda = \exp(\theta)$ is between 0.5 and 2 with 95% probability. The corresponding normal distribution has variance $\tau^2 = (\log(2)/1.96)^2 \approx 1/8$. Note that a normal distribution for the log relative risk corresponds to a so-called *log-normal distribution* for the relative risk, where explicit formulae for the mean and mode are available (see Appendix B).

Using Eq. 31.11, the posterior variance is $\tilde{\tau}^2 = 1/(x+8) \approx 0.02$ and the posterior mean is $\tilde{\nu} = \tilde{\tau}^2(\hat{\theta}_{ML}/\sigma^2) = 0.47$. This corresponds to a posterior mean of $\exp(\tilde{\nu} + \tau^2/2) = 1.62$ for the relative risk λ (see the formula for the mean of a log-normal distribution in Appendix B). The associated 95% equi-tailed credible interval for the relative risk is $[1.21, 2.12]$. Note that the relative prior sample size is $\tilde{\tau}^2/\tau^2 \approx 0.16$, that is, 16%.

If we combine the exact Poisson likelihood with a normal prior for the log relative risk parameter θ, then the posterior distribution is no longer analytically tractable. However, posterior characteristics can be computed using numerical techniques. One obtains the posterior mean 1.60 and the 95% credible interval [1.17,2.10] for λ. These results are very similar to those based on the approximate analysis.

If we let the prior variance τ^2 of a normal prior $N(0, \tau^2)$ for the log relative risk parameter θ go to infinity, we obtain a "locally uniform" or flat prior, $p(\theta) \propto 1$, which is sometimes described as "non-informative." In this case, the posterior Eq. 31.11 simplifies to

$$\theta \,|\, x \sim N(x, \sigma^2). \tag{31.12}$$

Therefore, the point estimate of θ is simply the MLE $\hat{\theta}_{ML}$, and the limits of the equi-tailed 95% credible interval are numerically identical to the limits of the standard 95% Wald confidence interval:

$$\hat{\theta}_{ML} \pm 1.96 \cdot \sigma.$$

Results based on this approximate analysis with a flat prior for the log relative risk parameter θ are similar to the ones based on the reference prior for the Poisson mean λ, see Table 31.2. A standard frequentist analysis can thus be regarded as a Bayesian analysis using a reference prior. This connection between frequentist and Bayesian parameter estimates can be established in many other situations, at least approximately. However, viewed from a Bayesian perspective, the frequentist approach uses a rather unrealistic prior which gives large weight to unrealistically extreme values of relative risk.

31.3.6 Bayesian Tail Probabilities

In classical hypothesis testing, a commonly encountered procedure is the so-called *one-sided hypothesis test* (see, e.g., Cox (2005)) where the evidence against a null hypothesis $H_0 : \theta \leq \theta_0$ is quantified using a p-value:

$$p\text{-value } = \Pr(T(X) \geq T(x) \,|\, \theta = \theta_0),$$

here $T(X)$ is a suitable summary of the data X, for example, the mean. The p-value obtained from such a one-sided hypothesis test has sometimes a Bayesian interpretation as the posterior probability of H_0:

$$\Pr(H_0 \,|\, x) = \Pr(\theta < \theta_0 \,|\, x).$$

For illustration, consider a simple scenario with $n = 1$ observation $T(X) = X$ from a normal distribution with unknown mean θ and known variance σ^2. Under the assumption of a reference prior $\mathrm{p}(\theta) \propto 1$, the posterior distribution is $\theta \,|\, x \sim N(x, \sigma^2)$, see (31.12). Therefore,

$$\Pr(H_0 \,|\, x) = \Pr(\theta < \theta_0 \,|\, x) = \Phi((\theta_0 - x)/\sigma),$$

where Φ denotes the cumulative distribution function of the standard normal distribution. On the other hand, the p-value against H_0 is

$$p\text{-value } = \Pr(X \geq x \,|\, \theta = \theta_0) = 1 - \Phi((x - \theta_0)/\sigma) = \Phi((\theta_0 - x)/\sigma),$$

so is numerically equal to the posterior probability $\Pr(H_0 \,|\, x)$.

Of course, posterior probabilities can be calculated also for other prior distributions, in which case the analogy between posterior probabilities and p-values will usually be lost.

Example 31.2. (Continued)
Table 31.2 lists the posterior probability $\Pr(\lambda < 1 \,|\, x)$ for different prior assumptions on the relative risk parameter λ. It can be seen that there is some variation of these tail probabilities depending on the prior distribution and on the usage of an exact or an approximate approach, respectively. The frequentist p-value based on the Poisson distribution is $\Pr(X \geq 41 \,|\, \lambda = 1, e = 23.3) = 0.00057$, so within the range of the reported tail probabilities.

Note that the posterior probability $\Pr(\lambda < 1 \,|\, x) = 0.00047$ using the approximate approach is somewhat different from the corresponding one calculated with the exact likelihood, which is 0.00170. The reason for this discrepancy is the approximation of the Poisson log likelihood through a quadratic function, which corresponds to the approximate normal distribution of the log relative risk. Figure 31.5 shows that the quadratic approximation is good around the MLE, but not so good for small values of θ, with larger values of the log likelihood than its quadratic approximation. This explains the difference between the approximate and the exact results.

In the following, we show that the relationship between p-values and posterior tail probability may also hold (approximately) in quite unexpected circumstances. However, it is important to emphasize that the apparent analogy between p-values and posterior tail probabilities holds only in special cases and does not extend to the commonly used *two-sided* hypothesis test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, as we will see later.

31.3.6.1 A Tail Probability for Case-Control Data

In 1877, the medical doctor Carl von Liebermeister (1833–1901) proposed a Bayesian approach for the analysis of counts in a 2 × 2 table (Liebermeister 1877). Carl von Liebermeister was at that time professor in the Medical Faculty at the University of Tübingen in southern Germany. A Bayesian approach was selected by Liebermeister, as it was the inferential method of its time, following the tradition of Thomas Bayes and Pierre-Simon Laplace.

In the following, we will adopt the notation from Sect. 31.3.4 on the Bayesian analysis of case-control data, with π_1 and π_0 denoting the probability that a case and a control was exposed, respectively. Liebermeister had the ingenious idea to consider the posterior probability

$$\Pr(\pi_1 \leq \pi_0 \mid x) \tag{31.13}$$

in order to assess if there is evidence for a "significant" difference between cases and controls with respect to the underlying exposure risk. Liebermeister selected independent uniform priors for the unknown probabilities π_1 and π_0, directly following the approach by Thomas Bayes. Note that in modern epidemiological terminology, Eq. 31.13 is the posterior probability that the *relative risk* π_1/π_0 is smaller or equal to one. Furthermore, this probability is identical to the posterior probability that the odds ratio $\pi_1(1 - \pi_0)/(\pi_0(1 - \pi_1))$ is smaller or equal to one, because $\pi_1/\pi_0 \leq 1$ if and only if $\pi_1(1 - \pi_0)/(\pi_0(1 - \pi_1)) \leq 1$. Analytical computation of Eq. 31.13 is far from trivial, as reviewed in Seneta (1994). Quite interestingly, it turns out that Eq. 31.13 is the p-value of Fisher's one-sided exact test when testing the null hypothesis $\pi_1 \leq \pi_0$ against the one-sided alternative $\pi_1 > \pi_0$, if the diagonal entries x_1 and y_0 of the 2 × 2 table (here we adapt the notation from Table 31.1) are both increased by one count. Note that addition of 1 on the diagonal increases the empirical odds ratio and hence decreases the p-value of the above test.

The close connection to Fisher's test, which was developed more than 50 years later (Fisher 1934), has led Seneta (1994) to call the Liebermeister approach a "Bayesian test procedure." Seneta and Phipps (2001) studied frequentist properties of Eq. 31.13, viewed as a classical p-value. They showed that it has better average frequentist properties than the p-value obtained from Fisher's original test.

Altham (1969) has derived formulae for Eq. 31.13 in the more general setting of two independent beta distributions for π_0 and π_1 with arbitrary parameters. Nurminen and Mutanen (1987) have further generalized these results and have provided formulae for the whole posterior distribution of the risk difference, the risk ratio and the odds ratio. An interesting review of the Bayesian analysis of the 2 × 2 table can be found in Howard (1998). Note that all these authors have apparently been unaware of the original work by Liebermeister.

An alternative approximate approach for Bayesian inference in the 2×2 table, sometimes called *semi-Bayes*, has also been suggested. The basic idea is to re-parameterize the model in terms of a parameter of interest (e.g., the log odds ratio) and a so-called nuisance parameter (e.g., the log odds in the control group). A posterior distribution is then derived for the parameter of interest, assuming a suitable prior distribution. It is well known that the likelihood function for the log odds ratio ψ is approximately normal with mean $\log(x_1 \cdot y_0)/(x_0 \cdot y_1)$ and variance $(1/x_1 + 1/y_1 + 1/x_0 + 1/y_0)$, see Clayton and Hills (1993, Chap. 17). Adopting a flat (improper) prior for ψ, the posterior distribution is therefore also approximately normal with that mean and variance, which allows for the computation of (approximate) Bayesian p-values based on the normal distribution function. Proper normal priors can be easily incorporated in this approach using the techniques described in Sect. 31.3.5.

Example 31.3. (Continued)
Consider again the case-control example described in Sect. 31.3.4. The p-value from Fisher's one-sided test is 0.108, whereas Liebermeister's probability Eq. 31.13, calculated as the p-value of Fisher's test applied to Table 31.1 with diagonal entries increased to 4 and 194, respectively, is 0.036. Using the approximate approach with a flat improper reference prior for ψ, the posterior probability that the log odds ratio is equal to or smaller than zero (and hence the odds ratio is equal to or smaller than one) turns out to be 0.048, so quite similar. Greenland (2008) suggests an informative mean-zero normal prior distribution for ψ with variance 1/2. This distribution implies that the prior probability for an odds ratio between 1/4 and 4 is (approximately) 95%. Then, the posterior probability that the odds ratio is equal to or smaller than one is 0.127, so larger than before.

31.4 Prior Criticism and Model Choice

Various statistical researchers have emphasized the importance of modeling and reporting uncertainty in terms of *observables*, as opposed to inference about (unobservable) parameters. However, the latter, more traditional approach to inference can be seen as a limiting form of *predictive* inference about observables (Bernardo and Smith 1994). Parametric inference can therefore be seen as an intermediate structural step in the predictive process.

A predictive model for observables, for example, future outcomes of a clinical trial, can be constructed easily within the Bayesian framework. As we will see in this section, the prior predictive distribution plays also a key role in prior criticism and Bayesian model choice.

31.4.1 Bayesian Prediction

Suppose we want to predict future data x^{new}, say, which is assumed to arise from the same likelihood function as the original data x. Bayesian prediction is based on the simple identity

$$\mathrm{p}(x^{\mathrm{new}} \mid x) = \int \mathrm{p}(x^{\mathrm{new}} \mid \theta) \times \mathrm{p}(\theta \mid x)\, d\theta, \tag{31.14}$$

so the *predictive distribution* of x^{new} given x is the integral of the likelihood function of x^{new} times the posterior distribution $\mathrm{p}(\theta \mid x)$ with respect to θ.

For example, consider the binomial model with unknown success probability π as described in detail in Sect. 31.3.2. Suppose we want to predict a future observation $X^{\mathrm{new}} \sim Bin(1, \pi)$. It is easy to show that $X^{\mathrm{new}} \mid x$ has a Bernoulli distribution with success probability equal to the mean of $\pi \mid x$.

The predictive distribution (31.14) is sometimes called *posterior predictive distribution* since it is conditional on the observed data x. In contrast, the *prior predictive distribution*

$$\mathrm{p}(x) = \int \mathrm{p}(x \mid \theta) \times \mathrm{p}(\theta)\, d\theta \tag{31.15}$$

is derived from the likelihood and the prior distribution alone. The prior predictive distribution plays a key role in Bayesian model criticism and model selection, as we will see in the following section.

Note that calculation of the prior predictive distribution requires that $\mathrm{p}(\theta)$ is proper; otherwise, $\mathrm{p}(x)$ would be undefined. Note also that $\mathrm{p}(x)$ is the denominator in Bayes' theorem (31.7). Therefore,

$$\mathrm{p}(x) = \frac{\mathrm{p}(x \mid \theta) \times \mathrm{p}(\theta)}{\mathrm{p}(\theta \mid x)}, \tag{31.16}$$

which holds for any value of θ. This formula is very useful if both prior and posterior are available in closed form, in which case the integration in definition (31.15) can be avoided. However, it is necessary to include all normalizing constants in $\mathrm{p}(x \mid \theta)$, $\mathrm{p}(\theta)$, and $\mathrm{p}(\theta \mid x)$, which makes the calculations slightly more tedious.

31.4.2 Prior Criticism

Box (1980) has suggested an approach to compare priors with subsequent data. The method is based on a p-value obtained from the prior predictive distribution and the actually observed datum. Small p-values indicate a *prior-data conflict*, that is, incompatibility of prior assumptions and the actual observations.

Box's p-value is defined as the probability of obtaining a result with prior predictive ordinate $\mathrm{p}(X)$ equal to or lower than at the actual observation x:

$$\Pr(\mathrm{p}(X) \leq \mathrm{p}(x)).$$

If both data and prior are normal, $X \mid \theta \sim N(\theta, \sigma^2)$ and $\theta \sim N(\nu, \tau^2)$, then the prior predictive distribution is $X \sim N(\nu, \sigma^2 + \tau^2)$. It can be shown that Box's p-value is then the upper tail probability of a chi-squared distribution with 1 degree of freedom (a more common name for a $Ga(1/2, 1/2)$ distribution) evaluated at

$$v^2 = \frac{(x-\nu)^2}{\sigma^2+\tau^2}.$$

Example 31.3. (Continued)
Let us revisit the case-control study from Sect. 31.3.4. The MLE of θ is $x = \log(3 \cdot 193/(5 \cdot 33)) \approx 1.255$ with standard error $\sigma = \mathrm{se}(\hat{\theta}_{ML}) = \sqrt{1/3 + 1/193 + 1/5 + 1/33} = 0.754$. Greenland's prior mean for the log odds ratio θ was $\nu = 0$ and the prior variance was $\tau^2 = 0.5$. We, hence, obtain

$$v^2 = \frac{1.255^2}{0.754^2 + 0.5} = 1.47$$

with an associated p-value of `1 - pgamma(1.47, 1/2, 1/2)` =0.22. Thus, by this check, the prior and the actually observed data appear to be fairly compatible, because Box's p-value is not remarkably small.

31.4.3 Bayesian Model Selection

Suppose we entertain two competing Bayesian models M_0 and M_1 and we are interested to know which one describes the data x better. Bayesian model choice is based on a variant of Eq. 31.4. Suppose we denote by $\Pr(M_0)$ and $\Pr(M_1)$ the prior probabilities of model M_0 and M_1, respectively, with $\Pr(M_0) + \Pr(M_1) = 1$. Then, the following fundamental equality holds:

$$\frac{\Pr(M_0 \mid x)}{\Pr(M_1 \mid x)} = \frac{\mathrm{p}(x \mid M_0)}{\mathrm{p}(x \mid M_1)} \times \frac{\Pr(M_0)}{\Pr(M_1)}. \tag{31.17}$$

Here, $\Pr(M_0)/\Pr(M_1)$ are the *prior odds*, $\Pr(M_0 \mid x)/\Pr(M_1 \mid x)$ are the *posterior odds*, and $\mathrm{p}(x \mid M_0)/\mathrm{p}(x \mid M_1)$ is the so-called *Bayes factor*, the ratio of the prior predictive distributions of the two models, both evaluated at the observed data x. The Bayes factor, which can be larger or smaller than one, summarizes the evidence of the data for the two models. If the Bayes factor is larger than one, then there is evidence for model M_0; otherwise, there is evidence for model M_1. Note that the Bayesian approach to model selection treats the two models M_0 and M_1 in a symmetric fashion, whereas classical hypothesis tests can only reject, but never accept the simpler model.

The term $\mathrm{p}(x \mid M)$ is also known as the *marginal likelihood*, to contrast it with the ordinary (conditional) likelihood $\mathrm{p}(x \mid \theta, M)$. The marginal likelihood can be calculated based on the prior predictive distribution (31.15).

Example 31.3. (Continued)
We revisit the approximate Bayesian analysis for case-control data and compare model M_0 with a fixed odds ratio of one with model M_1, where we use as before a $N(0, 0.5)$ prior for the log odds ratio ψ. This model comparison is the Bayesian version of the classical

two-sided hypothesis test for the null hypothesis that the odds ratio equals one. As before, we adopt an approximate Bayesian analysis assuming that the observed log odds ratio $\hat{\psi} = 1.26$ is normally distributed with known variance 0.57. The (marginal) likelihood in model M_0 is thus simply the density of a normal distribution with mean zero and variance 0.57, evaluated at $\hat{\psi} = 1.26$. This turns out to be 0.13. The prior predictive distribution in model M_1 is also normal with mean zero, but with variance $0.5 + 0.57 = 1.07$, so the marginal likelihood in model M_1 is 0.18. The Bayes factor of model M_0 relative to model M_1 is therefore $0.13/0.18 = 0.72$. Assuming 1 to 1 prior odds, the posterior odds for M_0 versus M_1 are therefore 0.72, and the corresponding posterior probability of model M_0 has decreased from 0.5 to $0.72/(1 + 0.72) = 0.42$ using the formula in Footnote 1 on page 1164.

It is somewhat surprising that the posterior probability has barely changed, despite a fairly small p-value obtained from Fisher's two-sided test ($p = 0.11$). The corresponding Wald test gives a similar result ($p = 0.096$). This illustrates that the correspondence between Bayesian model selection and p-values is typically lost for the standard two-sided hypothesis test (Berger and Sellke 1987). In particular, p-values cannot be interpreted as posterior probabilities of the null hypothesis.

In the following, we will study the two-sided hypothesis test $M_0 : \theta = 0$ versus $M_1 : \theta \neq 0$ in more detail, assuming that the MLE $\hat{\theta}_{ML}$ is normal distributed with unknown mean θ and known variance σ^2, equal to the squared standard error of $\hat{\theta}_{ML}$. This scenario reflects, at least approximately, many of the statistical procedures found in epidemiological journals.

We now have the possibility to calculate a *minimum Bayes factor* (MBF) (Edwards et al. 1963; Goodman 1999a,b), a lower bound on the evidence against the null hypothesis. The idea is to consider a whole family of prior distributions and to derive a lower bound for the Bayes factor in that family, the minimum Bayes factor. The approach can be taken to the limit by considering all possible prior distributions, in which case the minimum Bayes factor is a universal bound on the evidence of the data against the null hypothesis. Interestingly, the prior distribution $\mathrm{p}(\theta)$ in model M_1 with smallest Bayes factor is concentrated at the MLE $\hat{\theta}_{ML}$, that is, assumes $\theta = \hat{\theta}_{ML}$ a priori. If a z-value $z = \hat{\theta}_{ML}/\mathrm{se}(\hat{\theta}_{ML})$ has been calculated for this two-sided test, then the following formula can be used to calculate this universal minimum Bayes factor (Goodman 1999b):

$$\mathrm{MBF} = \exp\left(-\frac{z^2}{2}\right).$$

For example, if $z = 1.96$, where the two-sided p-value is 0.05, then $\mathrm{MBF} = \exp(-1.96^2/2) \approx 0.15$. If we assume 1 to 1 prior odds, then a universal lower bound on the posterior probability of the null hypothesis M_0 is therefore $0.15/(1+0.15) = 0.13$.

However, the above approach has been criticized since a prior distribution concentrated at the MLE is completely unrealistic since we do not know the MLE a priori. In addition, since the alternative hypothesis has all its prior density on one side of the null hypothesis, it is perhaps more appropriate to compare the outcome of this procedure with the outcome of a one-sided rather than a two-sided test, in which case MBF ≈ 0.26, so considerably larger.

Minimum Bayes factors can also be derived in more realistic scenarios. A particularly simple approach (Sellke et al. 2001) leads to the formula

$$\mathrm{MBF} = c \cdot p \log(p),$$

where $c = -\exp(1) \approx -2.72$ and p denotes the p-value from the two-sided hypothesis test (assumed to be smaller than $\exp(-1) \approx 0.37$). For example, for $p = 0.05$, we obtain MBF ≈ 0.41.

Example 31.3. (Continued)
In the case-control example, $z = 1.255/0.754 \approx 1.66$, and we obtain the minimum Bayes factor of $\exp(-1.66^2/2) \approx 0.25$, that is, 1 to 4, and a lower bound of 0.2 on the corresponding posterior probability of model M_0 (assuming 1 to 1 prior odds). Thus, it is impossible that the posterior probability of the null hypothesis is smaller than 0.2 if our prior probability was 0.5.

Of course, from the above, z-value a p-value can be easily calculated, which turns out to be $p = 0.096$. Using this p-value, the more realistic Sellke et al. (2001) approach gives a minimum Bayes factor of 0.61, which corresponds to a lower bound of 0.38 on the corresponding posterior probability. We conclude that for two-sided hypothesis tests, the evidence against the null hypothesis is by far not as strong as the p-value seems to suggest. This general finding is discussed extensively in the literature (Edwards et al. 1963; Berger and Sellke 1987; Goodman 1999a,b; Sellke et al. 2001; Goodman 2005).

31.5 Further Topics

We now discuss more advanced techniques of Bayesian inference: empirical Bayes approaches and Markov chain Monte Carlo methods.

31.5.1 Empirical Bayes Approaches

Empirical Bayes methods are a combination of the Bayesian approach with likelihood techniques. The general idea is to estimate parameters of the prior distribution $\mathrm{p}(\theta)$ from multiple experiments, rather than fixing them based on prior knowledge. Strictly speaking, this is not a fully Bayesian approach, but it can be shown that empirical Bayes estimates have attractive theoretical properties. Empirical Bayes techniques are often used in various applications. For a general discussion, see also Davison (2003, Sect. 11.5). Here, we sketch the idea in an epidemiological context discussing shrinkage estimates of age-standardized relative risks for use in disease mapping.

Suppose that for each region $i = 1, \ldots, n$ the observed number of cases x_i of a particular disease are available as well as the expected number e_i under the assumption of a constant disease risk. We now present a commonly used empirical Bayes procedure which is due to Clayton and Kaldor (1987).

Assume that $x_1, \dots, x_n$ are independent realizations from $Po(e_i \lambda_i)$ distributions with known expected counts $e_i > 0$ and unknown region-specific parameters λ_i. A suitable prior for the λ_i's is a gamma distribution, $\lambda_i \sim Ga(\alpha, \beta)$, due to the conjugacy of the gamma distribution to the Poisson likelihood. The posterior of λ_i turns out to be

$$\lambda_i \mid x_i \sim Ga(\alpha + x_i, \beta + e_i) \tag{31.18}$$

with posterior mean $(\alpha + x_i)/(\beta + e_i)$, compare Sect. 31.3.3. If α and β are fixed in advance, the posterior of λ_i does not depend on the data x_j and e_j from the other regions $j \neq i$.

Empirical Bayes estimates of λ_i are based on (31.18), but the parameters α and β of the prior distribution are not fixed in advance but estimated based on all available data. This is done by maximizing the implied prior predictive distribution or marginal likelihood, which depends only on α and β. One obtains MLEs $\hat{\alpha}_{ML}$ and $\hat{\beta}_{ML}$ of α and β, which are plugged into Formula (31.18). The resulting posterior mean estimates

$$\frac{\hat{\alpha}_{ML} + x_i}{\hat{\beta}_{ML} + e_i} \tag{31.19}$$

are called *empirical Bayes estimates* of λ_i. They will always lie between the MLEs x_i/e_i and the estimated mean $\hat{\alpha}_{ML}/\hat{\beta}_{ML}$ of the gamma prior; thus, the MLEs are shrunk toward the common value $\hat{\alpha}_{ML}/\hat{\beta}_{ML}$. This phenomenon is called *shrinkage*.

Example 31.4. Lip cancer in Scotland
Consider data on the incidence of lip cancer in $n = 56$ regions of Scotland, as reported in Clayton and Kaldor (1987). Here we obtain $\hat{\alpha}_{ML} = 1.88$ and $\hat{\beta}_{ML} = 1.32$. Figure 31.6 displays the empirical Bayes estimates and the corresponding 95% equi-tailed credible intervals, ordered with respect to the MLEs. Figure 31.6 shows clearly that the MLEs x_i/e_i are shrunk to the prior mean, that is, the empirical Bayes estimates lie between these two extremes. A map of Scotland with the empirical Bayes estimates is shown in Fig. 31.7.

31.5.2 Markov Chain Monte Carlo

Application of ordinary Monte Carlo methods is difficult if the unknown parameter is of high dimension. However, *Markov chain Monte Carlo* (MCMC) methods will then be a useful alternative. The idea is to simulate a *Markov chain* $\theta^{(1)}, \dots, \theta^{(m)}, \dots$ in a specific way such that it converges to the posterior distribution $p(\theta \mid x)$. After convergence, one obtains random samples from the target distribution, which can be used to estimate posterior characteristics as in ordinary Monte Carlo approaches. To ensure that the samples are taken from the target distribution, in practice, the first iterations, the so-called *burn-in*, are typically ignored. However, note that these samples will be dependent, an inherent feature of Markov chains.

The theory of MCMC is beyond the scope of this chapter, but we will illustrate the procedure in the context of disease mapping as discussed in Sect. 31.5.1. We

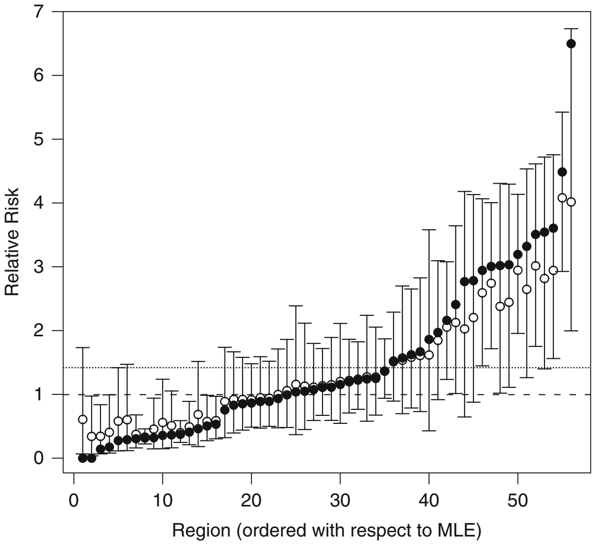


Fig. 31.6 Ninety-five percent equi-tailed credible intervals for Scottish lip cancer incidence rates λ_i $(i = 1, \dots, 56)$, calculated with an empirical Bayes approach. The *dotted line* marks the MLE $\hat{\alpha}_{ML}/\hat{\beta}_{ML} = 1.42$ of the prior mean. *Open circles* denote the posterior mean estimates of λ_i. The regions are ordered with respect to their MLEs x_i/e_i, shown as *filled circles*

now specify a prior on the log relative risks $\theta_i = \log(\lambda_i)$ which takes into account spatial structure and thus allows for spatial dependence (Besag et al. 1991). More specifically, we use a *Gaussian Markov random field* (GMRF), most easily specified through the conditional distribution of θ_i given $\theta_{j \neq i}$, that is, the log relative risks in all other regions $j \neq i$. A common choice is to assume that

$$\theta_i \mid \theta_{j \neq i}, \tau^2 \sim N\left(\bar{\theta}_i, \frac{\tau^2}{n_i}\right), \tag{31.20}$$

here $\bar{\theta}_i = n_i^{-1} \sum_{j \sim i} \theta_j$ denotes the mean of the n_i spatially neighboring regions of region i and τ^2 is an unknown variance parameter. Some decision has to be made to connect the two islands shown in Fig. 31.7 to the rest of Scotland. Here, we assume that they are both adjacent to the nearest mainland region.

To simulate from the posterior distribution a specific MCMC approach, the *Gibbs sampler*, iteratively updates the unknown parameters $\theta_1, \dots, \theta_n, \tau^2$. We omit details here but refer the interested reader to the relevant literature, for example, Rue and Held (2005).

Example 31.4. (Continued)
We now revisit the lip cancer incidence data in the $n = 56$ geographical regions of Scotland, allowing for spatial dependence between the relative risk parameters as described above. The following results are based on a Markov chain of length 100,000 where the first 10,000 samples were disregarded as burn-in. Figure 31.8 displays the corresponding posterior mean relative risks. Compared with the empirical Bayes estimates shown in Fig. 31.7, obtained from a model without spatial dependence, a spatially smoother picture can be observed.

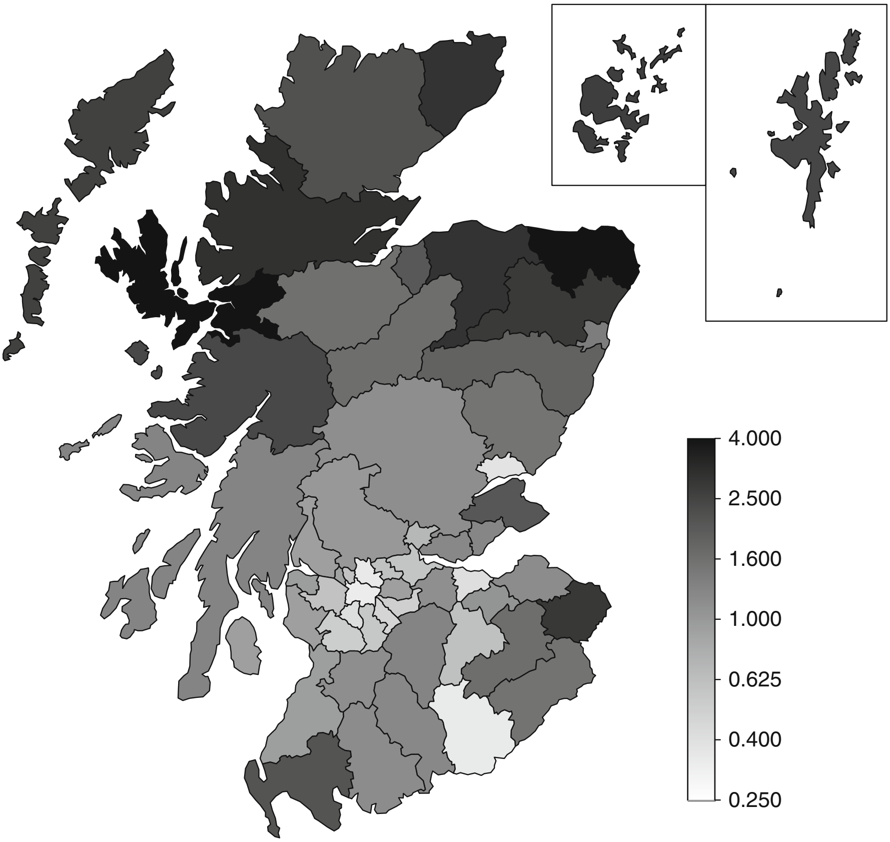


Fig. 31.7 Geographical distribution of the empirical Bayes estimates of the relative risk of lip cancer in Scotland

31.6 Conclusions

The Bayesian approach to statistical inference offers a coherent framework, in which both parameter estimation and model selection can be addressed. The key ingredient is the prior distribution, which reflects our knowledge about parameters or models before we integrate new data in our analysis. Bayesian statistics produces statements about the uncertainty of unknown quantities conditional on known data. This natural approach is in sharp contrast to frequentist procedures, which produce probability statements about hypothetical repetitions conditional on the unknown parameter and model.

The key to Bayesian statistics is the representation of prior beliefs through appropriate probability distributions. The key technique to update these prior beliefs in the light of new data is Bayes' theorem. Bayesian inference thus provides a

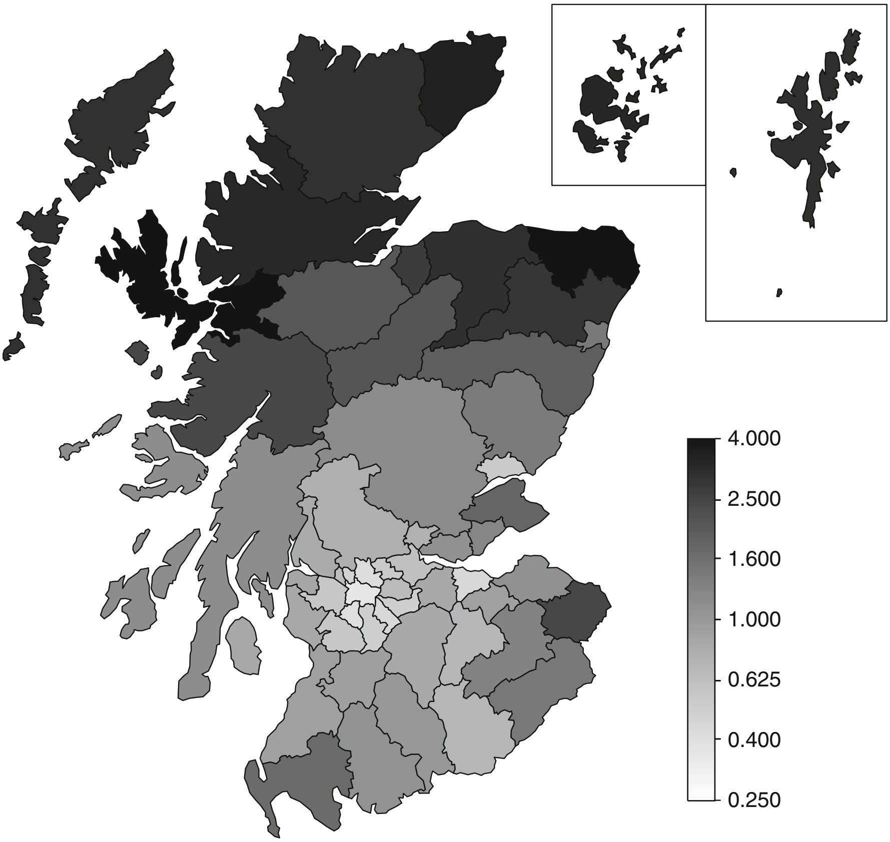


Fig. 31.8 Geographical distribution of the posterior mean relative risk estimates of lip cancer in Scotland, obtained through a GMRF approach

coherent way to update our knowledge in the light of new data. In the absence of conjugacy, the computation of the posterior distribution may require certain advanced numerical techniques such as Markov chain Monte Carlo.

I think it is important to emphasize relationships and differences between the frequentist and the Bayesian approach in order to appreciate what each of the different inference schools has to offer. A frequentist approach to parameter estimation based on the likelihood function alone can be regarded as a limited Bayesian form of inference in the absence of any prior knowledge. Such an approach typically leads to numerically similar results of both point and interval estimates. However, the possibility to specify a prior distribution is increasingly considered as something useful, avoiding implicit unrealistic assumptions of a frequentist analysis. Empirical Bayes approaches, which estimate a prior distribution from multiple experiments, are a compromise between the frequentist and Bayesian approach.

However, the frequentist and the Bayesian approach can lead to very different answers when it comes to hypothesis testing. A striking example is the two-sided hypothesis test, where the evidence against the null hypothesis, quantified by the Bayes factor, is by far not as strong as the p-value might suggest.

Appendix A. Rules of Probability

In this appendix, we summarize basic rules from probability theory. We also give a summary of important probability distributions.

A.1. Probabilities and Conditional Probabilities

Any experiment involving randomness can be modeled with probabilities. Probabilities are assigned to events such as "It will be raining tomorrow" or "I will suffer a heart attack in the next year." The *certain event* has probability one while the *impossible event* has probability zero. From a Bayesian perspective, probabilities are subjective in the sense that they quantify personal uncertainty that the event considered actually happens. Subjective probabilities can be elicited with a simple bet. If the actual realization of the event considered gives a return of 100 US dollar, say, and somebody is willing to bet up to but not more than p US dollars on that event happening, then his personal probability for the event is $p/100$.

Any event A has a disjoint, *complementary event* A^c such that $\Pr(A)+\Pr(A^c) = 1$. For example, if A is the event that "It will be raining tomorrow" then A^c is the event that "It will be not raining tomorrow." More generally, a series of events $A_1, A_2, \ldots, A_n$ is called a *partition* if the events are pairwise disjoint and if $\Pr(A_1) + \ldots + \Pr(A_n) = 1$.

Conditional probabilities $\Pr(A \mid B)$ are calculated to update the probability $\Pr(A)$ of a particular event under the additional information that a second event B has occurred. They can be calculated via

$$\Pr(A \mid B) = \frac{\Pr(A, B)}{\Pr(B)}, \tag{A.1}$$

where $\Pr(A, B)$ is the probability that both A and B occur. Rearranging this equation gives $\Pr(A, B) = \Pr(A \mid B)\Pr(B)$, but $\Pr(A, B) = \Pr(B \mid A)\Pr(A)$ must obviously also hold. Equating and rearranging these two formulas gives Bayes' theorem:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}. \tag{A.2}$$

Conditional probabilities behave like ordinary probabilities if the conditional event is fixed, so $\Pr(A \mid B) + \Pr(A^c \mid B) = 1$. It then follows that

$$\Pr(B) = \Pr(B \mid A) \Pr(A) + \Pr(B \mid A^c) \Pr(A^c), \tag{A.3}$$

and more generally,

$$\begin{aligned} \Pr(B) = \Pr(B \mid A_1) \Pr(A_1) + \Pr(B \mid A_2) \Pr(A_2) + \ldots \\ \ldots + \Pr(B \mid A_n) \Pr(A_n) \end{aligned} \tag{A.4}$$

if $A_1, A_2, \ldots, A_n$ is a partition. This is called the *law of total probability*. Equation (A.3) and (A.4) may be useful to calculate the denominator in Eq. (A.2).

A.2. Probability Functions

We now switch notation and replace Pr with p and events A and B with possible realizations x and y of *random variables* X and Y to indicate that the rules described in Appendix A.1 hold for any event considered. The formulas also hold if continuous random variables are considered, in which case $\mathrm{p}(\cdot)$ is a *density function*. For example, Eq. (A.1) now reads

$$\mathrm{p}(x \mid y) = \frac{\mathrm{p}(x, y)}{\mathrm{p}(y)} \tag{A.5}$$

while Bayes' theorem (A.2) translates to

$$\mathrm{p}(x \mid y) = \frac{\mathrm{p}(y \mid x)\, \mathrm{p}(x)}{\mathrm{p}(y)}. \tag{A.6}$$

Similarly, the law of total probability (A.4) now reads

$$\mathrm{p}(y) = \int \mathrm{p}(y \mid x)\, \mathrm{p}(x) dx, \tag{A.7}$$

where the integral $\int dx$ with respect to x is to be understood as a sum over x if $\mathrm{p}(x)$ is the probability function of a discrete random variable X. Combining equations (A.5) and (A.7) shows that the variable x has to be integrated out of the joint density $\mathrm{p}(x, y)$ to obtain the marginal density $\mathrm{p}(y)$ of Y:

$$\mathrm{p}(y) = \int \mathrm{p}(x, y)\, dx. \tag{A.8}$$

Appendix B. Important Probability Distributions

The following table gives some elementary facts about the probability distributions used in this chapter. A random variable is denoted by X, and its probability or density function is denoted by $\mathrm{p}(x)$. For each distribution, the mean $\mathrm{E}(X)$, variance $\mathrm{var}(X)$, and mode $\mathrm{mod}(X)$ is listed, if appropriate.

In the first row, we list the name of the distribution, an abbreviation, and the core of the corresponding `R`-function (*e.g.* `_norm`). Depending on the first letter, represented by the placeholder "`_`," these functions can be conveniently used as follows:

`r` stands for *r*andom and generates independent random numbers from that distribution. For example, `rnorm(n, mean = 0, sd = 1)` generates n random numbers from the standard normal distribution.

`d` stands for *d*ensity and returns the probability and density function, respectively. For example, `dnorm(x)` gives the density of the standard normal distribution.

`p` stands for *p*robability and gives the so-called *distribution function* of X. For example, if X is standard normal, then `pnorm(0)` returns 0.5 while `pnorm(1.96)` is 0.975.

`q` stands for *q*uantile and gives the *quantile function*. For example, `qnorm(0.975)` is $1.959964 \approx 1.96$.

The first argument `arg` depends on the particular function used. It is either the number n of random variables generated, a value x in the domain of the random variable, or a probability p with $0 < p < 1$.

Binomial: $Bin(n, \pi)$	$\texttt{\_binom(arg, size} = n, \texttt{prob} = \pi)$
$0 < \pi < 1, n \in \{1, \ldots, n\}$	$x \in \{0, \ldots, n\}$
$\mathrm{p}(x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}$	$\mathrm{L}_x(\pi) \propto \pi^x(1-\pi)^{n-x}$
$\mathrm{E}(X) = n\pi$	$\mathrm{var}(X) = n\pi(1-\pi)$
If $n = 1$ one obtains the *Bernoulli distribution*.	

Poisson: $Po(\lambda)$	$\texttt{\_pois(arg, lambda} = \lambda)$
$\lambda > 0$	$x \in \{0, 1, \ldots\}$
$\mathrm{p}(x) = \frac{\lambda^x}{x!}\exp(-\lambda)$	$\mathrm{L}_x(\lambda) \propto \lambda^x \exp(-\lambda)$
$\mathrm{E}(X) = \lambda$	$\mathrm{var}(X) = \lambda$

Beta: $Be(\alpha, \beta)$	$\texttt{\_beta(arg, shape1} = \alpha, \texttt{shape2} = \beta)$
$\alpha, \beta > 0$	$0 < x < 1$
$\mathrm{p}(x) = \mathrm{const} \cdot x^{\alpha-1}(1-x)^{\beta-1}$	
$\mathrm{E}(X) = \frac{\alpha}{\alpha+\beta}$	$\mathrm{mod}(X) = \frac{\alpha-1}{\alpha+\beta-2}$ if $\alpha, \beta > 1$
For $\alpha = \beta = 1$ one obtains the uniform distribution on the interval (0, 1).	

Gamma: $Ga(\alpha, \beta)$	$\texttt{\_gamma(arg, shape} = \alpha, \texttt{rate} = \beta)$
$\alpha, \beta > 0$	$x > 0$
$\mathrm{p}(x) = \mathrm{const} \cdot x^{\alpha-1}\exp(-\beta x)$	
$\mathrm{E}(X) = \alpha/\beta$	$\mathrm{mod}(X) = (\alpha-1)/\beta$ if $\alpha > 1$

Normal: $N(\theta, \sigma^2)$	$\_\mathtt{norm}(\mathtt{arg}, \mathtt{mu} = \theta, \mathtt{sd} = \sigma)$
$\sigma^2 > 0$	
$\mathrm{p}(x) = \text{const} \cdot \frac{1}{\sigma} \exp\left(-\frac{1}{2}\frac{(x-\theta)^2}{\sigma^2}\right)$	$\mathrm{L}_x(\theta) \propto \exp\left(-\frac{1}{2}\frac{(x-\theta)^2}{\sigma^2}\right)$
$E(X) = \theta$	$\mathrm{var}(X) = \sigma^2$
$N(0, 1)$ is called standard normal distribution.	
Log–normal: $LN(\theta, \sigma^2)$	$\_\mathtt{lnorm}(\mathtt{arg}, \mathtt{meanlog} = \theta, \mathtt{sdlog} = \sigma)$
$\sigma^2 > 0$	$x > 0$
$E(X) = \exp(\theta + \sigma^2/2)$	$\mathrm{mod}(X) = \exp(\theta - \sigma^2)$
$\mathrm{var}(X) = (\exp(\sigma^2) - 1) \cdot E(X)^2$	
If X is normal, i.e. $X \sim N(\theta, \sigma^2)$, then $\exp(X) \sim LN(\theta, \sigma^2)$.	

References

Altham PME (1969) Exact Bayesian analysis of a 2 × 2 contingency table and Fisher's "exact" significance test. J R Stat Soc B 31:261–269

Bayarri MJ, Berger JO (2004) The interplay of Bayesian and frequentist analysis. Stat Sci 19(1):58–80

Bayes T (1763) An essay towards solving a problem in the doctrine of chances. Philos Trans R Soc 53:370–418

Berger JO, Sellke T (1987) Testing a point null hypothesis: irreconcilability of P values and evidence (with discussion). J Am Stat Assoc 82:112–139

Bernardo JM, Smith AFM (1994) Bayesian theory. Wiley, Chichester

Besag J, York J, Mollié A (1991) Bayesian image restoration, with two applications in spatial statistics (with discussion). Ann Inst Stat Math 43:1–59

Boice JD, Monson RR (1977) Breast cancer in women after repeated fluroscopic examinations of the chest. J Natl Cancer Inst 59:823–832

Box GEP (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). J R Stat Soc A 143:383–430

Breslow NE (1996) Statistics in epidemiology: the case-control study. In: Armitage P, David HA (eds) Advances in biometry. Wiley, New York, pp 287–318

Clayton D, Hills M (1993) Statistical models in epidemiology. Oxford University Press, Oxford

Clayton D, Kaldor J (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43:671–681

Cornfield J (1951) A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. J Natl Cancer Inst 11:1269–1275

Cornfield J (1966) Sequential trials, sequential analysis and the likelihood principle. Am Stat 20:18–23

Cornfield J (1976) Recent methodological contributions to clinical trials. Am J Epidemiol 104:408–421

Cox DR (2005) Principles of statistical inference. Cambridge University Press, Cambridge

Davison AC (2003) Statistical models. Cambridge University Press, Cambridge

Edwards W, Lindman H, Savage LJ (1963) Bayesian statistical inference in psychological research. Psychol Rev 70:193–242

Fisher RA (1934) Statistical methods for research workers, 3rd edn. Oliver and Boyd, Edinburgh

Goodman SN (1999a) Towards evidence-based medical statistics. 1: the P value fallacy. Ann Int Med 130:995–1004

Goodman SN (1999b) Towards evidence-based medical statistics. 2: the Bayes factor. Ann Int Med 130:1005–1013

Goodman SN (2005) Introduction to Bayesian methods I: measuring the strength of evidence. Clin Trials 2:282–290

Greenland S (2006) Bayesian perspectives for epidemiological research: I. foundations and basic models. Int J Epidemiol 35:765–775

Greenland S (2007) Bayesian perspectives for epidemiological research: I. regression analysis. Int J Epidemiol 36:195–202

Greenland S (2008) Introduction to Bayesian statistics. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. LWW, Philadelphia, pp 328–344

Greenland S, Rothman KJ (2008) Introduction to categorical statistics. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. LWW, Philadelphia, pp 238–257

Howard JV (1998) The 2 × 2 table: a discussion from the Bayesian viewpoint. Stat Sci 4:351–367

Liebermeister C (1877) Über Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik. Sammlung klinischer Vorträge (Innere Medicin No. 31–64) 110:935–962

Mossman D, Berger JO (2001) Intervals for posttest probabilities: a comparison of 5 methods. Med Decis Making 21:498–507

Nurminen M, Mutanen P (1987) Exact Bayesian analysis of two proportions. Scand J Stat 14: 67–77

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) Uncertain judgements; eliciting experts' probabilities. Wiley, Chichester

Pawitan Y (2001) In all likelihood: statistical modelling and inference using likelihood. Oxford University Press, New York

Rothman KJ (2002) Epidemiology; an introduction. Oxford University Press, New York

Rue H, Held L (2005) Gaussian Markov random fields: theory and applications. Chapman and Hall, Boca Raton

Savitz DA, Wachtel H, Barnes FA, John EM, Tvrdik JG (1988) Case-control study of childhood cancer and exposure to 60-Hz magnetic fields. Am J Epidemiol 128:21–38

Sellke T, Bayarri MJ, Berger JO (2001) Calibration of p values for testing precise null hypotheses. Am Stat 55:62–71

Seneta E (1994) Carl Liebermeister's hypergeometric tails. Hist Math 21:453–462

Seneta E, Phipps MC (2001) On the comparison of two observed frequencies. Biom J 43:23–43

Spiegelhalter DJ, Abrams KR, Myles JP (2004) Bayesian approaches to clinical trials and health-care evaluation. Wiley, New York