Some One-parameter Models

Bayesian inference will be introduced by considering the simplest possible models, ones having only a single *scalar* parameter.

Models considered:

- Binomial experiment
- Poisson data
- Random sample from exponential distribution

Binomial experiment

The basic binomial experiment consists of a sequence of independent trials, each of which results in either 1 or 0, "success" or "failure."

It is assumed that the probability of a success remains constant from trial to trial, and we call this probability θ.

A useful function for defining probability distributions is the *indicator function*. Let A be a set of real numbers. The indicator function, I_A, is defined by $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ if $x \notin A$.

In the binomial experiment we have i.i.d. observations $Y_1, \ldots, Y_n$, where Y_i has the following Bernoulli distribution:

$$P(Y_i = y) = \theta^y (1-\theta)^{1-y} I_{\{0,1\}}(y).$$

The joint conditional distribution of the data given θ is

$$\begin{aligned} p(y_1, \ldots, y_n | \theta) &= \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} I_{\{0,1\}}(y_i) \\ &= \exp\left[\sum_{i=1}^{n} y_i \log\theta \right. \\ &\quad \left. + \left(n - \sum_{i=1}^{n} y_i \right) \log(1-\theta) \right] I_n, \end{aligned}$$

where $I_n = \prod_{i=1}^{n} I_{\{0,1\}}(y_i)$.

Let's say the given set of data is $y_1, \ldots, y_n$, where each y_i is either 0 or 1, and let $y = \sum_{i=1}^n y_i$. Then if the prior for θ is p, the posterior distribution is

$$\begin{aligned} p(\theta|y_1, \ldots, y_n) &= \frac{p(\theta)\theta^y(1-\theta)^{n-y}}{\int_0^1 p(t)t^y(1-t)^{n-y}\,dt} \\ &\propto p(\theta)\theta^y(1-\theta)^{n-y}. \end{aligned} \tag{4}$$

Fundamental fact:

> It is always true that the posterior distribution depends on the data *only through a sufficient statistic.*

A sufficient statistic is one that contains *all* the information in the data set.

In the binomial experiment the sufficient statistic is $\sum_{i=1}^n y_i$, the total number of successes.

Writing the posterior as being proportional to (4) is worth discussion. Suppose f is a nonnegative function that can be integrated. Then

$$g(x) = \frac{f(x)}{\int_{-\infty}^{\infty} f(t)\,dt}$$

is a density.

Now, suppose we know that f is proportional to some density h. This means that $f(x) = Ch(x)$ for all x and some constant C.

It follows that $f(x)/C$ is a density. The important point here is that *once we know that f is proportional to a "familiar" density, we don't have to do any integration to figure out the multiplicative constant that turns f into a density.*

Example 1 Suppose that

$$f(x) = \exp(-(5x^2 + 15x))$$

for all x. It should be clear that f is proportional to a density, because it's positive and integrable.

Is f proportional to a well-known density? The answer is yes, but we have to work a bit to see what that density is.

The function looks kind of like a normal density, but the argument of a normal density has the form $-(x-\mu)^2/(2\sigma^2)$.

We can make the argument of f look like this if we complete the square.

We have

$$\exp(-(5x^2 + 15x)) =$$

$$\exp(-5(x^2 + 3x)) =$$

$$\exp\left[-5\left(x^2 + 2\left(\tfrac{3}{2}\right)x + \tfrac{9}{4} - \tfrac{9}{4}\right)\right] =$$

$$\exp\left[-5(x + 3/2)^2 + \tfrac{45}{4}\right] =$$

$$\exp\left(\tfrac{45}{4}\right)\exp\left[-5(x + 3/2)^2\right] \propto$$

$$\exp\left[-5(x + 3/2)^2\right] =$$

$$\exp\left[-\tfrac{1}{2(1/10)}(x + 3/2)^2\right].$$

Without further ado, we can say that f *is proportional to a normal density with mean* $-3/2$ *and variance* $1/10$.

We don't need to do any integration to know what the appropriate normalizing constant is. We know that the $N(-3/2, 1/10)$ density is

$$\begin{aligned} h(x) &= \frac{1}{\sqrt{2\pi}(1/\sqrt{10})} \\ &\quad \times \exp\left[-\frac{1}{2(1/10)}(x+3/2)^2\right] \\ &= \frac{\sqrt{5}}{\sqrt{\pi}} \exp\left[-5(x+3/2)^2\right]. \end{aligned}$$

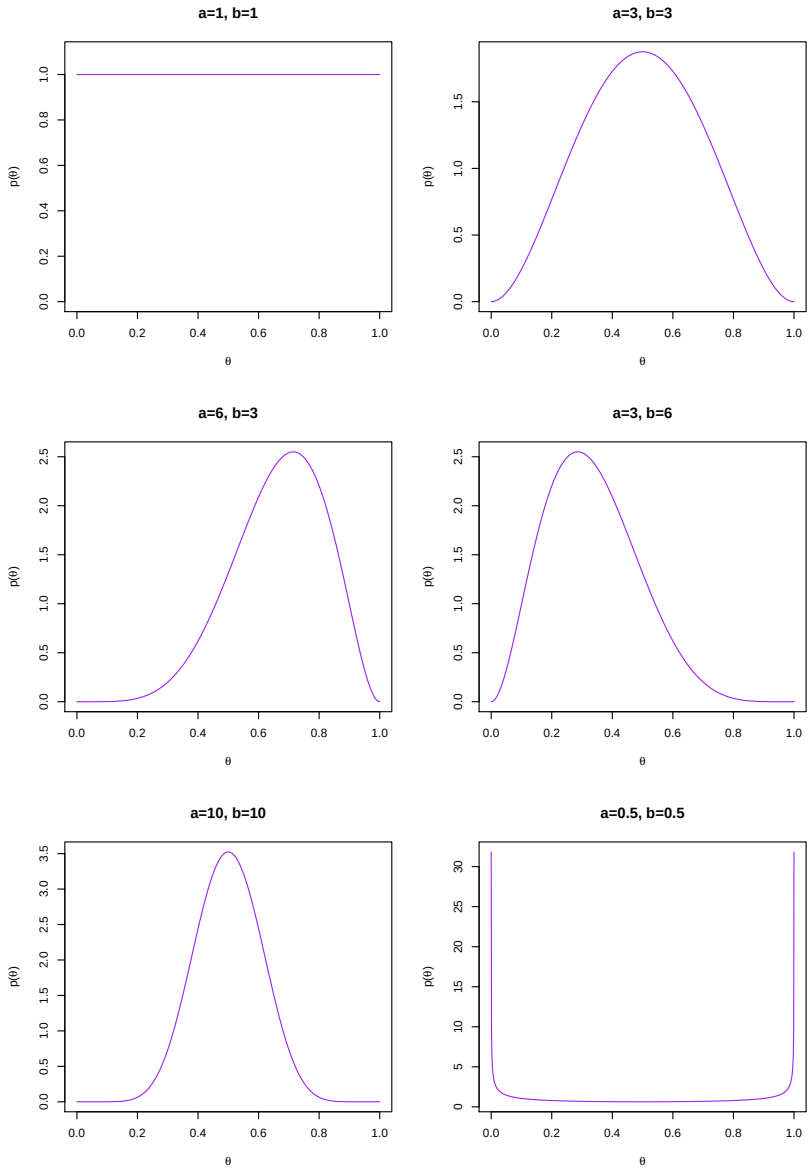
What shall we use as a prior distribution? A commonly used prior for the binomial experiment is the beta(a, b) density, which is

$$p(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\boldsymbol{I}_{(0,1)}(\theta),$$

where $a > 0$, $b > 0$ and Γ denotes the gamma function.

This family provides fair scope for expressing the experimenter's uncertainty about θ. By varying a and b, one can obtain a variety of distributional shapes and locations.

Various beta densities



Now, let's suppose that we've chosen to use a beta prior having specific parameters a and b. Then the posterior density is such that

$$\begin{aligned} p(\theta|y_1,\ldots,y_n) &\propto \theta^y(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{y+a-1}(1-\theta)^{n-y+b-1}. \end{aligned}$$

The last expression is proportional to a beta$(y+a, n-y+b)$ density, and so the posterior *must* be beta$(y+a, n-y+b)$.

So, when one uses a beta prior in a binomial experiment, the resulting posterior is also beta, but with different parameters. This type of prior is called a *conjugate prior*.

Definition 1 Let $\mathcal{F}$ be the class of densities of which it is assumed that $p(\boldsymbol{y}|\boldsymbol{\theta})$ is a member. A class $\mathcal{P}$ of prior distributions is said to be a *conjugate family for* $\mathcal{F}$ if $p(\cdot|\boldsymbol{y}) \in \mathcal{P}$ for all $p \in \mathcal{P}$ and for every possible data set $\boldsymbol{y}$.

Now that we know what the posterior is, how do we use it to make inferences about θ? In frequentist statistics the following are the "big three" of inference:

- Point estimation
- Confidence intervals
- Hypothesis testing

Each of these has its analog in the Bayesian world. For the moment, we'll just describe, in the context of the binomial experiment, a couple of ways that a Bayesian would obtain point estimates.

One popular Bayes point estimate is the *mode of the posterior distribution*. This is analogous to a *maximum likelihood estimate*.

In the binomial experiment with beta(a, b) prior, we found that the posterior is beta$(y + a, n - y + b)$.

It's easy to check that a beta(c, d) density has mode $(c - 1)/(c + d - 2)$ so long as $c > 1$ and $d > 1$. Therefore the mode of the posterior is $(y + a - 1)/(n + a + b - 2)$ whenever $n > 1$ and $0 < y < n$.

A classical frequentist estimate of θ in the binomial experiment is the sample proportion $\widehat{\theta} = y/n$.

How does the mode of the posterior compare with $\widehat{\theta}$?

$$\begin{aligned} \text{mode} &= \frac{y + a - 1}{n + a + b - 2} \\ &= w_n \widehat{\theta} + (1 - w_n) \left(\frac{a - 1}{a + b - 2} \right), \end{aligned}$$

where $w_n = n/(n + a + b - 2)$.

We can note the following facts:

- The mode is a weighted average of the classical frequentist estimator and the mode of the prior distribution.

- As n gets large, the weight on the frequentist estimator tends to 1, and hence the weight on the prior mode tends to 0.

The same sort of thing happens in most Bayesian analyses. This is an indication of something mentioned earlier:

As the sample size gets larger and larger the effect of the prior becomes smaller and smaller.

Another popular Bayes point estimate is the posterior mean.

The mean of a beta(c, d) density is $c/(c + d)$, which is true for all $c > 0$ and $d > 0$. So, the mean of our beta(a, b) prior is $a/(a + b)$, and the mean of the posterior is

$$\frac{y + a}{n + a + b} = \widehat{\theta}\left(\frac{n}{n + a + b}\right) + \left(\frac{a}{a + b}\right)\left(1 - \frac{n}{n + a + b}\right).$$

Note that this is a weighted average of $\widehat{\theta}$ and the mean of the prior density.

So the posterior mean has an interpretation similar to that of the posterior mode.

Example 2 Let θ be the proportion of all people suffering from a particular chronic illness who recover within one month when given a certain treatment.

A clinical trial is to be done in which 50 persons suffering from this illness are given the treatment of interest.

Suppose that a beta prior is to be used. We'll look at the effect of choices for a and b on a Bayes estimate of θ.

Suppose that 18 of the 50 persons receiving the treatment recover within one month. The classical point estimate of θ is

$$\hat{\theta} = \frac{18}{50} = 0.36.$$

Let's consider six different priors.

1. $\underline{a = 1,\ b = 1:}$ A uniform prior; gives equal weight to each value of θ.

$$\text{Bayes estimate } = \frac{19}{52} = 0.365$$

2. $\underline{a = 1/2,\ b = 1/2:}$ A particular type of non-informative prior.

$$\text{Bayes estimate } = \frac{18.5}{51} = 0.363$$

3. $\underline{a = 1,\ b = 2:}$ A fairly "vague" prior with mean 1/3.

$$\text{Bayes estimate } = \frac{19}{53} = 0.358$$

4. $\underline{a = 20,\ b = 40:}$ An informative prior with mean 1/3.

$$\text{Bayes estimate } = \frac{38}{110} = 0.345.$$

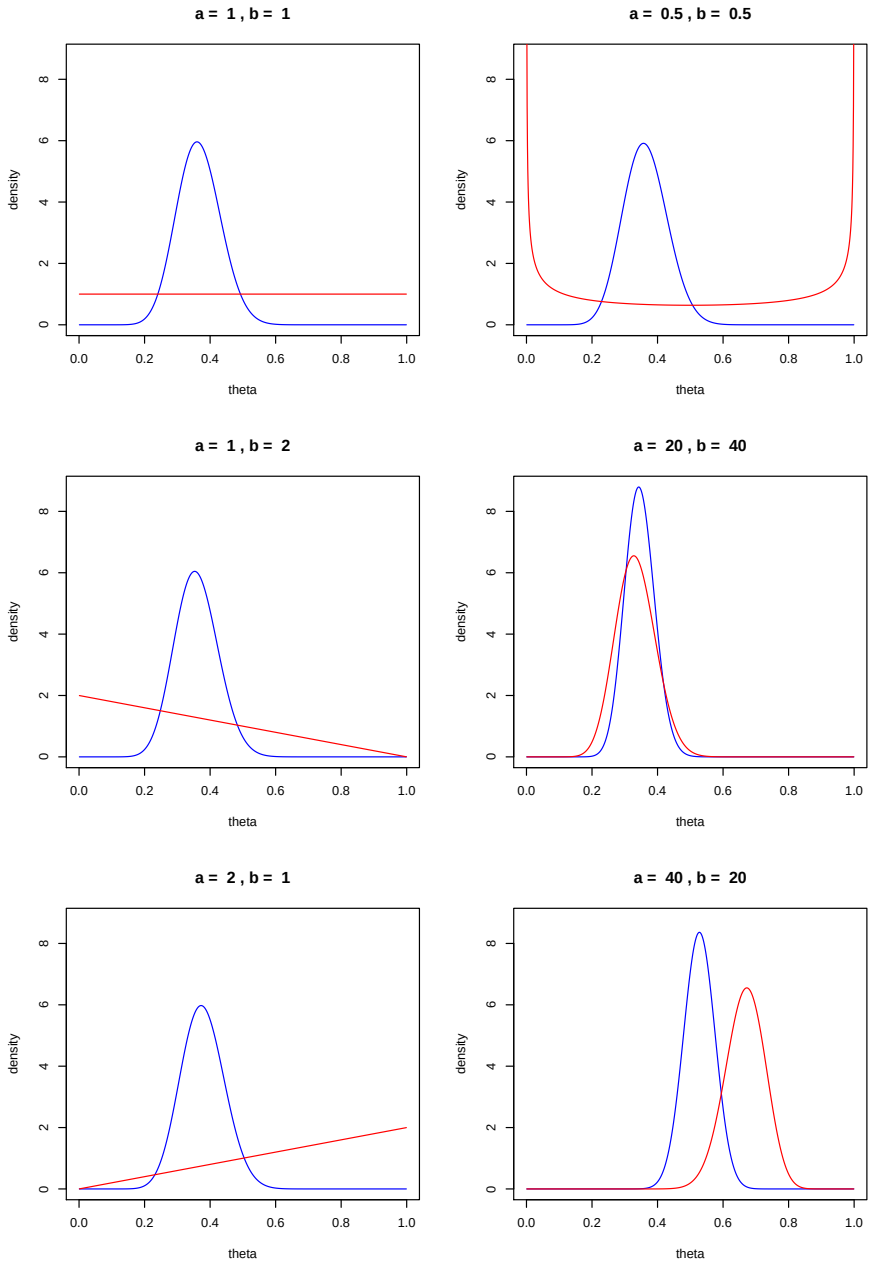
5. $\underline{a = 2,\ b = 1:}$ A vague prior with mean 2/3.

$$\text{Bayes estimate } = \frac{20}{53} = 0.377$$

6. $\underline{a = 40,\ b = 20:}$ An informative prior with mean 2/3.

$$\text{Bayes estimate } = \frac{58}{110} = 0.527$$

Priors and posteriors for Example 2



Prior: —— Posterior: ——

The effect of a beta(a, b) prior on the posterior *mean* has a nice interpretation. *The prior information is equivalent to that in a binomial experiment with a successes and b failures (and hence number of trials $a + b$).* (Recall: posterior mean $= (y + a)/(n + a + b)$.)

So, the uniform prior contains information equivalent to a binomial experiment with 2 trials and 1 success.

The other noninformative prior ($a = 1/2$, $b = 1/2$) contains information equivalent to a binomial experiment with 1 trial and 1/2 of a success. (Use your imagination!)

In Example 2, the prior with $a = 40$ and $b = 20$ effectively contains more information than the data itself, since $a + b = 60 > 50 = n$.

Noninformative priors

A noninformative prior is one that expresses ignorance as to the value of $\boldsymbol{\theta}$. Other terms for a noninformative prior are reference prior, diffuse prior and vague prior.

Suppose θ is a scalar parameter that is restricted to the interval (c, d). It seems like the uniform distribution over (c, d) would *have* to be the "correct" noninformative prior distribution for θ. But is it?

Example 3 Suppose my prior for θ is uniform on (c, d), or $U(c, d)$. Someone tells me that $\tau = \exp(\theta)$ has a better physical interpretation than θ, so I decide to parameterize the model in terms of τ.

Of course I need a prior for τ, and I want it to be noninformative, so I decide to use $U(e^c, e^d)$.

However, if I *really* believe that the $U(c, d)$ prior expresses my prior beliefs about θ, then I'm contradicting myself by using a $U(e^c, e^d)$ prior for τ. Why?

Let $c < x < y < d$. Then presumably I feel that

$$P(x < \theta < y) = \frac{(y - x)}{(d - c)}.$$

But

$$P(x < \theta < y) = P(e^x < \tau < e^y).$$

So, if I'm firm in my belief that $P(x < \theta < y) = (y - x)/(d - c)$, then logically I should also believe that $P(e^x < \tau < e^y) = (y - x)/(d - c)$.

However, if I use the $U(e^c, e^d)$ prior for τ, then

$$P(e^x < \tau < e^y) = \frac{(e^y - e^x)}{(e^d - e^c)} \neq \frac{(y-x)}{(d-c)}.$$

This example shows that the uniform distribution can't always be the correct noninformative prior.

How can we choose a prior that avoids the contradiction encountered in Example 3?

Sir Harold Jeffreys (1891-1989), an eminent astronomer/geophysicist/statistician, had the following simple, yet brilliant, idea:

Let M be a method for choosing a prior, and suppose we apply M to a parameter θ and obtain prior p. Now let g be a monotone and differentiable transformation. *If we apply M to find a prior for $\tau = g(\theta)$, then the resulting prior should be*

$$\pi(\tau) = p(g^{-1}(\tau)) \left| \frac{dg^{-1}(\tau)}{d\tau} \right|. \qquad (5)$$

The motivation for this criterion is that if we say that θ has density p, then logically τ must have the density π in (5).

Not only did Jeffreys propose the invariance-under-transformation criterion, he devised a method that guarantees invariance.

Let $p(\boldsymbol{y}|\boldsymbol{\theta})$ be our probability model for data $\boldsymbol{Y}$. The corresponding *information matrix,* $\boldsymbol{I}(\boldsymbol{\theta})$, is the matrix with (i, j) entry equal to

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

Now, suppose we take

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

Then this prior satisfies the invariance criterion on p. 39.

Example 4 *Jeffreys prior for the binomial experiment*

We have

$$\log p(y|\theta) = y \log \theta + (n-y) \log(1-\theta),$$

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

and

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

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*.

The prior in Example 4 is *proper*, because it satisfies the properties of a density (nonnegativity and integration to 1).

Unfortunately, the Jeffreys prior is often *improper*, meaning that when you integrate it you get ∞.

Many people don't mind using improper priors. However, everyone agrees that doing so is ok only if *the resulting posterior is proper*.

A proper posterior is one such that

$$\int_\Theta p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})\,d\boldsymbol{\theta} < \infty.$$

If this is true, then

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})}{\int_\Theta p(\boldsymbol{t})p(\boldsymbol{y}|\boldsymbol{t})\,d\boldsymbol{t}}$$

is a proper density even if the prior p is not.

Poisson Data

Suppose we observe Y, which has a Poisson distribution with unknown mean θ. Then

$$p(y|\theta) = \frac{e^{-\theta}\theta^y}{y!} I_{\{0,1,\ldots\}}(y),$$

and the parameter space is $\Theta = (0, \infty)$.

Effectively, this model could apply to two different situations:

- One observes a single Poisson count, or
- n i.i.d. Poisson variables.

Why?

A conjugate family for Poisson data

The gamma(a, b) density is

$$p(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}I_{(0,\infty)}(\theta),$$

where a and b are any positive constants.

Now suppose that in our Poisson model we use a gamma(a, b) prior for θ. Then

$$p(\theta|y) \propto \theta^{a-1}e^{-b\theta} \cdot e^{-\theta}\theta^y.$$

Notice that here I just dropped the constant multiplier in the gamma(a, b) density and also the term $1/y!$ from $p(y|\theta)$. Why is this ok?

We have

$$p(\theta|y) \propto \theta^{y+a-1}e^{-(b+1)\theta},$$

and so the posterior is a gamma$(y + a, b + 1)$ density.

Since the posterior is gamma, the gamma distribution is a conjugate family for Poisson data.

Jeffreys prior

The log-probability function is

$$\log p(y|\theta) = -\theta + y \log \theta - \log y!,$$

$$\frac{\partial}{\partial \theta} \log p(y|\theta) = -1 + \frac{y}{\theta}$$

and

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

It follows that

$$-E\left[\frac{\partial^2}{\partial \theta^2} \log p(Y|\theta)\right] = \frac{1}{\theta},$$

and so the Jeffreys noninformative prior is $p(\theta) \propto \theta^{-1/2}$, which is improper.

The posterior corresponding to the Jeffreys prior is such that

$$p(\theta|y) \propto \theta^{y-1/2}e^{-\theta}$$

and hence is Gamma(y+1/2, 1), which is proper regardless of the value of y.

For the gamma(a, b) prior, the posterior mean is

$$\frac{y+a}{b+1} = y \cdot \left(\frac{1}{b+1}\right) + \frac{a}{b} \cdot \left(1 - \frac{1}{b+1}\right).$$

As we saw in the binomial case, this a linear combination of a frequentist estimate (y, the MLE) and the prior mean (a/b).

The posterior mean for the Jeffreys prior is $y + 1/2$.

The mode of the posterior for a gamma(a, b) prior is

$$\text{mode} = \begin{cases} \frac{(y+a-1)}{(b+1),} & y + a > 1, \\ 0, & y + a \leq 1. \end{cases}$$

The posterior corresponding to Jeffreys prior has mode given by the last expression if we take $a = 1/2$ and $b = 0$.

Bayesian credible regions

Bayesian credible regions are the analogs of confidence regions in classical statistics. However, unlike the classical regions, *they have a proper probability interpretation.*

Definition 2 A $100(1-\alpha)\%$ *credible region* for $\boldsymbol{\theta}$ is a subset C of Θ such that

$$P(\boldsymbol{\theta} \in C | \boldsymbol{Y} = \boldsymbol{y}) \geq 1 - \alpha.$$

Two principles are often used in constructing C.

- The volume of C should be as small as possible.

- The posterior density should be greater for every $\boldsymbol{\theta} \in C$ than it is for any $\boldsymbol{\theta} \notin C$.

It turns out that these two criteria are equivalent.

Definition 3 The $100(1-\alpha)\%$ *highest posterior density (HPD)* region for $\boldsymbol{\theta}$ is a subset C of Θ such that

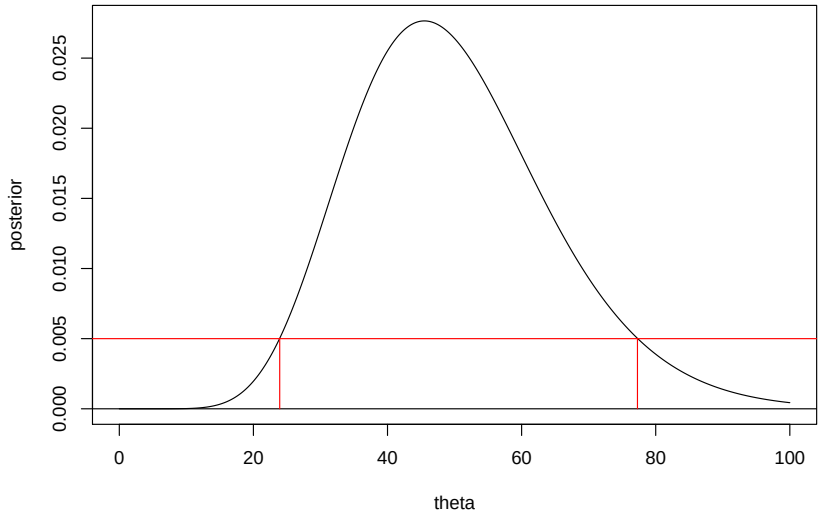
$$C = \{\boldsymbol{\theta} : p(\boldsymbol{\theta}|\boldsymbol{y}) \geq k_\alpha\},$$

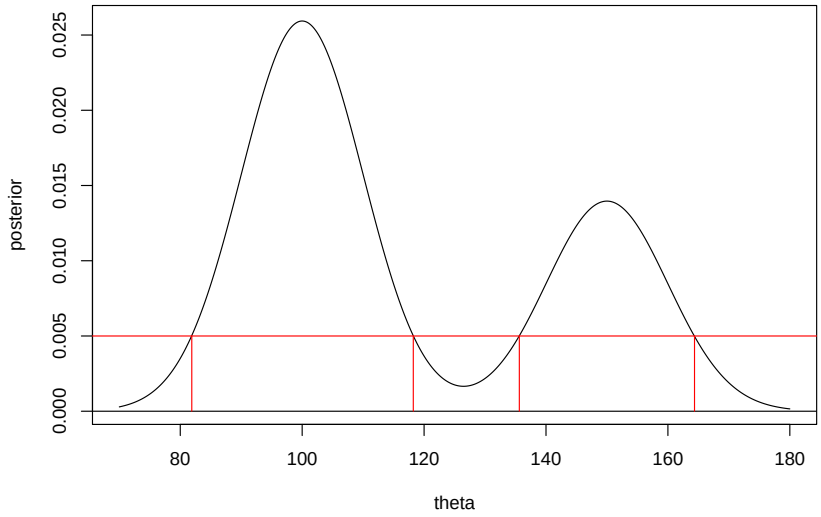
where k_α is the largest constant for which

$$P(\boldsymbol{\theta} \in C|\boldsymbol{Y} = \boldsymbol{y}) \geq 1 - \alpha.$$

An HPD region has the smallest volume of all regions with a given probability content.

Finding an HPD region





In each graph, the set of θ values at which the density exceeds the horizontal line at 0.005 forms an HPD region.

Bayesian hypothesis testing

Let $\Theta = \Theta_0 \cup \Theta_1$, where $\Theta_0 \cap \Theta_1 = \emptyset$. Suppose we want to test

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\theta} \in \Theta_1.$$

Define

$$\alpha_0 = P(\boldsymbol{\theta} \in \Theta_0 | \boldsymbol{Y} = \boldsymbol{y}) \quad \text{and} \quad \alpha_1 = 1 - \alpha_0.$$

The *posterior odds ratio* is α_0/α_1. In Bayesian hypothesis testing, one simply uses α_0/α_1 to assess the relative plausibility of H_0 and H_1.

Let p_0 and p_1 denote the prior probabilities of H_0 and H_1, respectively. Another approach to testing is to use the *Bayes factor*:

$$\begin{aligned} \text{Bayes factor} &= \frac{\text{posterior odds ratio}}{\text{prior odds ratio}} \\ &= \frac{\alpha_0/\alpha_1}{p_0/p_1} \\ &= \frac{\alpha_0 p_1}{p_0 \alpha_1} = B. \end{aligned}$$

$B < 1 \Rightarrow$ degree of belief in H_0 has decreased upon observing the data.

It's of interest to consider testing simple hypotheses, i.e.,

$$\Theta_0 = \{\boldsymbol{\theta}_0\} \quad \text{and} \quad \Theta_1 = \{\boldsymbol{\theta}_1\}.$$

This is the setting where the Neyman-Pearson lemma provides a most powerful test of given size.

In the simple vs. simple case,

$$\alpha_0 = \frac{p_0 p(\boldsymbol{y}|\boldsymbol{\theta}_0)}{p_0 p(\boldsymbol{y}|\boldsymbol{\theta}_0) + p_1 p(\boldsymbol{y}|\boldsymbol{\theta}_1)},$$

the posterior odds ratio is

$$\frac{\alpha_0}{\alpha_1} = \frac{p_0 p(\boldsymbol{y}|\boldsymbol{\theta}_0)}{p_1 p(\boldsymbol{y}|\boldsymbol{\theta}_1)},$$

and the Bayes factor is

$$B = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}_0)}{p(\boldsymbol{y}|\boldsymbol{\theta}_1)}.$$

Two interesting facts about this Bayes factor:

- B is the likelihood ratio statistic upon which the most powerful test is based.

- B is free of any prior probabilities.

Jeffreys' scale for interpreting Bayes factors

This scale appeared in

> Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford, p. 432.

B	Evidence against H_0
$10^{-1/2}$ to 1	Not worth more than a bare mention
10^{-1} to $10^{-1/2}$	Substantial
$10^{-3/2}$ to 10^{-1}	Strong
10^{-2} to $10^{-3/2}$	Very strong
Less than 10^{-2}	Decisive

Exponential data

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, where $Y_1, \ldots, Y_n$ are a random sample from an exponential density with unknown mean.

There are two ways of parameterizing the exponential distribution.

$$f(y|\theta) = \theta e^{-\theta y} I_{(0,\infty)}(y) \qquad (A)$$

$$f(y|\theta) = \frac{1}{\theta} e^{-y/\theta} I_{(0,\infty)}(y) \qquad (B)$$

In either case the (unrestricted) parameter space is $\Theta = (0, \infty)$.

A conjugate family of priors for case (A) is the family of gamma distributions, while for case (B) the inverse-gamma family is conjugate.

Let's find the Jeffreys noninformative prior for case (A). The likelihood function is

$$\begin{aligned} p(\boldsymbol{y}|\theta) &= \theta^n \prod_{i=1}^{n} e^{-\theta y_i} \\ &= \theta^n e^{-\theta n \bar{y}}. \end{aligned}$$

(As an aside we note that the sample mean $\bar{Y}$ is a sufficient statistic.)

Now,

$$\log p(\boldsymbol{y}|\theta) = n \log \theta - \theta n \bar{y},$$

$$\frac{\partial}{\partial \theta} \log p(\boldsymbol{y}|\theta) = \frac{n}{\theta} - n\bar{y},$$

and

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

So, *the Jeffreys noninformative prior for θ is proportional to θ^{-1}*, and hence is improper.

Verify that the Jeffreys prior for case (B) is the same, i.e., it is proportional to θ^{-1}.

Let's compute the posterior for case (A) using a gamma(a, b) prior.

$$p(\theta|\boldsymbol{y}) \propto \theta^{a-1}e^{-b\theta}\theta^{n}e^{-\theta n\bar{y}} = \theta^{n+a-1}e^{-\theta(b+n\bar{y})}.$$

It follows that the posterior is gamma$(n+a, b+n\bar{y})$.

Gamma$(n, n\bar{y})$ is the posterior one obtains with the improper prior $p(\theta) = \theta^{-1}$. If one insisted on a proper noninformative prior, one could always take a and b to be *very* small, but positive.

The mean of the gamma(a, b) distribution is a/b, and hence the posterior mean is

$$\frac{n + a}{b + n\bar{y}} = \frac{1 + a/n}{\bar{y} + b/n}.$$

Regardless of the values of a and b, when n is sufficiently big,

$$\text{posterior mean} \approx \frac{1}{\bar{y}} = \text{MLE}.$$

Also, *if we use the Jeffreys prior, the posterior mean is precisely the same as the MLE*.

Whenever a is bigger than 1, the mode of the gamma(a, b) distribution is $(a - 1)/b$, and so the mode of the posterior is guaranteed to be $(n + a - 1)/(b + n\bar{y})$ whenever $n \geq 2$. This estimate will generally differ very little from the mean of the posterior.

Example 5 HPD region for exponential rate

A company wants to obtain information about the distribution of lifetimes of a certain electronic component. It is assumed that the lifetime of a randomly selected component has density

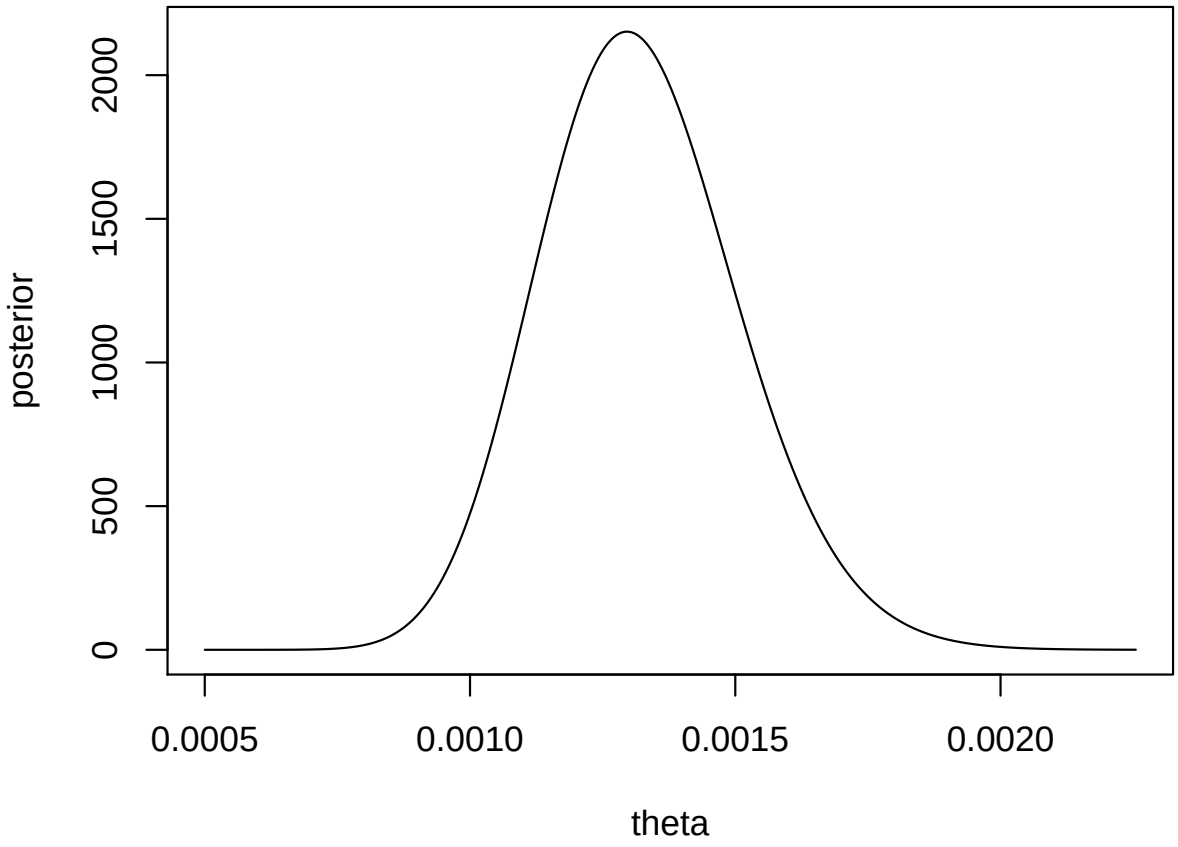
$$f(y|\theta) = \theta e^{-\theta y} I_{(0,\infty)}(y).$$

The company decides to use the noninformative (and improper) prior $p(\theta) = \theta^{-1}$.

Fifty components are put on test until they all fail. The average lifetime of the components was 756.3 hours.

The company wishes to find a 95% HPD region for θ.

The posterior for θ is gamma$(50, 37815)$, which is pictured below.



The mean of this distribution is $1/756.3 = 0.001322$, which of course is the posterior mean.

The standard deviation of the distribution is $\sqrt{50}/[50(756.3)] = 0.0001870$.

To find a 95% HPD region, one can use numerical approximation. The following algorithm was used:

- Choose an initial guess, $\tilde{k}$, for the value of $k_{.05}$ (See p. 48N.)

- Use Newton's method to solve the equation $p(\theta|\bar{y}) = \tilde{k}$ for its two solutions, call them $\theta_1 < \theta_2$.

- Evaluate $prob = P(\theta < \theta_1|\bar{y}) + P(\theta > \theta_2|\bar{y})$.

- If $prob$ is sufficiently close to 0.05, then stop and use (θ_1, θ_2) as the HPD region. Otherwise, repeat the previous steps starting from a different guess $\tilde{k}$.

The previous algorithm was applied starting from an initial $\tilde{k}$ of 250. Initial guesses for the solutions of

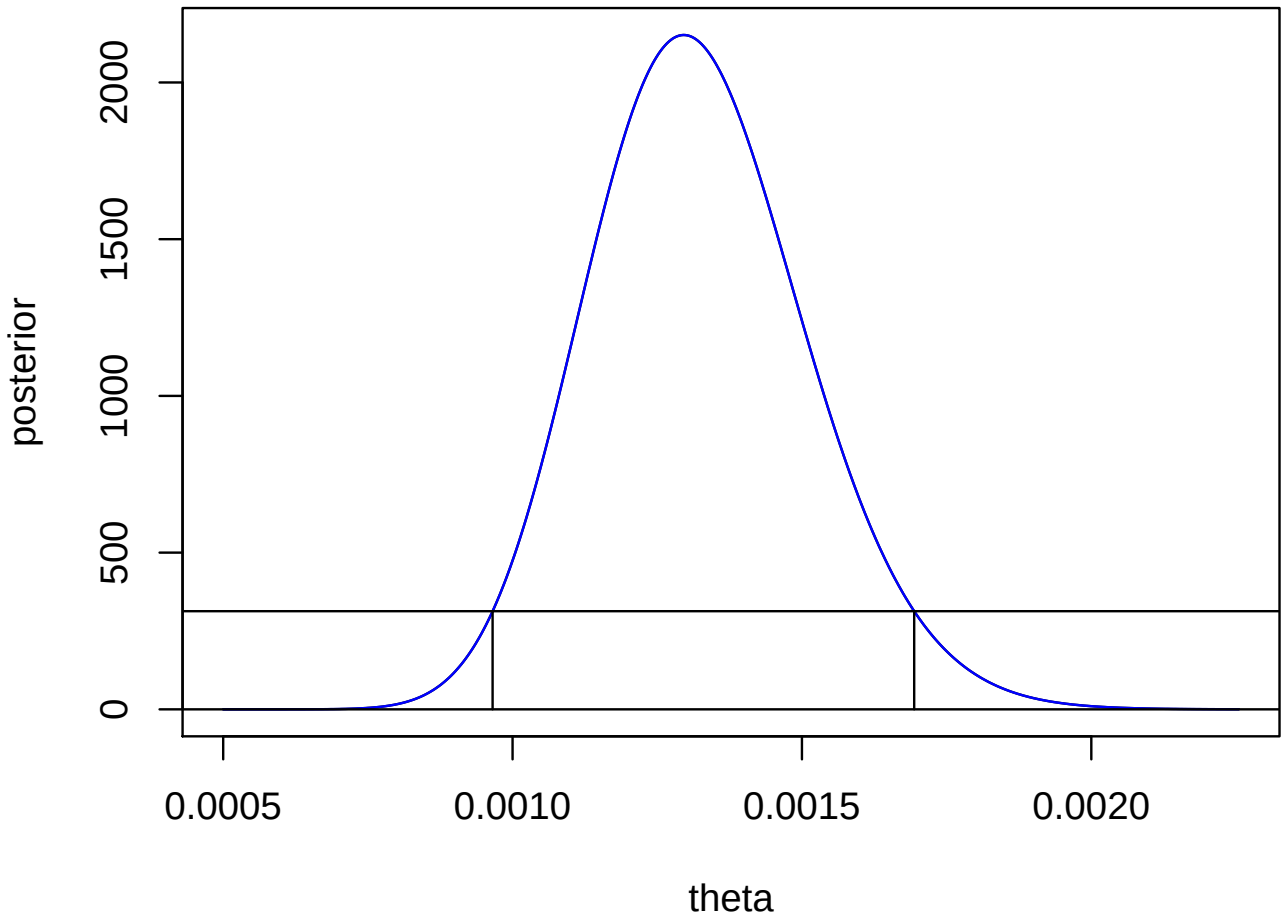
$$p(\theta|\bar{y}) = 250 \qquad (H)$$

were 0.001 and 0.0017.

The solutions to equation (H) were found to be 0.0009487 and 0.0017187, and the area between these two values is 0.9617. Since $1 - 0.9617 = 0.0383$ is smaller than 0.05, $k_{0.05}$ is bigger than 250.

Continuing to refine the choice of $\tilde{k}$, it was found that $k_{0.05} \approx 313.14$ and the 95% HPD interval is $(0.0009655, 0.0016940)$.

Illustration of 95% HPD region for Example 5



At the course website you can find the *R* function used to obtain this region.

Example 6 Normal data with unknown mean and known variance

Suppose $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, where $Y_1, \ldots, Y_n$ are i.i.d. as $N(\theta, \sigma^2)$, θ is unknown and σ^2 is known.

For a prior, we will use $p \equiv N(\mu_0, \sigma_0^2)$. *By taking σ_0^2 sufficiently large, this prior will be essentially noninformative.*

The posterior is

$$\begin{aligned} p(\theta|\boldsymbol{y}) &\propto \exp\left(-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right) \\ &\quad \times \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma_0^2}(\theta^2-2\mu_0\theta)\right) \\ &\quad \times \exp\left(-\frac{n}{2\sigma^2}(\theta^2-2\bar{y}\theta)\right). \end{aligned}$$

By completing the square in the exponent, we have

$$p(\theta|\boldsymbol{y}) \propto \exp\left(-\frac{1}{2\sigma_n^2}(\theta-\mu_n)^2\right),$$

where

$$\mu_n = \frac{\bar{y} + \mu_0\left(\frac{\sigma^2}{n\sigma_0^2}\right)}{1+\frac{\sigma^2}{n\sigma_0^2}}$$

and

$$\sigma_n^2 = \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}.$$

It follows that the posterior is $N(\mu_n, \sigma_n^2)$.

Of course, the mean and mode of the normal distribution are one and the same, and hence both of these Bayesian point estimates of θ are μ_n.

We have

$$\mu_n = w_n \bar{y} + (1 - w_n)\mu_0,$$

where

$$w_n = \left(1 + \frac{\sigma^2}{n\sigma_0^2}\right)^{-1},$$

and hence the posterior mean has the properties noted on p. 30N.

Note also that

$$\sigma_n^2 = w_n \cdot \frac{\sigma^2}{n}.$$

HPD region

Because the normal density is unimodal and symmetric about its mean, a $100(1-\alpha)\%$ HPD region for θ is

$$\mu_n \pm z_{\alpha/2}\sigma_n,$$

where z_p is the $(1-p)$th quantile of the standard normal distribution.

Regardless of how small n is, if we take σ_0 to be sufficiently big (meaning that the prior for θ is noninformative), then

$$\mu_n \approx \bar{y}, \quad \sigma_n \approx \frac{\sigma}{\sqrt{n}}$$

and the HPD region is approximately

$$\bar{y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

The interesting thing here is that the last HPD region has exactly the same form as the usual frequentist confidence interval in this setting.

The Jeffreys noninformative prior for the normal model under consideration is constant over the real line, and hence improper.

However, the corresponding posterior is $N(\bar{y}, \sigma^2/n)$, and thus proper. So, by using the improper prior, an HPD region is *precisely* the usual frequentist confidence interval.

Hypothesis testing

$Y_1, \ldots, Y_{100}$ are a random sample from $N(\theta,$ 1000). Want to test

$$H_0 : \theta < 20 \quad \text{vs.} \quad H_1 : \theta \geq 20.$$

The prior for θ is $N(24, 30)$. This means

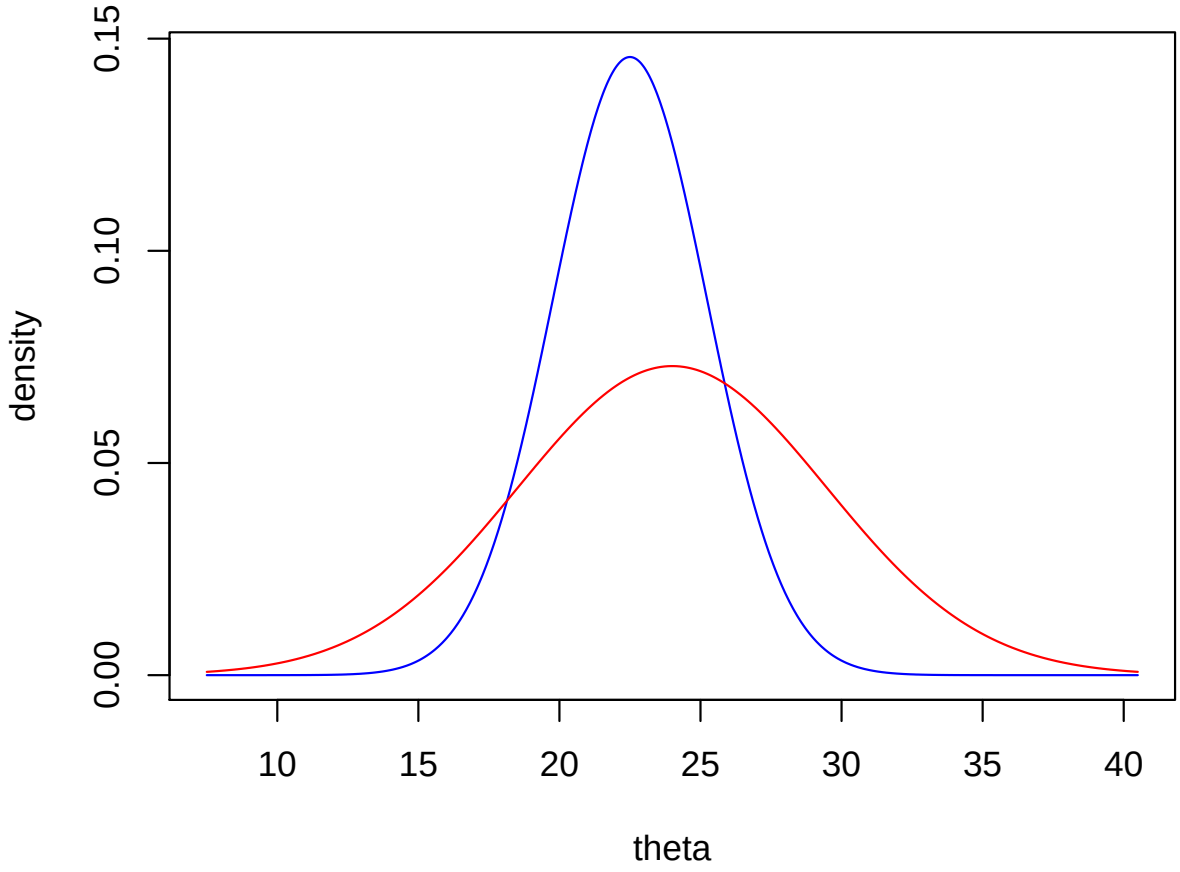
$$w_n = \left(1 + \frac{1000}{100(30)}\right)^{-1} = 3/4,$$

and hence the posterior is $N(6 + 3\bar{y}/4, 7.5)$.

Using the prior, one may check that the prior probabilities of H_0 and H_1 are $p_0 = 0.233$ and $p_1 = 0.767$.

Suppose $\bar{y}$ is observed to be 22. Then the posterior is $N(22.5, 7.5)$.

Prior and posterior for Example 6



Prior: Posterior:

The posterior probability of H_1 is

$$\begin{aligned}\alpha_1 &= P(\theta \geq 20|\bar{Y} = 22) \\ &= P\left(Z \geq \frac{20 - 22.5}{\sqrt{7.5}}\right) \\ &= 0.819.\end{aligned}$$

So, $\alpha_0 = 0.181$ and the posterior odds ratio is

$$\frac{\alpha_0}{\alpha_1} = \frac{0.181}{0.819} = 0.221.$$

The Bayes factor is

$$B = \frac{0.221}{0.233/0.767} = 0.727.$$

The odds of H_0 being true have declined in light of the data.

As a point of interest, note that the classical P-value in this example is

$$P\left(Z > \frac{22 - 20}{\sqrt{10}}\right) = 0.264.$$

Let's look more closely at how the posterior probability of H_0 compares with the P-value.

Suppose we're testing

$$H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0$$

in the setting of Example 6. We have

$$\begin{aligned} P(\theta < \theta_0 | \bar{y}) &= P\left(\frac{\theta - \mu_n}{\sigma_n} < \frac{\theta_0 - \mu_n}{\sigma_n} \Big| \bar{y} \right) \\ &= P\left(Z < \frac{\theta_0 - \mu_n}{\sigma_n} \right) \\ &= P\left(Z > \frac{\mu_n - \theta_0}{\sigma_n} \right). \end{aligned}$$

When n is large or we use the improper constant prior for θ, *the last quantity is exactly equal to the frequentist P-value*.

Example 7 Testing a point null hypothesis

Suppose we want to test

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0.$$

If θ is a continuous parameter, then continuous priors and posteriors assign probability 0 to θ_0 (and hence to H_0).

There are two ways around this problem:

- Change H_0 to $H_0 : \theta \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$.

- Assign a prior probability of $p_0 > 0$ to θ_0, and then distribute the remaining probability $1 - p_0$ over the other values of θ.

In the second case, let Π_1 be an absolutely continuous cdf over all of Θ, and let Π_0 be the cdf that assigns probability 1 to θ_0. Then the prior cdf, Π, is a mixture of Π_0 and Π_1:

$$\Pi(\theta) = p_0\Pi_0(\theta) + (1 - p_0)\Pi_1(\theta) \quad \forall\ \theta \in \Theta.$$

Letting π_1 be the density corresponding to Π_1, the posterior probability of H_0 is

$$\alpha_0 = P(\theta_0|\boldsymbol{Y} = \boldsymbol{y}) = \frac{p_0 f(\boldsymbol{y}|\theta_0)}{m(\boldsymbol{y})},$$

where

$$\begin{aligned} m(\boldsymbol{y}) &= p_0 f(\boldsymbol{y}|\theta_0) \\ &\quad + (1 - p_0) \int_\Theta f(\boldsymbol{y}|\theta)\pi_1(\theta)\, d\theta. \end{aligned}$$

Suppose now that $Y_1, \ldots, Y_n$ are a random sample from $N(\theta, \sigma^2)$, where θ is unknown and σ^2 known.

We want to test

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0,$$

and we take $p_0 = p_1 = 1/2$.

Let π_1 be $N(\theta_0, \sigma^2)$, and suppose

$$z = \frac{\bar{y} - \theta_0}{\sigma/\sqrt{n}} = 1.96.$$

The following table is from Table 4.2 in Berger, *Statistical Decision Theory and Bayesian Analysis* and applies to the above situation:

n	5	20	100	1000
α_0	0.33	0.42	0.60	0.80

So, in a situation where a frequentist would usually say that there is convincing evidence *against* H_0, the posterior probability of H_0 can actually be quite large when $p_0 = 1/2$.

In general, *if one truly believes that a point null hypothesis has nonzero probability, then the frequentist P-value often overstates the significance of the evidence against H_0.*

This is referred to as *Lindley's paradox*.

See Section 4.3.3 of Berger, *Statistical Decision Theory and Bayesian Analysis* for a detailed discussion of this issue.
