

Prior Distributions

There are three main ways of choosing a prior.

- Subjective
- Objective and informative
- Noninformative

Subjective

As mentioned previously, the prior may be determined subjectively. In this case the prior expresses the experimenter's personal probability that θ lies in (essentially) any given subset of Θ .

See Chapter 3 of Berger, *Statistical Decision Theory and Bayesian Analysis* for a discussion of methods for subjectively choosing a prior.

Objective and informative

The experimenter may have information or data that can be used to help formulate a prior. This could take at least two forms:

1. Historical data on the distribution of parameter values.
2. Data from experiments done prior to the one being undertaken.

An example of 1 is as follows. A company wants to estimate the proportion of all parts produced on a particular day that are defective. They will take only a sample of the day's production to estimate this proportion.

From production records we have $\theta_1, \theta_2, \dots, \theta_N$, which are (approximate) proportions of defective parts for a sequence of N days.

One may fit a distribution to the observations $\theta_1, \dots, \theta_N$ and use this as a prior. The usual options are available for fitting the distribution: parametric methods, histograms, kernel density estimates, etc.

Rather than having observations on the parameter itself, one may have previous data that merely contains information about θ . In this case we could use the previous posterior as the prior for the upcoming experiment. This is summarized in the following maxim:

“Today’s posterior is tomorrow’s prior.”

Let's look at the maxim a bit closer. Define the following quantities:

- \mathbf{Y}_1 : the random vector whose value, \mathbf{y}_1 , has already been observed
- \mathbf{Y}_2 : the random vector we are to observe in the next experiment
- π : the prior previous to the first experiment (in which we obtained \mathbf{y}_1)
- $f_1(\mathbf{y}_1|\boldsymbol{\theta})$: the distribution of \mathbf{Y}_1 given $\boldsymbol{\theta}$
- $f_2(\mathbf{y}_2|\boldsymbol{\theta})$: the distribution of \mathbf{Y}_2 given $\boldsymbol{\theta}$
- $f(\mathbf{y}_1, \mathbf{y}_2|\boldsymbol{\theta})$: the joint distribution of \mathbf{Y}_1 and \mathbf{Y}_2 given $\boldsymbol{\theta}$

Two possible approaches for inferring θ having observed both \mathbf{y}_1 and \mathbf{y}_2 :

1. Use the posterior, $\pi_1(\theta|\mathbf{y}_1)$, from experiment 1 as the prior leading into experiment 2 where we will observe a value of \mathbf{Y}_2 . (This is literally what the maxim says.)
2. Treat $(\mathbf{y}_1, \mathbf{y}_2)$ as one big set of data with likelihood f and compute posterior

$$\pi(\theta|\mathbf{y}_1, \mathbf{y}_2) \propto f(\mathbf{y}_1, \mathbf{y}_2|\theta)\pi(\theta).$$

When do the two approaches coincide?

In approach 1, the posterior after experiment 2 is

$$\begin{aligned}\pi_2(\boldsymbol{\theta}|\mathbf{y}_2) &\propto \pi_1(\boldsymbol{\theta}|\mathbf{y}_1)f_2(\mathbf{y}_2|\boldsymbol{\theta}) \propto \\ &f_1(\mathbf{y}_1|\boldsymbol{\theta})f_2(\mathbf{y}_2|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).\end{aligned}$$

In general, approaches 1 and 2 will give the same posterior only when

$$f(\mathbf{y}_1, \mathbf{y}_2|\boldsymbol{\theta}) = f_1(\mathbf{y}_1|\boldsymbol{\theta})f_2(\mathbf{y}_2|\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta},$$

which requires \mathbf{Y}_1 and \mathbf{Y}_2 to be statistically independent.

So, in order for the maxim to be true in the strictest sense, the data not yet observed should be independent of those already observed.

Noninformative priors

A noninformative prior is one that expresses ignorance as to the value of θ . Other terms for a noninformative prior are reference prior, diffuse prior and vague prior.

In general, a noninformative prior is one which is dominated by the likelihood function. In other words, such a prior

- does not change much over the region in which the likelihood is appreciable, and
- does not assume large values outside that region.

A prior having the two properties above is said to be *locally uniform*. Box and Tiao, *Bayesian Inference in Statistical Analysis* give an excellent account of locally uniform priors.

The Jeffreys noninformative prior

Suppose that $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$. The *Fisher information matrix*, $I(\boldsymbol{\theta})$, is the $p \times p$ matrix with (i, j) element

$$-E \left[\frac{\partial^2 \log f(\mathbf{Y}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right].$$

The Jeffreys noninformative prior is

$$\pi(\boldsymbol{\theta}) \propto \det(I(\boldsymbol{\theta}))^{1/2},$$

where $\det(\mathbf{A})$ denotes the determinant of matrix \mathbf{A} .

The motivation for this prior is a certain invariance argument. Consider a 1-1 transformation of the parameter: $\boldsymbol{\phi} = h(\boldsymbol{\theta})$.

Now, if our prior for $\boldsymbol{\theta}$ is π , then the corresponding density of $\boldsymbol{\phi} = h(\boldsymbol{\theta})$ is

$$g(\boldsymbol{\phi}) = \pi(h^{-1}(\boldsymbol{\phi}))|J(\boldsymbol{\phi})|,$$

where J is the Jacobian of the transformation.

Let M_P be a method for finding a noninformative prior. Now, suppose M_P is used to obtain a prior π for θ . Jeffreys argued that if M_P is used to find a prior π^* for $\phi = h(\theta)$, then it should be true that

$$\pi^*(\phi) = g(\phi) \quad \forall \phi,$$

where g is defined at the bottom of the previous page.

The Jeffreys prior satisfies this property. Let's check this in the case $p = 1$. We have

$$\pi(\theta) \propto \left\{ -E \left[\frac{\partial^2 \log f(\mathbf{Y}|\theta)}{\partial \theta^2} \right] \right\}^{1/2}$$

and

$$g(\phi) = \pi(h^{-1}(\phi)) \left| \frac{dh^{-1}(\phi)}{d\phi} \right|.$$

The Fisher information when we use the parameterization $\phi = h(\theta)$ is

$$-E \left[\frac{\partial^2 \log f(\mathbf{Y} | h^{-1}(\phi))}{\partial \phi^2} \right],$$

and so the Jeffreys prior for ϕ is proportional to the square root of the last expression, which is

$$\begin{aligned} & -E \left[\frac{\partial^2 \log f(\mathbf{Y} | h^{-1}(\phi))}{\partial h^{-1}(\phi)^2} \cdot \frac{\partial h^{-1}(\phi)^2}{\partial \phi^2} \right] = \\ & - \left(\frac{\partial h^{-1}(\phi)}{\partial \phi} \right)^2 E \left[\frac{\partial^2 \log f(\mathbf{Y} | h^{-1}(\phi))}{\partial h^{-1}(\phi)^2} \right]. \end{aligned}$$

But the last expression is proportional to $g^2(\phi)$ (as defined on the previous page), and hence Jeffreys invariance property holds.

Example 7 *Jeffreys prior for the binomial experiment*

We have

$$\log f(y|\theta) = \log \binom{n}{y} + y \log \theta + (n-y) \log(1-\theta),$$

$$\frac{\partial \log f(y|\theta)}{\partial \theta} = \frac{y}{\theta} - \frac{(n-y)}{(1-\theta)},$$

and

$$\frac{\partial^2 \log f(y|\theta)}{\partial \theta^2} = -\frac{y}{\theta^2} - \frac{(n-y)}{(1-\theta)^2}.$$

So,

$$\begin{aligned} -E \left[\frac{\partial^2 \log f(Y|\theta)}{\partial \theta^2} \right] &= \frac{n\theta}{\theta^2} + \frac{(n-n\theta)}{(1-\theta)^2} \\ &= \frac{n}{\theta(1-\theta)}. \end{aligned}$$

So, the Jeffreys noninformative prior for θ is proportional to $[\theta(1-\theta)]^{-1/2}$, and hence must be a Beta(1/2, 1/2) density.

Example 8 *Jeffreys prior for normal random sample*

Let $\mathbf{Y} = (Y_1, \dots, Y_n)$, where Y_1, \dots, Y_n are i.i.d. $N(\theta_1, \theta_2^2)$. We have

$$f(\mathbf{y}|\boldsymbol{\theta}) = \left(\frac{1}{\sqrt{2\pi\theta_2}} \right)^n \exp \left(-\frac{1}{2\theta_2^2} \sum_{i=1}^n (y_i - \theta_1)^2 \right),$$

and

$$\log f(\mathbf{y}|\boldsymbol{\theta}) = -n \log(\sqrt{2\pi\theta_2}) - \frac{1}{2\theta_2^2} \sum_{i=1}^n (y_i - \theta_1)^2.$$

Now,

$$\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \theta_1} = \frac{1}{\theta_2^2} \sum_{i=1}^n (y_i - \theta_1),$$

$$\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \theta_2} = -\frac{n}{\theta_2} + \frac{1}{\theta_2^3} \sum_{i=1}^n (y_i - \theta_1)^2,$$

$$\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \theta_1^2} = -\frac{n}{\theta_2^2},$$

$$\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \theta_2^2} = \frac{n}{\theta_2^2} - \frac{3}{\theta_2^4} \sum_{i=1}^n (y_i - \theta_1)^2,$$

and

$$\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} = -\frac{2}{\theta_2^3} \sum_{i=1}^n (y_i - \theta_1).$$

The Fisher information matrix is thus

$$\begin{bmatrix} n/\theta_2^2 & 0 \\ 0 & 2n/\theta_2^2 \end{bmatrix}.$$

The determinant of this matrix is $2n^2/\theta_2^4$, and hence the Jeffreys noninformative prior is such that

$$\pi(\theta_1, \theta_2) \propto \frac{1}{\theta_2^2} I_{(-\infty, \infty)}(\theta_1) I_{(0, \infty)}(\theta_2).$$

The “prior” at the bottom of the previous page is called *improper* since it is not integrable. This is an example of the unfortunate fact that Jeffreys noninformative prior is sometimes improper.

Note that the form of Jeffreys prior in this case implies that θ_1 and θ_2 are a priori independent with

$$\pi_1(\theta_1) = \text{constant} \quad \forall \theta_1$$

and

$$\pi_2(\theta_2) = \frac{1}{\theta_2^2} I_{(0, \infty)}(\theta_2).$$

So, it is the prior for the location parameter θ_1 that is improper. The prior for θ_2 is proper.
