Chapter 14
Bayesian Inference

14.1 Frequentist vs Bayesian

In the frequentist approach to parametric statistical inference:

1. Probability models are based on the relative frequency interpretation of probabilities.
2. Parameters of the resulting probability models are assumed to be fixed, unknown constants.
3. Observations on random variables with a probability model depending on the parameters are used to construct statistics. These are used to make inferential statements about the parameters.
4. Inferences are evaluated and interpreted on the basis of the sampling distribution of the statistics used for the inference. Thus an interval which claims to be a 95 % confidence interval for θ has the property that it contains θ 95 % of the time in repeated use.
5. In all cases inferences are evaluated on the basis of **data not observed**.

Bayesian statistics, on the other hand:

1. Allows probabilities to be degrees of belief.
2. Probability statements can be made about parameters and represent degrees of belief about a parameter.
3. From observations with a given probability model and a prior distribution we determine the probability distribution of the parameter given the observed data using Bayes theorem.
4. Any inferential statements are then based on this distribution.
5. Since inferences depend on the prior the degrees of belief and hence inferences can differ for a given set of observations.

C.A. Rohde, *Introductory Statistical Inference with the Likelihood Function*, DOI 10.1007/978-3-319-10461-4_14

14.2 The Bayesian Model for Inference

The basic parametric statistical model is

$$(\mathcal{X}, f(x;\theta), \Theta)$$

We observe X which has sample space $\mathcal{X}$. The probability density for $X = x$ is $f(x;\theta)$; θ is a parameter(s) having values in the parameter space Θ.

In the Bayesian approach

1. $f(x;\theta)$ is interpreted as the conditional probability density of x given θ.
2. We interpret $f(x;\theta)$ as $f(x|\theta)$ implicitly replacing the ; by a | indicating conditioning on θ.
3. A **prior** density, $p(\theta)$, of θ which describes our beliefs about θ before the data is observed, is assumed.
4. Bayes theorem is then used to obtain the **posterior** distribution, $\mathbf{P}(\theta|x)$ of θ given the data x, i.e.,

$$\mathbf{P}(\theta|x) = \frac{f(x;\theta)p(\theta)}{f(x)}$$

1. Where

$$f(x) = \int_{\Theta} f(x;\theta)p(\theta)\mu(d\theta)$$

is a normalizing constant (the marginal distribution of x).

The posterior result may also be written as

$$\mathbf{P}(\theta|x) \quad \propto \quad \mathscr{L}(\theta;x)g(\theta)$$

where $\mathscr{L}(\theta;x)$ is the likelihood of θ having observed x. The posterior represents what we believe about θ after we have observed the data. It represents an updating of our beliefs about θ having observed the data.

14.3 Why Bayesian? Exchangeability

If $X_1, X_2, \ldots, X_n$ are independent and identically distributed (iid) then for any $(x_1, x_2, \ldots, x_n)$ the joint density $f(x_1, x_2, \ldots, x_n)$ can be written as

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$$

where $f(x)$ is the density function of any X at x.

A **permutation** of $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, $\pi(\mathbf{x})$, is any rearrangement of $\mathbf{x}$. For example, $(x_{i_1}, x_{i_2}, \ldots, x_{i_n})$. There are $n!$ such permutations.

Definition 14.3.1. If

$$f(\pi(\mathbf{x})) = f(\mathbf{x})$$

for every permutation and every $\mathbf{x}$ the random variables $X_1, X_2, \ldots, X_n$ are said to be **exchangeable**.

Clearly if $X_1, X_2, \ldots, X_n$ are iid then they are exchangeable.

Theorem 14.3.1. *If $X_1, X_2, \ldots, X_n$ are iid given θ then they are exchangeable*

Proof. For any $x_1, x_2, \ldots, x_n$ we have

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= \int_\Theta f(x_1, x_2, \ldots, x_n|\theta) g(\theta) d\mu(\theta) \\ &= \int_\Theta \prod_{i=1}^n f(x_i|\theta) g(\theta) d\mu(\theta) \\ &= \int_\Theta \prod_{j=1}^n f(x_{i_j}|\theta) g(\theta) d\mu(\theta) \\ &= \int_\Theta f(x_{i_1}, x_{i_2}, \ldots, x_{i_n}|\theta) g(\theta) d\mu(\theta) \\ &= f(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) \end{aligned}$$

A famous theorem due to de Finetti provides a partial converse to the fact that conditionally iid random variables are exchangeable. □

Theorem 14.3.2 (de Finetti). *If $X_1, X_2, \ldots$ is a sequence of random variables which are exchangeable for every n then there is a distribution g such that*

$$f(x_1, x_2, \ldots, x_n) = \int_\Theta \prod_{i=1}^n f(x_i|\theta) g(\theta) d\mu(\theta)$$

i.e., $X_1, X_2, \ldots, X_n$ can be viewed as conditionally independent given θ where θ has distribution defined by g

Jose Bernardo, a leading proponent of the use of Bayesian statistics, states:

> It is important to realize that if the observations are conditionally independent, -as it is implicitly assumed when they are considered to be a random sample from some model-, then they are necessarily exchangeable. The representation theorem, -a pure probability theory result- proves that if observations are judged to be *exchangeable*, then they must indeed be a random sample from some model *and* there *must exist* a prior probability distribution over the parameter of the model, hence requiring a *Bayesian* approach.

> Note however that the representation theorem is an existence theorem: it generally does not specify the model, and it never specifies the required prior distribution. The additional assumptions which are usually necessary to specify a particular model are described in particular representation theorems. An additional effort is necessary to assess a prior distribution for the parameter of the model.

The key point is that exchangeability implies the existence of a prior and provides a powerful justification for the use of Bayesian methods to describe beliefs.

Other justifications for the use of Bayesian methods are based on the concept of utilities and decision making and rely on the concept of coherence:

1. One important point in using Bayesian methods is that the choice of prior need not reflect true prior knowledge about the parameter.
2. This is the basis for the **objective Bayes** approach. The prior in this case represents that function of the parameter which has minimal impact on the posterior.
3. It need not be a proper probability distribution, but the posterior is required to be a proper probability distribution.

In fact some Bayesians are even more forthright:

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

[31]

Let X be binomial with parameters n and θ where we assume that n is known. That is

$$f(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

for $x = 0, 1, \ldots, n$. Suppose we represent prior information by a distribution of the form

$$g(\theta;\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathbf{B}(\alpha,\beta)}$$

where α and β are both positive.

This prior is a Beta distribution with parameters α and β and

$$\mathbf{B}(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

It is known that $\mathbf{B}(\alpha,\beta)$ is given by

$$\mathbf{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

where

$$\Gamma(\delta) = \int_0^\infty t^{\delta-1} e^{-t} dt$$

for $\delta > 0$ is the Gamma function.

With this choice of prior we have that

$$\begin{aligned} f(x) &= \int_\Theta f(x;\theta) g(\theta;\alpha,\beta) d\theta \\ &= \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathbf{B}(\alpha,\beta)} d\theta \\ &= \frac{\binom{n}{x}}{\mathbf{B}(\alpha,\beta)} \int_0^1 \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} d\theta \\ &= \frac{\binom{n}{x}\mathbf{B}(x+\alpha-1, n-x+\beta-1)}{\mathbf{B}(\alpha,\beta)} \end{aligned}$$

Thus the posterior density of θ is given by

$$g(\theta|x) = \frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\mathbf{B}(x+\alpha, n-x+\beta)}$$

i.e., a beta distribution with parameters $x+\alpha$ and $n-x+\beta$. Note that the posterior distribution depends on the parameters α and β

It is known that the Beta distribution with parameters $\alpha^{'}$ and $\beta^{'}$ has expected value given by

$$\frac{\alpha^{'}}{(\alpha^{'}+\beta^{'})}$$

Thus the expected value of the posterior is given by

$$\frac{(x+\alpha)}{(n+\alpha+\beta)}$$

which is a natural Bayes estimate of θ,

Note that this estimate is not the same as the conventional estimate x/n.

In fact

$$\frac{x+\alpha}{n+\alpha+\beta} = (1-w_n)\frac{x}{n} + w_n \frac{\alpha}{\alpha+\beta}$$

where

$$w_n = \frac{\alpha + \beta}{n + \alpha + \beta}$$

i.e., the posterior mean is a weighted combination of the prior mean and the usual estimate.

This may also be written as

$$\frac{x + \alpha}{n + \alpha + \beta} = \frac{x}{n} - w_n \left(\frac{x}{n} - \frac{\alpha}{\alpha + \beta} \right)$$

which shows that the usual estimate is "shrunk" toward the prior mean. Note that the shrinkage factor, w_n, approaches 0 for large sample sizes.

Suppose that we have observed $x_1, x_2, \ldots, x_n$ assumed to be realized values of independent random variables which are Poisson with parameter λ. Then the density given λ is

$$f(x_1, x_2, \ldots, x_n; \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{n\overline{x}} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

Assume a prior for λ of the form

$$p(\lambda) = \frac{\lambda^{\alpha - 1} e^{-\lambda/\beta}}{\Gamma(\alpha) \beta^{\alpha}}$$

i.e., a Gamma distribution with parameters $\alpha > 0$ and $\beta > 0$

Then the posterior of λ is proportional to

$$\lambda^{n\overline{x} + \alpha - 1} e^{-\lambda(n + 1/\beta)}$$

and hence the posterior is Gamma with parameters

$$a = n\overline{x} + \alpha \quad \text{and} \quad b = \frac{1}{n + \frac{1}{\beta}}$$

One natural (Bayes) estimate of λ is the mean of the posterior given by

$$ab = \frac{n\overline{x} + \alpha}{n + \frac{1}{\beta}}$$

Note that the posterior mean can be written as

$$(1 - w_n)\overline{x} + w_n \alpha \beta$$

where

$$w_n = \frac{1}{1 + n\beta}$$

Thus the posterior mean is a linear combination of the prior mean and the maximum likelihood estimate. This can also be written as

$$\overline{x} - w_n \left(\overline{x} - \alpha\beta \right)$$

which again shows the shrinkage of the usual estimate to the prior mean. Again note that for large n the posterior mean is very nearly equal to the maximum likelihood estimate.

Another estimate is the mode of the posterior which, in this case, is given by

$$\frac{\overline{x} + \frac{\alpha - 1}{n}}{1 + \frac{1}{n\beta}}$$

For large n this is very nearly equal to the maximum likelihood estimate (this result is true quite generally).

Suppose that $\mathbf{Y}$ obeys a general linear model, i.e., $\mathbf{Y}$ is normal with

$$\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \text{ and } \mathbb{V}(\mathbf{Y}) = \mathbf{I}\sigma^2$$

Suppose further that the prior distribution for $\boldsymbol{\beta}$ is also normal with mean and variance-covariance matrix given by

$$\mathbb{E}(\boldsymbol{\beta}) = \boldsymbol{\beta}_0 \text{ and } \mathbb{V}(\boldsymbol{\beta}) = \mathbf{V}$$

The joint distribution of $\mathbf{Y}$ and $\boldsymbol{\beta}$ is thus also normal with

$$\mathbb{E}\left(\begin{bmatrix} \mathbf{Y} \\ \boldsymbol{\beta} \end{bmatrix} \right) = \begin{bmatrix} \mathbf{X} \\ \mathbf{I} \end{bmatrix} \boldsymbol{\beta}$$

and

$$\mathbb{V}\left(\begin{bmatrix} \mathbf{Y} \\ \boldsymbol{\beta} \end{bmatrix} \right) = \begin{bmatrix} \mathbf{I}\sigma^2 + \mathbf{X}\mathbf{V}\mathbf{X}^\top & \mathbf{X}\mathbf{V} \\ \mathbf{V}\mathbf{X}^\top & \mathbf{V} \end{bmatrix}$$

It then follows that the posterior distribution of $\boldsymbol{\beta}$ given $\mathbf{Y} = \mathbf{y}$ is normal with

$$\mathbb{E}(\boldsymbol{\beta}) = \boldsymbol{\beta}_0 + \mathbf{V}\mathbf{X}^\top \left[\mathbf{I}\sigma^2 + \mathbf{X}\mathbf{V}\mathbf{X}^\top \right]^{-1} \mathbf{X}\mathbf{V}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)$$

and

$$\mathbb{V}(\boldsymbol{\beta}) = \left[\mathbf{I} + \mathbf{X}\mathbf{V}\mathbf{X}^\top\right]^{-1}$$

Lots of complex, but routine, matrix algebra shows that the mean of the posterior distribution of $\boldsymbol{\beta}$ can be written as

$$\left[(\mathbf{X}^\top\mathbf{X})\sigma^{-2} + \mathbf{V}^{-1}\right]^{-1}\left\{\sigma^{-2}\mathbf{X}^\top\mathbf{X}\mathbf{b} + \mathbf{V}^{-1}\boldsymbol{\beta}_0\right\}$$

where $\mathbf{b}$ is the least squares estimate

$$\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

i.e., the mean of the posterior is a weighted linear combination of the prior mean $\boldsymbol{\beta}_0$ and the maximum likelihood estimate $\mathbf{b}$.

Examples 1–3 have priors which are **conjugate**.

If a prior and a posterior belong to the same family of distributions they are said to be conjugate.

14.4 Stable Estimation

Intuitively, with large amounts of data, the impact of the prior should be small.

More formally, if the likelihood $\mathscr{L}(\theta|x)$ is highly concentrated over a region $\Theta_s \subset \Theta$, then the posterior will satisfy

$$\pi(\theta|x) \approx \frac{\mathscr{L}(\theta; x)}{\int_{\Theta_s} \mathscr{L}(\theta; x) d\mu(\theta)} \quad \theta \in \Theta_s$$

Thus the prior has essentially no impact on the posterior and we have robustness to prior misspecification.

This is called **stable estimation**.

14.5 Bayesian Consistency

If the posterior converges to a distribution which is concentrated at the true value of the parameter θ_0 we have Bayesian consistency.

Under weak regularity conditions most commonly used models with sensible priors lead to a posterior which is consistent.

14.6 Relation to Maximum Likelihood

1. Suppose that $X_1, X_2, \ldots, X_n$ are iid with pdf $f(x;\theta)$ where

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$$

2. Assume that the prior $\pi(\boldsymbol{\theta})$ and $f(\mathbf{x};\boldsymbol{\theta})$ are positive and twice differentiable near the maximum likelihood estimate of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$.
3. Then under suitable regularity conditions (similar to those for maximum likelihood estimation) we have the Bernstein-von Mises result as stated in Berger (1987).

   The posterior density of $\boldsymbol{\theta}$

$$\pi_n(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x};\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\mathbf{x})}$$

can be approximated for large n in the following four ways:

(i) π_n is approximately MVN $(\boldsymbol{\mu}^*(\mathbf{x}), \mathbb{V}(\mathbf{x}))$ where $\boldsymbol{\mu}^*(\mathbf{x})$ and $\mathbb{V}^*(\mathbf{x})$ are the posterior mean and posterior covariance matrix.

(ii) π_n is approximately MVN $\left(\widehat{\boldsymbol{\theta}}^*; [\mathcal{I}^\pi(\mathbf{x})]^{-1}\right)$ where $\widehat{\boldsymbol{\theta}}^*$ is the generalized maximum likelihood estimate for $\boldsymbol{\theta}$, i.e., the maximum likelihood estimator for $\boldsymbol{\theta}$ is the model with likelihood

$$\mathscr{L}^*(\boldsymbol{\theta}, \mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

and $\mathcal{I}^\pi(\mathbf{x})$ is the $p \times p$ matrix with (i,j) element given by

$$\mathcal{I}^\pi_{ij}(\mathbf{x}) = -\left[\frac{\partial^2 \ln[f(\mathbf{x};\boldsymbol{\theta})\pi(\boldsymbol{\theta})]}{\partial\theta_i \partial\theta_j}\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^*}$$

(iii) π_n is approximately MVN $\left(\widehat{\boldsymbol{\theta}}; [\widehat{\mathcal{I}}(\mathbf{x})]^{-1}\right)$ where $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimate for $\boldsymbol{\theta}$ and $\widehat{\mathcal{I}}(\mathbf{x})$ is the $p \times p$ observed Fisher information matrix with (i,j) element given by

$$\widehat{\mathcal{I}}_{ij}(\mathbf{x}) = -\left[\frac{\partial^2 \ln[f(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\sum_{k=1}^{n}\left[\frac{\partial^2 \ln[f(x_k;\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

(iv) π_n is approximately MVN $\left(\widehat{\boldsymbol{\theta}}; [\mathcal{I}(\widehat{\boldsymbol{\theta}})]^{-1}\right)$ where $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimate for $\boldsymbol{\theta}$ and $\mathcal{I}(\boldsymbol{\theta})$ is the $p \times p$ expected Fisher information matrix with (i,j) element given by

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = -n\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \ln[f(X_1;\boldsymbol{\theta})]}{\partial\theta_i\partial\theta_j}\right]$$

In general the approximations are ordered (i)–(iv) with (i) better than (ii), (ii) better than (iii), and (iii) better than (iv).

If the prior is "uninformative" then the result gives "objective" posterior approximations.

It follows that large sample posterior intervals are approximately

$$\widehat{\theta}_n \pm z_{1-\alpha/2}\sqrt{\frac{1}{n\, i(\widehat{\theta})}}$$

It also follows that large sample posterior intervals for $g(\theta)$ are given by

$$g(\widehat{\theta}_n) \pm z_{1-\alpha/2}\sqrt{\frac{|g^{(1)}(\widehat{\theta}_n)|}{n\, i(\widehat{\theta})}}$$

Remember Efron's statement that estimate plus or minus two standard errors has good credentials in any theory of statistical inference.

14.7 Priors

Many scientists and statisticians often say that they would use Bayesian methods if they could find or justify use of a particular prior. What they are arguing about is the choice of prior not the basic methodology, i.e., there is an acceptance of the treatment of parameters as random. Exchangeability ensures that there is a prior.

Much has been written on the choice of priors and much more will surely be written.

Example. Consider a room which we are told is square and between 10 and 20 feet on a side. If θ_1 is the parameter representing the length of a side, pleading ignorance leads to a prior of the form

$$p(\theta_1) = \begin{cases} \frac{1}{10} \text{ if } 10 \le \theta_1 \le 20 \\ 0 \qquad\qquad \text{elsewhere} \end{cases}$$

If θ_2 is the parameter representing the area of the room then again pleading ignorance leads to a prior of the form

$$p(\theta_2) = \begin{cases} \frac{1}{400} \text{ if } 100 \le \theta_2 \le 400 \\ 0 \qquad\qquad \text{elsewhere} \end{cases}$$

Note that the probability that the length of the side is between 10 and 15 feet is $\frac{1}{2}$ which corresponds to an area between 100 and 225 square feet.

The probability assigned to this area under the ignorance model for the area is

$$\int_{100}^{225} \frac{1}{400} d\theta_2 = \frac{\theta_2}{400}\Bigg|_{100}^{225} = \frac{225-100}{400} = \frac{125}{400} = \frac{5}{8}$$

Thus, the two ignorance assignments are not compatible.

Reference: Bayesianism: Its Scope and Limits, Elliott Sobel

Many view the above example as a key counterexample to the use of Bayesian statistics. If it were, interest in Bayesian statistics would have waned decades ago.

What has happened is the search for priors which are **transformation invariant** and yet, in some sense, do not convey "much" prior information, i.e., the prior is dominated by the likelihood, even for small samples.

14.7.1 Different Types of Priors

Basically there are four approaches:

1. A formal mathematical approach which uses **conjugate priors**
2. An ad hoc approach which uses **vague**, **flat**, or **uniform priors** to represent "ignorance"
3. A formal approach using **reference priors** which are designed to "let the data speak for themselves"
4. A formal approach which elicits information to determine a truly **subjective prior**

14.7.1.1 Conjugate Priors

If, in a given problem, there is a prior which, when combined with the likelihood, yields a posterior which is in the same family as the prior, then the prior is said to be a **conjugate prior**. It is tacitly assumed that a case can be made that this prior represents prior beliefs about the parameter.

Conjugate priors have the great advantage that closed forms can be obtained for the posterior and hence inferences are computationally simple.

Conjugate priors are not often available, but they are in one special family of distributions called **the exponential family**.

For example:

(i) If the likelihood is binomial then the beta distribution is a conjugate prior.
(ii) If the likelihood is Poisson then the Gamma distribution is a conjugate prior.
(iii) If the likelihood is normal then the normal is a conjugate prior.

Thus many of the simplest inference problems have conjugate priors and hence closed form expressions for the posterior distributions can be found.

More generally linear combinations of conjugate priors can be used to find priors. There are compendiums of conjugate priors available, e.g, Fink [15]

14.7.2 Vague Priors

Vague priors, also called flat or non-informative priors, are priors which are such that they are constant over the range of parameter values for which the likelihood is moderate in size.

Thus the posterior is essentially the likelihood normalized so as to integrate or sum to 1.

Whenever the prior is of this type it is usually not a density. There is, for example, no constant density that integrates to 1 over the interval $[0, \infty)$.

Thus one always needs to check that the posterior is in fact a density function when flat priors are used. Such a posterior is called **proper**.

Vague priors are supposed to represent **ignorance**, i.e., any parameter value is considered to be equally likely a priori. However this means that while we are ignorant about θ we are not ignorant about $g(\theta)$ since its density will not be uniform. (The Jacobian of the transformation form θ to $g(\theta)$ is not, in general, a constant.)

Thus vague priors are not **transformation invariant**. Reread the slides on the problem of ignorance.

14.7.2.1 Jeffrey's Priors

Jeffrey's priors are a class of default priors which are translation invariant.

These priors, when applicable, choose the prior

$$p(\theta) \propto \sqrt{i(\theta)}$$

where $i(\theta)$ is Fisher's information. For the multiple parameter case these priors are of the form

$$p(\boldsymbol{\theta}) \propto \sqrt{\det(i(\boldsymbol{\theta})}$$

i.e., to the determinant of Fisher's information matrix for a sample of size 1.

Example. For the Bernoulli we know that Fisher's information is $i(\theta) = 1/[\theta(1-\theta)]$ so that the Jeffrey's prior is given by

$$i(\theta) = \frac{1}{\theta^{1/2}(1-\theta)^{1/2}}$$

and hence the posterior for a binomial using Jeffrey's prior is given by

$$p(\theta|x) \propto \theta^{x-\frac{1}{2}}(1-\theta)^{n-x-\frac{1}{2}}$$

i.e., a beta distribution with parameter $a = x + 1/2$ and $b = n - x + 1/2$.

Example. For the normal with known variance, Jeffrey's prior is given by

$$p(\theta) = 1$$

which is an improper prior. For the normal with unknown mean and variance Jeffrey's prior is given by

$$g(\theta, \sigma^2) = \frac{1}{\sigma}$$

which is also an improper prior. In both cases the posterior is a proper prior.

Jeffrey's priors are not flat, but they are transformation invariant and are widely used.

14.7.2.2 Reference Priors

In the last three decades much work has been done on developing a class of prior distributions, called **reference priors**, which "let the data speak for themselves." Essentially these priors maximize the distance between the prior and the posterior, using distance specified by the Kullback-Leibler divergence.

> "Intuitively, a reference prior for θ is one which maximizes what is **not known** about θ, **relative** to what could possibly be learnt from the result of a particular experiment. More formally a reference prior for θ is defined to be one which maximizes, within some class of candidate priors, the **missing information** about the quantity of interest θ, defined as a limiting form of the amount of information about its value which data from the assumed model could possibly provide."

The amount of missing information is defined in terms of the Kullback-Leibler divergence. Determination of these priors is quite technical, but often they turn out to be Jeffrey's prior.

> "Reference priors are not descriptions of personal beliefs; they are proposed as formal **consensus** priors to be used as standards for scientific communication."

The quotations above are from papers by Bernardo. His website has excellent papers on objective Bayes procedures.

Here is a short summary of reference priors for common statistical problems:

Likelihood	Prior	Posterior
Binomial	Beta(1/2,1/2)	Beta(x+1/2,n-x+1/2)
Poisson	$\lambda^{-1/2}$	Gamma(x+1/2,1)
Normal (known σ^2)	Constant	Normal($\overline{x}, \sigma^2/n$)
Normal (unknown σ^2)	σ^{-1}	Student'st

A complete list of reference priors appears in [56].

14.7.2.3 Subjective Priors

One important method of obtaining priors is to convene a group of experts in the field under study and have them assess a prior distribution. Much has been written about this under the name **prior elicitation**.