MAS3301 Bayesian Statistics

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 2, 2008-9

19 Odds and Bayes Factors

19.1 Hypotheses

Hypothesis: A *hypothesis* is a proposition (statement) which may or may not be true.

Simple hypothesis: If a hypothesis specifies everything about a model, including all parameter values, it is called a *simple hypothesis.* E.g. "$X \sim$ binomial(10, 0.5)".

Composite hypothesis: If a hypothesis leaves some parameter values unknown, it is called a *composite hypothesis.* E.g. "$X \sim$ binomial(10, θ) where θ is unknown".

19.2 Comparing two simple hypotheses

This is the simplest case but it is relatively rare.

The prior probabilities for the two hypotheses are π_1, π_2.

The likelihoods are L_1, L_2.

The posterior probabilities are then

$$p_1 = \frac{\pi_1 L_1}{\pi_1 L_1 + \pi_2 L_2}, \qquad p_2 = \frac{\pi_2 L_2}{\pi_1 L_1 + \pi_2 L_2}.$$

The prior odds in favour of hypothesis 1 are π_1/π_2.

The posterior odds in favour of hypothesis 1 are

$$\frac{p_1}{p_2} = \frac{\pi_1 L_1}{\pi_2 L_2}.$$

Notice that **posterior odds = prior odds $\times$ likelihood ratio**.

The likelihood ratio (also called the *Bayes factor),* is L_1/L_2, and does not depend on the prior probabilities. It is therefore an objective measure of the weight of evidence in favour of hypothesis 1.

If we are choosing between k simple hypotheses with prior probabilities $\pi_1, \ldots, \pi_k$ and likelihoods $L_1, \ldots, L_k$ then we can easily work out $\tilde{p}_i = \pi_i L_i$ and then

$$p_i = \frac{\tilde{p}_i}{\sum_{j=1}^{k} \tilde{p}_j}.$$

19.3 Examples

19.3.1 Example 1

In the "animals" example in Section 3.2 we really had two simple hypotheses, $H_1: \theta = 0.1$, with prior probability $\pi_1 = 2/3$, and $H_2: \theta = 0.4$, with prior probability $\pi_2 = 1/3$. We observe twenty animals and count the number Y with a particular gene. Under H_1 we have $Y \sim$ bin(20, 0.1) and, under H_2 we have $Y \sim$ bin(20, 0.4). We observe that $y = 3$ animals have the gene. So we find

$$L_j = \binom{20}{3} \theta_j^3 (1-\theta_j)^{17}$$

where $\theta_1 = 0.1$ and $\theta_2 = 0.4$. Hence the likelihood ratio in favour of H_1 is

$$\frac{L_1}{L_2} = \frac{0.1^3 0.9^{17}}{0.4^3 0.6^{17}} = \frac{1.667718 \times 10^{-4}}{1.083306 \times 10^{-5}} = 15.39471.$$

The prior odds in favour of H_1 are $\pi_1/\pi_2 = 2.0$. So the posterior odds are $2.0 \times 15.39471 = 30.78941$. Hence the posterior probability of H_1 is $30.78941/(1 + 30.78941) = 0.969$.

19.3.2 Example 2

A person A is suspected of having committed a crime. As a result of forensic science work, evidence E is found. Let H_1 be the hypothesis that A did commit the crime and H_2 be the hypothesis that A did not commit the crime. Under H_1 the probability of the evidence E being found is L_1 and, under H_2, it is L_2. For example, suppose that $L_1 = 0.8$ and $L_2 = 0.0002$. Then the likelihood ratio in favour of H_1 is $0.8/0.0002 = 4000$. This appears to be strong evidence in favour of the guilt of A. However our posterior probability for H_1 also depends on our prior probabilities. There may be many people who could have committed the crime and, in the absence of any other evidence, we should perhaps assign equal prior probabilities to them. Suppose that 10000 other people, apart from A, could have committed the crime. Then the prior odds in favour of H_1 are $1/10000 = 0.0001$. The posterior odds are therefore $4000/10000 = 0.4$ and the posterior probability of H_1 is $0.4/1.4 = 0.286$. This may seem surprisingly small but it illustrates the importance of not confusing the probability of the evidence given that the suspect is innocent, which is very small, with the probability that the suspect is innocent given the evidence, which, in this case, is much larger.

How large does π_1 have to be so that the posterior probability $p_1 > 0.5$? We require $p_1/p_2 > 1$ so $\pi_1 L_1 > \pi_2 L_2$ and $\pi_1/\pi_2 > L_2/L_1 = 1/4000$, the reciprocal of the likelihood ratio. So we require

$$\pi_1 > \frac{L_2/L_1}{1 + L_2/L_1} = \frac{L_2}{L_1 + L_2} = \frac{0.0002}{0.8002} = 0.0002499.$$

19.4 Composite hypotheses

In a composite hypothesis there are parameters with values which are not specified by the hypothesis.

Example: We are interested in the mean of a normal distribution and have hypotheses about its value but we do not specify the precision (or variance).

- Model: Given the parameters, $Y_1, \ldots, Y_n$ are normal with mean μ and precision τ.
- Hypotheses: $H_1 : \ \mu < \mu_0$ and $H_2 : \ \mu \geq \mu_0$. These are both composite hypotheses both because we do not specify completely the value of μ and because we do not specify the value of τ.

In general:

Hypotheses: $H_1, \ H_2$, with prior probabilities $\pi_1, \ \pi_2$ and unspecified parameters θ.

Conditional prior density $f_j^{(0)}(\theta)$ for θ given hypothesis H_j.

Marginal likelihood: The condititional probability (density) of observing data $\underline{y}$ given hypothesis j and the value of θ gives us the likelihood $L_j(\theta; \underline{y})$. The probability (density) of observing data $\underline{y}$ given hypothesis j is therefore

$$\int f_j^{(0)}(\theta) L_j(\theta; \underline{y}) \ d\theta.$$

This is called the *marginal likelihood* of hypothesis H_j. Notice that it is the prior predictive density evaluated at the observed value of the data, $\underline{y}$.

The posterior odds in favour of H_1 are therefore

$$\frac{p_1}{p_2} = \frac{\pi_1}{\pi_2} \times \frac{\int f_1^{(0)}(\theta) L_1(\theta; \underline{y}) \ d\theta}{\int f_2^{(0)}(\theta) L_2(\theta; \underline{y}) \ d\theta}.$$

The Bayes factor is the ratio of the marginal likelihoods. The Bayes factor is also equal to

$$\frac{p_1/p_2}{\pi_1/\pi_2}.$$

Notice that it now depends on the priors, $f_1^{(0)}, \ f_2^{(0)}$, as well as the likelihoods so it is not "objective." However, in some cases, the priors have little effect on the Bayes factor.

19.5 Examples

19.5.1 Example 1

In an industrial quality control application we are interested in the precision of a particular dimension of amnufactured components. Twenty components are measured and the difference between the actual measurement and the nominal value is recorded in each case. The values are given in μm, where $1\mu m$ is $10^{-6}m$).

Our model for these values, y_i, is that, given the values of parameters $\mu,\ \tau$,

$$y_i \sim N(\mu,\ \tau^{-1})$$

and y_i is independent of y_j for $i \neq j$.

Our prior distribution for $\mu,\ \tau$ is as follows.

The distribution of τ is

$$\tau \sim \text{gamma}(d_0/2,\ d_0 v_0/2)$$

that is

$$d_0 v_0 \tau \sim \chi^2_{d_0}$$

where $d_0 = 6$ and $v_0 = 2500$.

The conditional distribution of $\mu \mid \tau$ is

$$\mu \mid \tau \sim N(m_0,\ [c_0\tau]^{-1})$$

where $m_0 = 0$ and $c_0 = 0.5$.

We have two hypotheses:

- $H_A:\ \ \tau < 0.0004$
- $H_B:\ \ \tau \geq 0.0004$

The data are as follows.

163	51	87	70	-31	-85	30	37	-26	65
-3	81	-50	-3	-21	64	92	-26	71	72

Find the Bayes factor in favour of H_A.

Solution:

We have the standard conjugate prior.
In the prior:

$$\begin{aligned} \Pr(\tau < 0.0004) &= \Pr(d_0 v_0 \tau < d_0 v_0 \times 0.0004) \\ &= \Pr(d_0 v_0 \tau < 6) \\ &= 0.57681 \end{aligned}$$

since $d_0 v_0 \tau \sim \chi^2_6$.

(E.g. use `pchisq(6,6)` in R).

Hence the prior odds are

$$\frac{0.57681}{1 - 0.57681}.$$

In the posterior:

$$d_1 v_1 \tau \sim \chi^2_{d_1}$$

where

$$\begin{aligned} d_1 &= d_0 + n = 26 \\ s_n^2 &= \frac{1}{n}\sum (y_i - \bar{y})^2 = \frac{1}{n}\left\{\sum y_i^2 - n\bar{y}^2\right\} = 3481.19 \\ r^2 &= \frac{1}{n}\sum (y_i - m_0)^2 = (\bar{y} - m_0)^2 + s_n^2 = 4498.8 \\ v_d &= \frac{c_0 r^2 + n s_n^2}{c_0 + n} = 3506.01 \\ v_1 &= \frac{d_0 v_0 + n v_d}{d_0 + n} = 3273.85 \end{aligned}$$

Hence

$$\begin{aligned} \Pr(\tau < 0.0004) &= \Pr(d_1 v_1 \tau < d_1 v_1 \times 0.0004) \\ &= \Pr(d_1 v_1 \tau < 34.048) \\ &= 0.86618 \end{aligned}$$

since $d_1 v_1 \tau \sim \chi^2_{26}$.

(E.g. use `pchisq(34.048,26)` in R).

Hence the posterior odds are

$$\frac{0.86618}{1 - 0.86618}.$$

So the Bayes factor is

$$\frac{0.86618}{(1 - 0.86618)} \frac{(1 - 0.57681)}{0.57681} = \underline{4.749}.$$

19.5.2 Sufficiency

Suppose that we wish to find a Bayes factor to compare two hypotheses concerning a, possibly vector, parameter θ. In some cases we will have available a sufficient statistic T for θ. Then we can

write the conditional density of the data $\underline{Y}$ given θ as

$$f_{Y|\theta}(\underline{y} \mid \theta) = f_{T|\theta}(t \mid \theta) f_{Y|t}(\underline{y} \mid t).$$

(See Section 12).

In such a case, when we find the two marginal likelihoods, $f_{Y|t}(\underline{y} \mid t)$ will be left as a common factor in both, since it does not involve θ. Therefore it will cancel from the Bayes factor. It follows that we can find the Bayes factor by considering the predictive distributions of the sufficient statistic T.

19.5.3 Example 2: The mean of a normal distribution

Model: Sample $Y_1, \ldots, Y_n$ from a normal sampling distribution, $N(\mu,\ \tau)$, where the observations are independent given the parameters. Suppose that τ is known.

Hypotheses:
- $H_0:\ \ \mu = m^*$, where m^* is a specified value,
- $H_1:\ \ \mu \neq m^*$.

Prior: Under H_1, we have a normal prior distribution for μ with mean m_0 and precision p_0. For convenience we write $p_0 = c_0\tau$.

We know that the sample mean $\bar{Y} = \sum_{i=1}^{n} Y_i/n$ is sufficient for μ so we find its predictive distributions under the two hypotheses.

The conditional distribution of $\bar{Y}$ given μ is $N(\mu,\ [n\tau]^{-1})$.

We can represent the joint distribution of μ and $\bar{Y}$ in terms of μ and D where $D = \bar{Y} - \mu \sim N(0,\ [n\tau]^{-1})$ and D is independent of μ. It is easy to check that this gives the correct means, variances and covariances.

Therefore, under H_1,

$$\bar{Y} = \mu + D \sim N(m_0,\ [c_0\tau]^{-1} + [n\tau]^{-1}) = N(m_0,\ [nc_0\tau/c_1]^{-1})$$

where $c_1 = c_0 + n$. Under H_0, $\bar{Y} \sim N(m^*, [n\tau]^{-1})$ since there is no contribution to the variance from μ.

So the marginal likelihood under H_1 is

$$\bar{L}_1 = (2\pi)^{-1/2}[nc_0\tau/c_1]^{1/2} \exp\left\{-\frac{nc_0\tau}{2c_1}(\bar{y} - m_0)^2\right\}.$$

The likelihood under H_0 is

$$L_0 = (2\pi)^{-1/2}[n\tau]^{1/2} \exp\left\{-\frac{n\tau}{2}(\bar{y} - m^*)^2\right\}.$$

Hence the Bayes factor in favour of H_1 is

$$K = \frac{\bar{L}_1}{L_0} = \left(\frac{c_0}{c_1}\right)^{1/2} \exp\left\{-\frac{n\tau}{2}\left[\frac{c_0}{c_1}(\bar{y} - m_0)^2 - (\bar{y} - m^*)^2\right]\right\}.$$

If $m_0 = m^*$, which would often be the case, then the Bayes factor simplifies to

$$K = \left(\frac{c_0}{c_1}\right)^{1/2} \exp\left\{\frac{n^2\tau}{2c_1}(\bar{y} - m^*)^2\right\}.$$

Example: Suppose that we give twenty nine-year-old children from a population of interest a reading test. The standard score in this test, for a nine-year-old, is 100. The standard deviation of scores is known to be 10 so $\tau = 10^{-2} = 0.01$. We wish to look at the evidence for and against the hypothesis H_0 that the mean for children in this population is $\mu = m^* = 100$, compared to the more general hypothesis H_1. Our conditional prior distribution for μ under H_1 is $N(m_0,\ p_0^{-1})$ where $m_0 = m^* = 100$ and $p_0^{-1} = 20^2 = 400$ so $p_0 = c_0\tau = 0.0025$ and $c_0 = 0.25$.

Our data give $n = 20$, $\bar{y} = 97.5$ and $S_d = \sum_{i=1}^{20}(y_i - \bar{y})^2 = 2130.72$. Hence $c_1 = c_0 + 20 = 20.25$ and the Bayes factor in favour of H_1 is

$$K = \left(\frac{0.25}{20.15}\right)^{1/2} \exp\left\{\frac{20^2 \times 0.01}{2 \times 20.25}(97.5 - 100)^2\right\} = 0.206.$$

This Bayes factor would be interpreted as giving some evidence *against* H_1 and *in favour* of H_0. It may seem surprising that an observed value of $\bar{y}$ which lies in the set of values of μ allowed by H_1 but not by H_0 should provide evidence in favour of H_0 but this is a feature of the use of sharp hypotheses. Remember that the probability of observing $\bar{Y}$ equal to m^* is zero so it will never happen.

20 Vague Priors

20.1 Introduction

Probably the most frequent objection to the use of Bayesian inference in statistics concerns the use of the prior distribution. It is argued that this introduces an element of subjectivity into the analysis and that this is undesirable. Even without this objection, people sometimes feel that they have little or no prior information and that the prior distribution should reflect this ignorance. For these reasons people sometimes try to use prior distributions which have one or more of the following properties.

- The prior distribution has, in some sense, little effect on the posterior distribution. This is often taken to mean that the posterior distribution is (at least almost) proportional to the likelihood. It could also mean, for example, that changing the prior mean has little effect on the posterior mean.
- The prior distribution conveys little information about the value of the parameter. Typically this means that the prior distribution has a large variance.
- The prior distribution is a "standard" prior which is automatically chosen in some way. Such a prior is sometimes called a *reference* prior.

Prior distributions satisfying these requirements are sometimes described as "noninformative", although, as we will see, this description may be misleading.

20.2 Example: Normal mean with known precision

Suppose we wish to learn about the mean μ of a normal $N(\mu,\ \tau^{-1})$ distribution where the value of τ is known. We will observe a sample of n observations $y_1, \ldots, y_n$ which are independent given μ. Suppose the prior distribution for μ is $\mu \sim N(M_0,\ P_0^{-1})$. The posterior distribution is then

$$\mu \mid y_1, \ldots, y_n \sim N(M_1,\ P_1)$$

where

$$\begin{aligned} M_1 &= \frac{P_0 M_0 + P_d \bar{y}}{P_0 + P_d} \\ P_1 &= P_0 + P_d \\ P_d &= n\tau \end{aligned}$$

(see Lecture 14).

Suppose we make P_0 very small to reflect prior ignorance about the value of μ. (That is, we make the prior variance very large). The resulting prior distribution would be called a *vague* or *diffuse* prior distribution. Then

$$\begin{aligned} M_1 &\approx \bar{y} \\ P_1 &\approx P_d. \end{aligned}$$

The choice of M_0 then has (virtually) no effect on the posterior distribution.

20.3 Proper and improper priors

Suppose, in 20.2, we let $P_0 \to 0$. Then it would appear that we have exactly

$$\begin{aligned} M_1 &= \bar{y} \\ P_1 &= P_d. \end{aligned}$$

However, consider what happens to the prior pdf of μ as $P_0 \to 0$. The pdf is

$$f^0(\mu) = (2\pi)^{-1/2} P_0^{1/2} \exp\left\{-\frac{P_0}{2}(\mu - M_0)^2\right\}.$$

Clearly, as $P_0 \to 0$, $f^0(\mu) \to 0$ for all μ. In the limit the pdf is zero everywhere and we can no longer integrate it and get a total probability of 1.

This seems to be a problem but, since we only need to work in terms of proportionality when we apply Bayes' theorem, it is sometimes still possible to get "sensible" answers in some case, even when we use such an "impossible" prior distribution.

Definition

(The definitions are given in terms of a real scalar variable. There are similar definitions for vectors and variables on other sets).

Proper pdf : A pdf $f(x)$ is a *proper* pdf if and only if $f(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f(x)\ dx$ exists and

$$\int_{-\infty}^{\infty} f(x)\ dx = 1.$$

Improper pdf : For the purposes of this module, a function $f(x)$ such that $f(x) \geq 0$ for all x but $\int_{-\infty}^{\infty} f(x)\ dx$ does not exist or

$$\int_{-\infty}^{\infty} f(x)\ dx \neq 1.$$

is an *improper* pdf.

Proper and improper priors : A prior distribution with a proper pdf is a *proper prior distribution.* A prior distribution with an improper pdf is an *improper prior distribution.*

According to this definition, if $\int_{-\infty}^{\infty} f(x)\ dx = 2$, for example, then $f(x)$ is not a proper pdf. However this would easily be put right by dividing $f(x)$ by 2. However we do come across the use of prior "distributions" with improper densities where the integral is undefined. The normal example in 20.2 is an example of this when we let $P_0 \to 0$. In this case, the corresponding posterior distribution *is* defined and is proper. It is $N(\bar{y},\ P_d^{-1})$.

20.4 Example 20.2 continued

Suppose that $\tau = 0.0625$. That is $\sigma^2 = 16$. Suppose that

$$n = 20, \qquad \sum y = 350.27, \qquad \bar{y} = \frac{350.27}{20} = 17.5135.$$

Suppose that we have a vague (improper) uniform prior.
Posterior mean:

$$M_1 = \bar{y} = 17.5135$$

Posterior precision:

$$P_1 = n\tau = 20 \times 0.0625 = 1.25$$

Posterior variance:

$$V_1 = P_1^{-1} = \frac{1}{1.25} = 0.8 = \frac{\sigma^2}{n}$$

So

$$\mu \sim N(17.5135,\ 0.8)$$

95% posterior credible interval:

$$\bar{y} \pm 1.96\sqrt{\frac{\sigma^2}{n}}$$

That is

$$16.65 < \mu < 18.37$$

20.5 Uniform priors

It is often thought that the uniform distribution provides a suitable noninformative prior (but see 20.7 below). For example, if we wish to learn about the parameter θ in a binomial(N, θ) distribution then a uniform(0,1) prior distribution for θ would mean that the posterior distribution would be exactly proportional to the likelihood and this might seem appropriate since we know that $0 \leq \theta \leq 1$.

In cases where the parameter has an infinite range, e.g. the mean μ in a normal distribution, where $-\infty < \mu < \infty$, or the mean λ of a Poisson distribution, where $0 < \lambda < \infty$, the uniform distribution presents a problem since it has a finite range.

Suppose, for example, that we want to learn about the mean height of people in some faraway country and we propose that, given the parameters, the height y, in metres, of an individual has a normal $N(\mu, \tau^{-1})$ distribution. Then we could give μ a uniform prior distribution on a suitably

wide, but finite, range. For example we could use $\mu \sim \mathrm{U}(0, 10)$. We are hardly likely to find any individuals with heights less than 0 or greater than 10m! The posterior density will then be proportional to the likelihood between the limits and zero outside the limits. Provided that the limits are wide enough the resulting posterior distribution (if τ is assumued known) will be approximately $N(\bar{y},\ P_d^{-1})$. (It is only approximate because the normal distribution is truncated at the prior limits but this might be a very small effect).

The uniform prior in this example is proper but we only get an approximately normal posterior and we have to choose the limits of the uniform distribution. If we let the limits tend to $-\infty$ and ∞, then the prior becomes improper but the limiting posterior distribution is exactly normal $N(\bar{y},\ P_d^{-1})$.

20.6 Example: Normal mean with unknown precision, uniform priors

(See lecture 14 for comparison).

Our model is

$$Y \sim N(\mu, \tau^{-1}).$$

Both μ and τ are unknown.

We will observe $y_1, \ldots, y_n$.

Suppose that we give μ and τ independent uniform priors.

$$\mu \sim \mathrm{U}(c_1, c_2), \qquad \tau \sim \mathrm{U}(0, d)$$

where $c_1,\ c_2,\ d$ are chosen so that the posterior density is effectively proportional to the likelihood. That is d very large, $c_1 << \mathrm{E}_0(Y),\ c_2 >> \mathrm{E}_0(Y)$.

The posterior density is then approximately proportional to the likelihood

$$\begin{aligned}
L(\mu, \tau) &\propto \tau^{n/2} \exp\left\{-\frac{\tau}{2}[S_d + n(\bar{y} - \mu)^2]\right\} \\
&\propto \tau^{n/2} e^{-(S_d/2)\tau} \exp\left\{-\frac{n\tau}{2}(\bar{y} - \mu)^2\right\} \\
&\propto \tau^{(n-1)/2} e^{-(S_d/2)\tau} (2\pi)^{-1/2} (n\tau)^{1/2} \exp\left\{-\frac{n\tau}{2}(\mu - \bar{y})^2\right\} \\
&\propto \frac{(S_d/2)^{(n+1)/2}}{\Gamma[(n+1)/2]} \tau^{(n+1)/2-1} e^{-(S_d/2)\tau} (2\pi)^{-1/2} (n\tau)^{1/2} \exp\left\{-\frac{n\tau}{2}(\mu - \bar{y})^2\right\}
\end{aligned}$$

So, in the posterior, the conditional distribution of μ given τ is

$$\mu \mid \tau \sim N(\bar{y},\ (n\tau)^{-1})$$

and the marginal distribution of τ is

$$\tau \sim \mathrm{gamma}([n+1]/2,\ S_d/2).$$

That is

$$S_d\tau = (n+1)s^2\tau \sim \chi^2_{n+1}$$

where

$$s^2 = \frac{S_d}{n+1} = \frac{\sum(y - \bar{y})^2}{n+1}.$$

The marginal distribution of μ is given by

$$\frac{\mu - \bar{y}}{\sqrt{s^2/n}} \sim t_{n+1}.$$

20.7 The problem of transformations

In 20.5 we suggested that a uniform prior distribution might be considered "noninformative." However a uniform distribution for θ implies a non-uniform distribution for, e.g. θ^2 or for $\exp(\theta)$. Consider, for example, a uniform U(0,1) prior for the parameter θ of a geometric(θ) distribution. The mean of this distribution (given θ) is $\mu = \theta^{-1}$. The implied prior density for μ is

$$f_\mu^{(0)}(\mu) = f_\theta^{(0)}(\theta)/|J|$$

where $f_\theta^{(0)}(\theta)$ is the prior density of θ so $f_\theta^{(0)}(\theta) = 1$ for $0 < \theta < 1$ and $J = d\mu/d\theta = -\theta^{-2} = -\mu^2$. Therefore the prior density of μ is

$$f_\mu^{(0)}(\mu) = \mu^{-2} \qquad (1 < \mu < \infty).$$

This is a proper pdf but it is certainly *not* uniform. So how can a prior be noninformative for θ but not noninformative for $1/\theta$?

20.8 Jeffreys priors

As a way of overcoming the difficulty described in 20.7 Jeffreys proposed using a prior which turns out to be the same whether or not we transform the parameter (considering 1-1 transformations).

Let us restrict our attention to scalar parameters. The *Jeffreys prior* for a parameter θ has density proportional to $\sqrt{I(\theta)}$. Here $I(\theta)$ is the Fisher information

$$I(\theta) = -\mathrm{E}\left(\frac{d^2 \log(L)}{d\theta^2}\right)$$

where L is the likelihood. We notice straight away that this depends on the likelihood but it does not depend on the actual values of the data because we take expectations over the distribution of the data given θ.

Suppose $\phi = g(\theta)$ where g is a 1-1 function. Then the Jeffreys prior for ϕ is the same whether we derive it directly or whether we first find the Jeffreys prior for θ and then apply the transformation.

Proof : Let $l = \log(L)$. Then

$$\begin{aligned} \frac{dl}{d\phi} &= \frac{dl}{d\theta}\frac{d\theta}{d\phi} \\ \frac{d^2 l}{d\phi^2} &= \frac{dl}{d\theta}\frac{d^2\theta}{d\phi^2} + \frac{d^2 l}{d\theta^2}\left(\frac{d\theta}{d\phi}\right)^2 \end{aligned}$$

It can be shown that

$$\mathrm{E}\left(\frac{dl}{d\theta}\right) = 0$$

so

$$I(\phi) = -\mathrm{E}\left(\frac{d^2 l}{d\phi^2}\right) = I(\theta)\left(\frac{d\theta}{d\phi}\right)^2.$$

Therefore the Jeffreys prior for ϕ would have density proportional to

$$\sqrt{I(\theta)}\left|\frac{d\theta}{d\phi}\right|$$

but this is exactly what we would get if we gave the Jeffreys prior with density proportional to $\sqrt{I(\theta)}$ to θ.

Local history : Sir Harold Jeffreys (1891-1989) was born in Fatfield, County Durham. He was a student at Armstrong College from 1907 to 1910 when he graduated. At that time Armstrong College was part of Durham University but, of course, was located here in Newcastle. It eventually became Newcastle University. There is a plaque in the ground floor corridor of the Armstrong building leading from the quadrangle.

Examples

In the Poisson and normal cases we assume a sample of n independent (given the parameter) observations.

Poisson :

$$\begin{aligned} L &\propto e^{-n\lambda}\lambda^{\sum x} \\ l = \log(L) &= \text{constant } -n\lambda + \sum x \log(\lambda) \\ \frac{dl}{d\lambda} &= -n + \frac{\sum x}{\lambda} \\ \frac{d^2l}{d\lambda^2} &= -\frac{\sum x}{\lambda^2} \\ I(\lambda) = -\mathrm{E}\left(\frac{d^2l}{d\lambda^2}\right) &= \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda} \end{aligned}$$

Therefore the Jeffreys prior has density proportional to

$$\sqrt{\frac{1}{\lambda}} = \lambda^{-1/2}.$$

This is improper but it does lead to a proper posterior.

Normal mean with known precision τ :

$$\begin{aligned} L &\propto \exp\left\{-\frac{\tau}{2}\sum(y_i-\mu)^2\right\} \\ l = \log(L) &= \text{constant } -\frac{\tau}{2}\sum(y_i-\mu)^2 \\ \frac{dl}{d\mu} &= \tau\sum(y_i-\mu) \\ \frac{d^2l}{d\mu^2} &= -n\tau \\ I(\mu) = -\mathrm{E}\left(\frac{d^2l}{d\mu^2}\right) &= n\tau \end{aligned}$$

Therefore the Jeffreys prior has density proportional to

$$\sqrt{\tau},$$

i.e. a constant. This is improper but it does lead to a proper posterior. (See 20.5).

Binomial : We have an observation x from bin$(n,\ \theta)$.

$$\begin{aligned} L &\propto \theta^x(1-\theta)^{n-x} \\ l = \log(L) &= \text{constant } + x\log(\theta) + (n-x)\log(1-\theta) \\ \frac{dl}{d\theta} &= \frac{x}{\theta} - \frac{n-x}{1-\theta} \\ \frac{d^2l}{d\theta^2} &= -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2} \\ I(\theta) = -\mathrm{E}\left(\frac{d^2l}{d\theta^2}\right) &= \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)} \end{aligned}$$

Therefore the Jeffreys prior has density proportional to

$$\sqrt{\theta^{-1}(1-\theta)^{-1}} = \theta^{-1/2}(1-\theta)^{-1/2}.$$

This is a beta$(1/2,\ 1/2)$ distribution. This is proper.