Set 11 - Normal model

STAT 401 (Engineering) - Iowa State University

February 15, 2017

Outline

- Normal model with known variance
- Normal model with known mean
- Normal model

Normal model with known variance

Suppose $Y_i \stackrel{ind}{\sim} N(\mu, s^2)$ and we assume the default prior $p(\mu) \propto 1$.

This "prior" is actually not a distribution at all, since its integral is not finite. Nonetheless, we can still use it to derive a posterior.

If you work through the math (lots of algebra and a little calculus), you will find

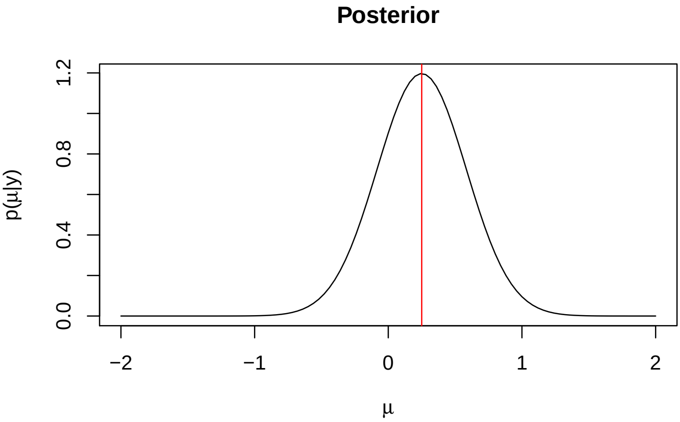
$$\mu|y \sim N(\overline{y}, s^2/n).$$

This looks exactly like the likelihood, but now it is normalized, i.e. it integrates to 1 and therefore it is a valid probability density function.

The Bayes estimator is

$$E[\mu|y] = \overline{y}.$$

```r
n <- 9
y <- rnorm(n) # default is mean is 0 and sd is 1
curve(dnorm(x, mean = mean(y), sd = sqrt(1/n)), -2, 2,
      xlab = expression(mu),
      ylab = expression(paste("p(",mu,"|y)")),
      main = "Posterior")
abline(v=mean(y), col='red')
```



Credible intervals

We can obtain credible intevals directly.

```
a <- .05
qnorm(c(a/2,1-a/2), mean(y), sd = sqrt(1/n))
```

```
[1] -0.4032876  0.9033550
```

Or we can use the fact that

$$\frac{\mu - \overline{y}}{s/\sqrt{n}} = Z \sim N(0,1)$$

to construct the interval using

$$\overline{y} \pm z_{.025} s/\sqrt{n}$$

where $a = \int_{z_a}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$, i.e. the area to the right of z_a under the pdf of a standard normal is a.

```
mean(y) + c(-1,1)*qnorm(.975)*sqrt(1/n)
```

```
[1] -0.4032876  0.9033550
```

Normal model with known mean

Suppose $Y_i \stackrel{ind}{\sim} N(m, \sigma^2)$ and we assume the default prior $p(\sigma^2) \propto \frac{1}{\sigma^2}\mathrm{I}(\sigma^2 > 0)$.

Again, this "prior" is actually not a distribution at all, since its integral is not finite. Nonetheless, we can still use it to derive a posterior.

If you work through the math (lots of algebra and a little calculus), you will find

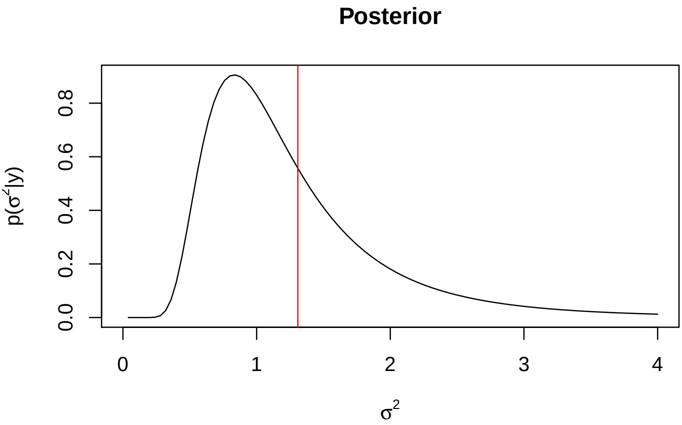
$$\sigma^2|y \sim IG\left(\frac{n}{2}, \frac{\sum_{i=1}^n (y_i - m)^2}{2}\right)$$

where IG indicates an inverse gamma distribution.

The Bayes estimator is

$$E[\sigma^2|y] = \frac{\frac{\sum_{i=1}^n (y_i - m)^2}{2}}{\frac{n}{2} - 1} = \frac{\sum_{i=1}^n (y_i - m)^2}{n-2} \text{ for } n > 2$$

```r
s <- sum((y-0)^2)
curve(MCMCpack::dinvgamma(x, shape = n/2, scale = s/2), 0, 4,
      xlab = expression(sigma^2),
      ylab = expression(paste("p(",sigma^2,"|y)")),
      main = "Posterior")
abline(v = (s/2)/((n/2)-1), col='red')
```



Credible intervals

We don't have a quantile function for this inverse gamma distribution. So we'll obtain estimates of the interval endpoints by taking a bunch of simulated draws from the inverse gamma distribution and finding their sample quantiles.

```
draws <- MCMCpack::rinvgamma(1e5, shape = s/2, scale = n/2)
quantile(draws, c(a/2, 1-a/2))


     2.5%     97.5%
0.4671976 3.2267781
```

If you don't have the MCMCpack library, you can draw from the gamma distribution and then invert the draws. It is slightly confusing because the 'scale' parameter for the inverse gamma is the 'rate' parameter for the gamma.

```
draws <- rgamma(1e5, shape = s/2, rate = n/2)
quantile( 1/draws, c(a/2, 1-a/2))


     2.5%     97.5%
0.4686875 3.2518544
```

Normal model

Suppose $Y_i \stackrel{ind}{\sim} N(\mu, \sigma^2)$ and we assume the default prior $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}\mathrm{I}(\sigma^2 > 0)$.

Again, this "prior" is actually not a distribution at all, since its integral is not finite. Nonetheless, we can still use it to derive a posterior.

If you work through the math (lots of algebra and a little calculus), you will find

$$\begin{aligned} \mu|\sigma^2, y &\sim N(\overline{y}, \sigma^2/n) \\ \sigma^2|y &\sim IG\left(\frac{n-1}{2}, \frac{\sum_{i=1}^n (y_i - \overline{y})^2}{2}\right) \end{aligned}$$

The joint posterior is obtained using

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

The Bayes estimator is

$$\begin{aligned} E[\mu|y] &= \overline{y} \\ E[\sigma^2|y] &= \frac{\frac{\sum_{i=1}^n (y_i - \overline{y})^2}{2}}{\frac{n-1}{2} - 1} = \frac{\sum_{i=1}^n (y_i - \overline{y})^2}{n-3} \text{ for } n > 3 \end{aligned}$$

Focusing on μ

Typically, the main quantity of interest in the normal model is the mean, μ. Thus, we are typically interested in marginal posterior for μ:

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

If

$$\mu|\sigma^2, y \sim N(\overline{y}, \sigma^2/n) \quad \text{and} \quad \sigma^2|y \sim IG\left(\frac{n-1}{2}, \frac{\sum_{i=1}^n (y_i - \overline{y})^2}{2}\right),$$

then

$$\mu|y \sim t_{n-1}(\overline{y}, S^2/n) \quad \text{where} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \overline{y})^2$$

that is, $\mu|y$ has a t distribution with $n-1$ degrees of freedom, location parameter $\overline{y}$ and scale parameter S^2/n.

t distribution

Definition

A t distributed random variable, $T \sim t_v(m, s^2)$ has probability density function

$$f_T(t) = \frac{\Gamma([v+1]/2)}{\Gamma(v/2)\sqrt{v\pi}s}\left(1 + \frac{1}{v}\left[\frac{x-m}{s}\right]^2\right)^{-(v+1)/2}$$

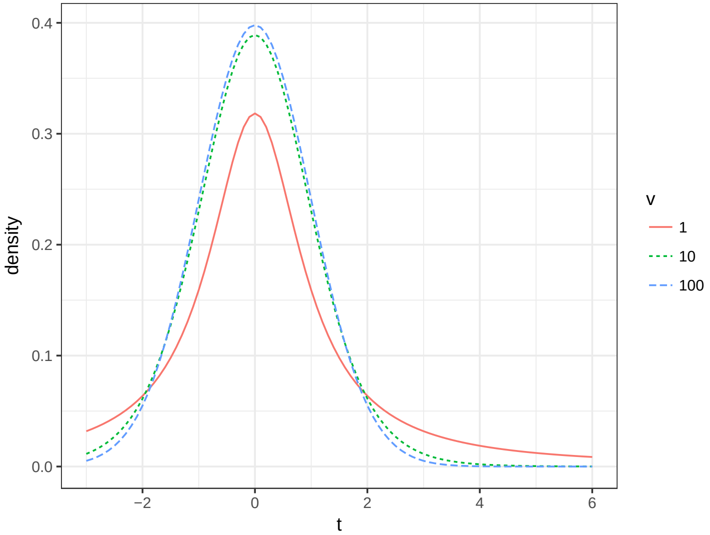
with degrees of freedom v, location m, and scale s^2. It has

$$\begin{array}{rll} E[T] & = m & v > 1 \\ Var[T] & = s^2\frac{v}{v-2} & v > 2. \end{array}$$

In addition,

$$t_v(m, s^2) \stackrel{d}{\to} N(m, s^2) \quad \text{as} \quad v \to \infty.$$

t distribution as v changes



Credible intervals

In R, there is no way to obtain t credible intervals directly. Thus we can use the fact that

$$\frac{\mu - \overline{y}}{S/\sqrt{n}} = t \sim t_{n-1}(0,1)$$

to construct the interval using

$$\overline{y} \pm t_{n-1,.025} S/\sqrt{n}$$

where the area to the right of $t_{n-1,a}$ under the pdf of a standard t is a.

```
mean(y) + c(-1,1)*qt(.975, df=n-1)*sd(y)/sqrt(n)


[1] -0.546741  1.046808
```

Corn yield

In evaluating corn yield for a particular year, the yield on a number of fields is measured. (For simplicity, assume that fields are standardized in size.) We measure 100 randomly selected fields in Iowa and find the average is 198 bushels per acre and the sample standard deviation is 20 bushels per acre. Provide a 90% credible interval for the mean yield across all fields in Iowa.

Let Y_i be the yield in field i and assume

$$Y_i \stackrel{ind}{\sim} N(\mu, \sigma^2).$$

If we assume the default prior $p(\mu, \sigma^2) \propto 1/\sigma^2$, then we have

$$\mu|y \sim t_{n-1}(\overline{y}, S^2/n)$$

where $n = 100$, $\overline{y} = 198$, and $S = 20$. A 90% interval is

```
n    <- 100
ybar <- 198
s    <- 20
a    <- 0.1

ybar +c(-1,1)*qt(1-a/2, df=n-1)*s/sqrt(n)


[1] 194.6792 201.3208
```