Non-informative Priors Multiparameter Models

Statistics 220

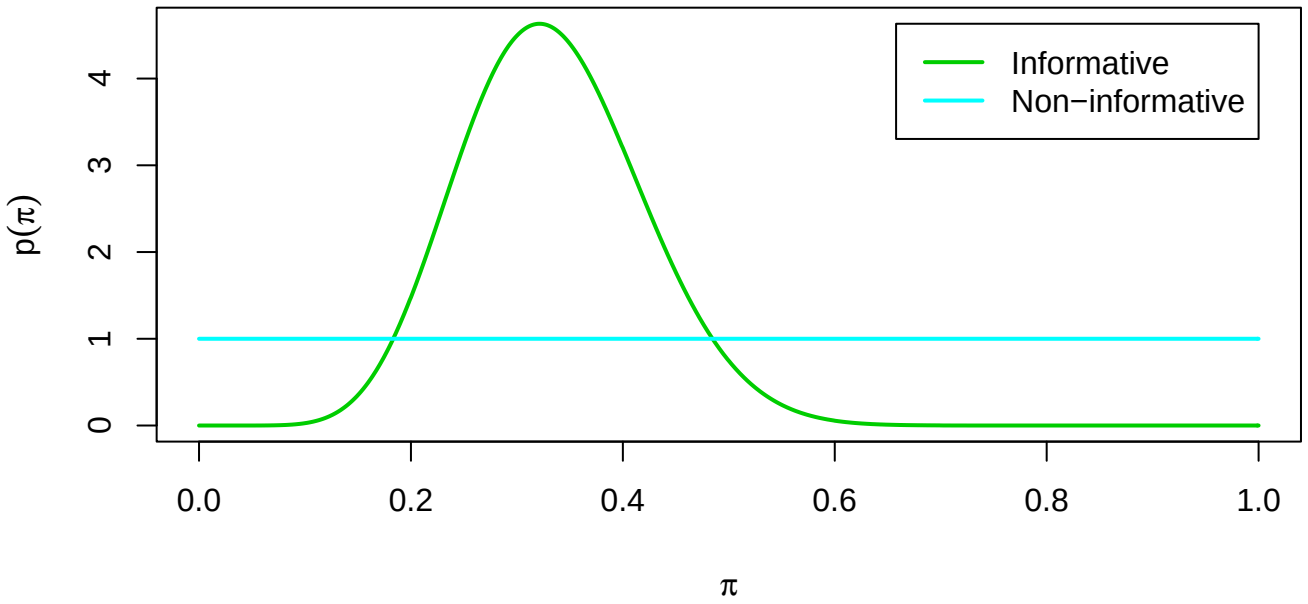
Spring 2005

Copyright ©2005 by Mark E. Irwin

Prior Types

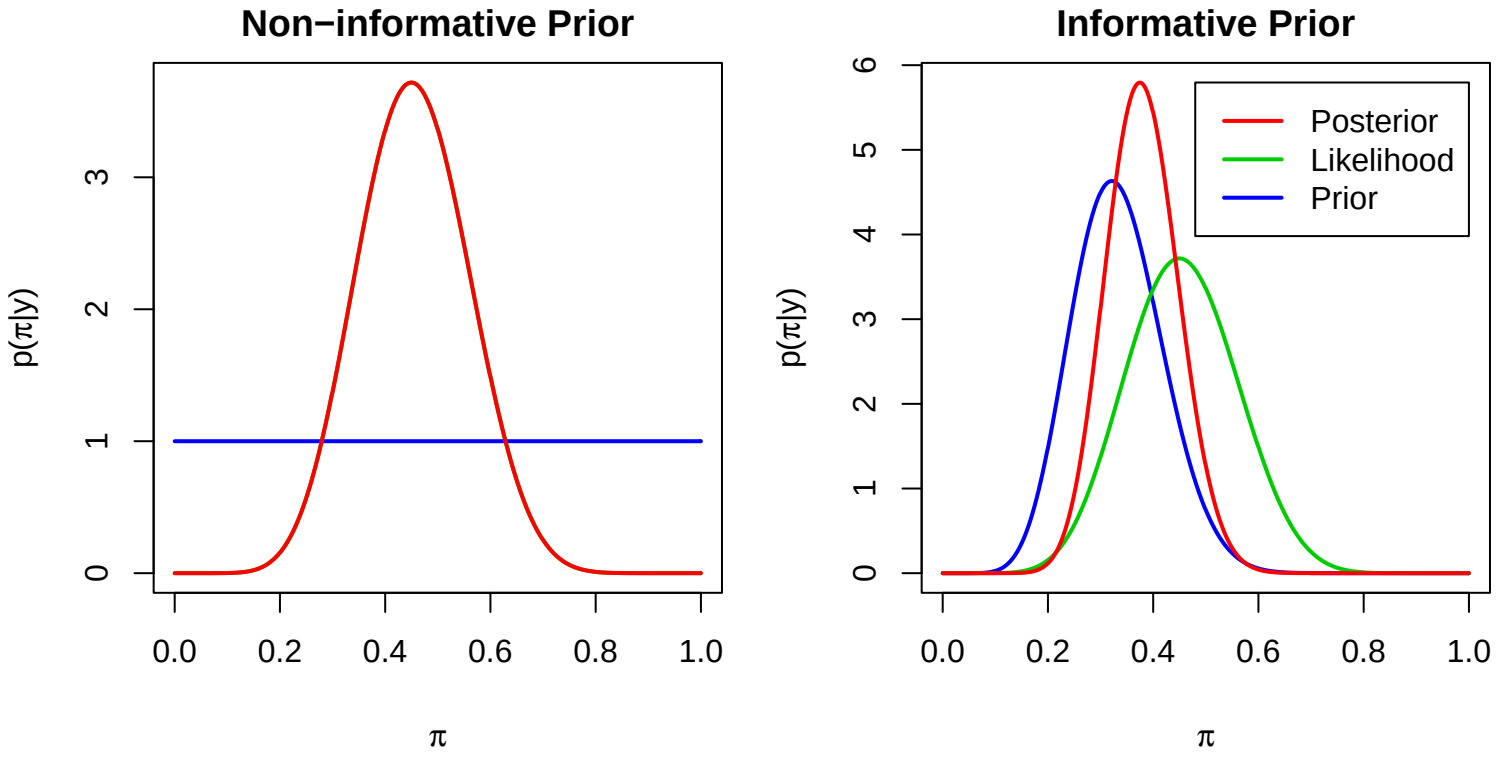
- Informative vs Non-informative

  There has been a desire for a prior distributions that play a minimal in the posterior distribution. These are sometime referred to a non-informative or reference priors.



These priors are often described as vague, flat, or diffuse.

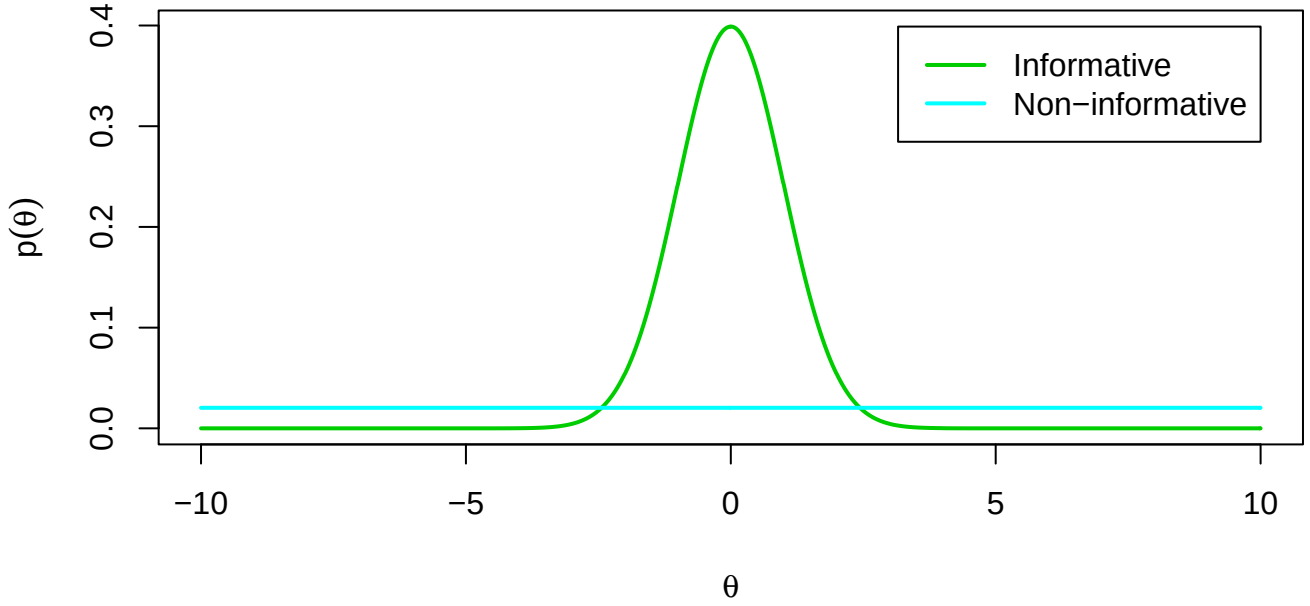
In the case when the parameter of interest exists on a bounded interval (e.g. binomial success probability π), the uniform distribution is an "obvious" non-informative prior.



For this example, with the non-informative prior, Posterior = Likelihood

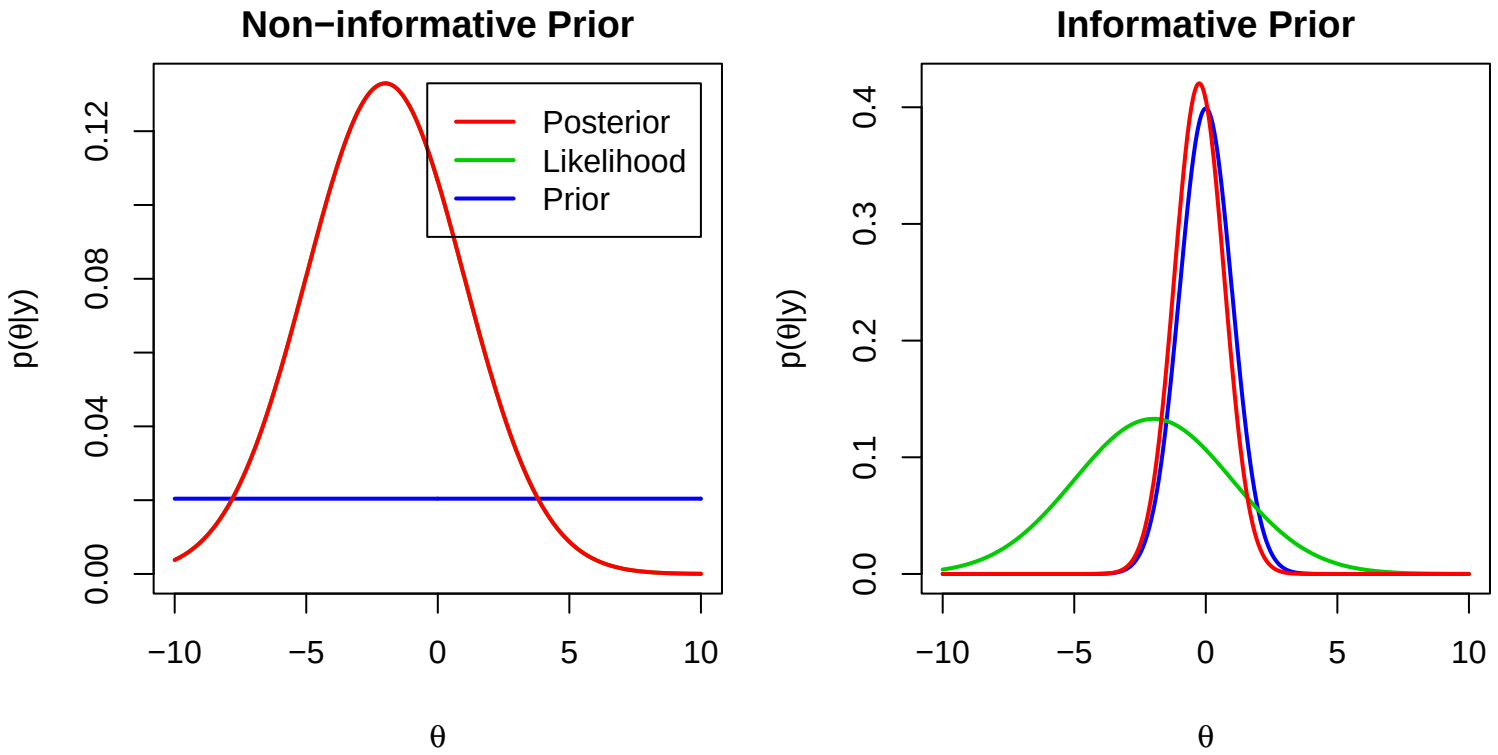
However for a parameter that occurs on an infinite interval (e.g. a normal mean θ), using a uniform prior on θ is problematic.

For the normal mean example, lets use the conjugate prior $N(\mu_0, \tau_0^2)$, but with a very big variance τ_0^2



The posterior mean and precision are

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$



So if we let $\tau_0^2 \to \infty$, then

$$\mu_n \to \bar{y} \quad \text{and} \quad \frac{1}{\tau_n^2} \to \frac{n}{\sigma^2}$$

This equivalent to the posterior being proportional to the likelihood, which is what we get if $p(\theta) \propto 1$ (e.g. uniform).

This does not describe a valid probability density as

$$\int_{-\infty}^{\infty} d\theta = \infty$$

- Proper vs Improper

  A prior is called proper if it is a valid probability distribution

$$p(\theta) \geq 0, \ \forall \theta \in \Theta \quad \text{and} \int_{\Theta} p(\theta) d\theta = 1$$

  (Actually all that is needed is a finite integral. Priors only need to be defined up to normalization constants.)

  A prior is called improper if

$$p(\theta) \geq 0, \ \forall \theta \in \Theta \quad \text{and} \int_{\Theta} p(\theta) d\theta = \infty$$

  If a prior is proper, so must the posterior.

If a prior is improper, the posterior often is, i.e.

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

is a proper distribution for all y. Note that an improper prior may lead to an improper prior. For many common problems, popular improper reference priors will usually lead to proper posteriors, assuming there is enough data.

For example

$$\begin{aligned} y_1, \ldots, y_n|\theta &\overset{iid}{\sim} N(\theta, \sigma^2) \\ p(\theta) &\propto 1 \end{aligned}$$

will have a proper posterior as long n is at least 1.

Non-informative Priors

While it may seem that picking a non-informative prior distribution might be easy, (e.g. just use a uniform), its not quite that straight forward.

Example: Normal observations with known mean, but unknown variance

$$\begin{aligned} y_1, \ldots y_n|\sigma &\overset{iid}{\sim} N(\theta, \sigma^2) \\ p(\sigma) &\propto 1 \end{aligned}$$

What is the equivalent prior on σ^2

Aside: Let θ be a random variable with density $p(\theta)$ and let $\phi = h(\theta)$ be a one-one transformation. Then the density of ϕ satisfies

$$f(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| = p(\theta)|h'(\theta)|^{-1} \quad \text{where } \theta = h^{-1}(\phi)$$

If $h(\sigma) = \sigma^2, h'(\sigma) = 2\sigma$, then a uniform prior on σ leads to

$$p(\sigma^2) = \frac{1}{2\sigma}$$

which clearly isn't uniform. This implies that our prior belief is that the variance should be small

Similarly, if there is a uniform prior on σ^2, the equivalent prior on σ is

$$p(\sigma) = 2\sigma$$

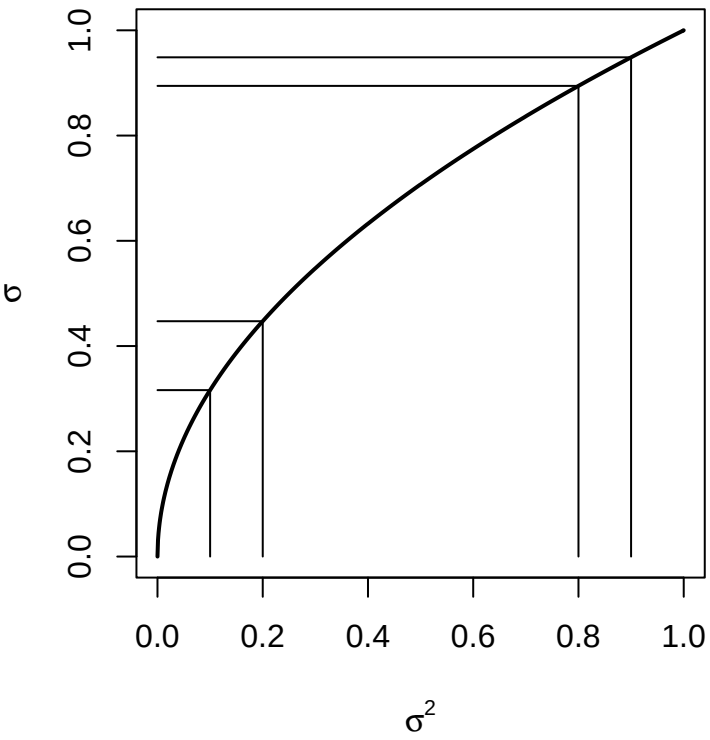
This implies that we believe sigma to be large.

One way to think about what is happening is to look at what happens to intervals of equal measure.

In the case σ^2 being uniform, an interval $[a, a + 0.1]$ must have the same prior measure as the interval $[0.1, 0.2]$.

When we transform to σ, the prior measure on it must have intervals $[\sqrt{a}, \sqrt{a + 0.1}]$ having equal measure.

But note that the length of the interval $[\sqrt{a}, \sqrt{a + 0.1}]$ is a decreasing function of a, which agrees with the increasing density in σ.



So when talking about non-informative priors you need to think about on what scale.

Jeffreys' Priors

Can we pick a prior where the scale the parameter is measured in doesn't matter.

Jeffreys' principle states that any rule for determining the prior density $p(\theta)$ should yield an equivalent result if applied to the transformed parameter.

That is applying

$$p(\phi) = p(\theta) \left| \frac{d\theta}{d\phi} \right| = p(\theta)|h'(\theta)|^{-1} \quad \text{where } \theta = h^{-1}(\phi)$$

should give the same answer as dealing directly with the transformed model

$$p(y, \phi) = p(\phi)p(y|\phi)$$

Applying this principle gives

$$p(\theta) = [J(\theta)]^{1/2}$$

where $J(\theta)$ is the $Fisher\ information$ for θ

$$J(\theta) = E\left[\left(\frac{d \log p(y|\theta)}{d\theta}\right)^2 |\theta\right] = -E\left[\frac{d^2 \log p(y|\theta)}{d\theta^2}|\theta\right]$$

Why does this work?

It can be shown that (see page 63)

$$J(\phi) = J(\theta)\left|\frac{d\theta}{d\phi}\right|^2$$

so

$$p(\phi) = p(\theta) \left| \frac{d\theta}{d\phi} \right|$$

For example, for the normal example with unknown variance, the Jeffreys' prior for the standard deviation σ is

$$p(\sigma) \propto \frac{1}{\sigma}$$

Alternative descriptions under different parameterizations for the variability are

$$\begin{aligned} p(\sigma^2) &\propto \frac{1}{\sigma^2} \\ p(\log \sigma^2) &\propto p(\log \sigma) \propto 1 \end{aligned}$$

For exponential data ($y_i \overset{iid}{\sim} Exp(\theta); \theta = \frac{1}{E[y|\theta]}$), the Jeffreys' prior is

$$p(\theta) = \frac{1}{\theta}$$

If you wish to parameterize in terms of the mean ($\lambda = \frac{1}{\theta}$), the Jeffreys' prior is

$$p(\lambda) = \frac{1}{\lambda}$$

For parameters with infinite parameter spaces (like a normal mean or variance), the Jeffrey's prior is often improper under the usual parameterizations.

As we have seen, different approaches may lead to different non-informative priors.

Pivotal Quantities

There are some situations where the common approaches give the same non-informative distributions.

- Location Parameter

  Suppose that the density of $p(y-\theta|\theta)$ is a function that is free of θ, call it $f(u)$. For example, if $y \sim N(\mu, 1)$,

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

  Then $y - \theta$ is known as a pivotal quantity and θ is known as a pure location parameter.

  In this situation, a reasonable approach would assume that a non-informative prior would give $f(y-\theta)$ as the posterior density of $y - \theta|y$.

This gives

$$p(y-\theta|y) \propto p(\theta)p(y-\theta|\theta)$$

which implies $p(\theta) \propto 1$ (i.e. θ is uniform)

- Scale parameters

  Suppose that the density of $p(y/\theta|\theta)$ is a function that is free of θ, call it $g(u)$. For example, if $y \sim N(0, \sigma^2)$,

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

  In this case y/θ is also a pivotal quantity and θ is known as a pure scale parameter.

If we follow the same approach as to above to where $g(y/\theta)$ as the posterior, this gives

$$p(\theta|y) = \frac{y}{\theta} p(y|\theta)$$

which implies $p(\theta) \propto \frac{1}{\theta}$

The standard deviation from a normal distribution and the mean of an exponential distribution are scale parameters.

Using the earlier result for the standard deviation, it implies that in some sense, the "right" scale for a scale parameter θ is $\log \theta$ as

$$\begin{aligned} p(\theta) &\propto \frac{1}{\theta} \\ p(\theta^2) &\propto \frac{1}{\theta^2} \\ p(\log \theta) &\propto 1 \end{aligned}$$

Note that pivotal quantities also come into standard frequentist inference. Examples involving $y_1, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$ are

$$\sqrt{n}\frac{\bar{y} - \mu}{s} \sim t_{n-1} \qquad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

The standard confidence intervals and hypothesis tests use the fact that these are pivotal quantities.

Multiparameter Models

Most analyzes we wish to perform involve multiple parameters

- $y_i \overset{iid}{\sim} N(\mu, \sigma^2)$
- Multiple Regression: $y_i|x_i \overset{ind}{\sim} N(x_i^t\beta, \sigma^2)$
- Logistic Regression: $y_i|x_i \overset{ind}{\sim} Bern(p_i)$ where $\text{logit}(p_i) = \beta_0 + \beta_1 x_i$

In these cases we want to assume all of the parameters are unknown and want to perform inference on some or all of them.

An example of the case, where only some of them may be of interest is multiple regression. Usually only the regression parameters β are of interest. The measurement variance σ^2 is often considered as a nuisance parameter.

Lets consider the case with two parameters θ_1 and θ_2 and that only θ_1 is of interest. An example of this would be $N(\mu, \sigma^2)$ data where $\theta_1 = \mu$ and $\theta_2 = \sigma^2$.

Want to base our inference on $p(\theta_1|y)$. We can get at this a couple of ways. First we can start with the joint posterior

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)p(\theta_1, \theta_2)$$

This gives

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)d\theta_2$$

We can also get it by

$$p(\theta_1|y) = \int p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2$$

This implies that distribution of θ_1 can be considered a mixture of the conditional distributions, averaged over the nuisance parameter.

Note that this marginal conditional distribution is often difficult to determine explicitly. Normally it needs to be examined by Monte Carlo methods.

Example: Normal Data

$$y_i \overset{iid}{\sim} N(\mu, \sigma^2)$$

For a prior, lets assume that μ and σ^2 are independent and use the standard non-informative priors

$$p(\mu, \sigma^2) = p(\mu)p(\sigma^2) \propto \frac{1}{\sigma^2}$$

So the joint posterior satisfies

$$\begin{aligned} p(\mu, \sigma^2) &\propto \frac{1}{\sigma^2} \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2 \right) \\ &= \frac{1}{\sigma^{n+2}} \exp\left(-\frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 \right] \right) \\ &= \frac{1}{\sigma^{n+2}} \exp\left(-\frac{1}{2\sigma^2} \left[(n-1)s^2 + n(\bar{y} - \mu)^2 \right] \right) \end{aligned}$$

where s^2 is the sample variance of the y_i's. Note that the sufficient statistics are $\bar{y}$ and s^2.

- The conditional distribution $p(\mu|\sigma, y)$

  Note that we have already derived this as this is just the fixed and known variance case. So

$$\mu|\sigma, y \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)$$

We can also get it by looking at the joint posterior. The only part that contains μ looks like

$$p(\mu|\sigma, y) \propto \exp\left(-\frac{n}{2\sigma^2}(\mu - \bar{y})^2\right)$$

which is proportional to a $N\left(\bar{y}, \frac{\sigma^2}{n}\right)$ density.

- The marginal posterior distribution $p(\sigma^2|y)$

  To get this, we must integrate μ out of the joint posterior.

$$
\begin{aligned}
p(\sigma^2|y) &\propto \int \frac{1}{\sigma^{n+2}} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) d\mu \\
&= \frac{1}{\sigma^{n+2}} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \int \exp\left(-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right) d\mu
\end{aligned}
$$

The piece left inside the integral is $\sqrt{2\pi\sigma^2/n}$ times the $N\left(\bar{y}, \frac{\sigma^2}{n}\right)$ density which gives

$$
\begin{aligned}
p(\sigma^2|y) &\propto \frac{1}{\sigma^{n+2}} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \sqrt{2\pi\sigma^2/n} \\
&\propto \frac{1}{(\sigma^2)^{(n+1)/2}} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right)
\end{aligned}
$$

Which is a scaled inverse-χ^2 density

$$\sigma^2|y \sim Inv-\chi^2(n-1, s^2)$$

A random variable $\theta \sim Inv-\chi^2(n-1, s^2)$ if

$$\frac{(n-1)s^2}{\theta} \sim \chi^2_{n-1}$$

Note that this result agrees with the standard frequentist result on the sample variance. However this shouldn't be surprising using the results on non-informative priors, particularly the result involving pivotal quantities.

- The marginal posterior distribution $p(\sigma^2|y)$

  Now that we have $p(\mu|\sigma^2, y)$ and $p(\sigma^2|y)$, inference on μ isn't difficult.

One method is to use the Monte Carlo approach discussed earlier

1. Sample σ_i^2 from $p(\sigma^2|y)$
2. Sample μ_i from $p(\mu|\sigma_i^2, y)$

Then $\mu_1, \ldots, \mu_m$ is a sample from $p(\mu|y)$.

Note that in this case, it is actually possible to derive the exact density of $p(\mu|y)$.

In this case

$$p(\mu|y) = \int p(\mu, \sigma^2|y) d\sigma^2$$

is tractable. With the substitution $z = \frac{A}{2\sigma^2}$ where $A = (n-1)s^2 + n(\bar{y} - \mu)^2$, leaves a integral involving the gamma density (see the book, page 76).

Cranking though this leaves

$$p(\mu|y) \propto \frac{1}{\left[1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right]^{n/2}}$$

a $t_{n-1}(\bar{y}, \frac{s^2}{n})$ density.

Or

$$\frac{\mu - \bar{y}}{s/\sqrt{n}}|y \sim t_{n-1}$$

which corresponds to the standard result used for inference on a population mean

$$\frac{\bar{y} - \mu}{s/\sqrt{n}}|\mu \sim t_{n-1}$$