I04 - Normal model

STAT 401 (Engineering) - Iowa State University

February 15, 2018

Outline

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

Corn yield

For the following examples, we will consider measuring corn yield on fields. We will base our analyses on the following values:

- Mean yield per field is 200 bushels per acre
- Standard deviation of yield per field is 20 bushels per acre

In the following analyses, we will be assuming

- **Population mean is unknown** while population SD is known to be 20
- Population mean is known to be 200 while **population SD is unknown**
- **Both are population mean and population SD are unknown**

Normal model with known population variance

Suppose $Y_i \stackrel{ind}{\sim} N(\mu, v^2)$ 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

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

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.

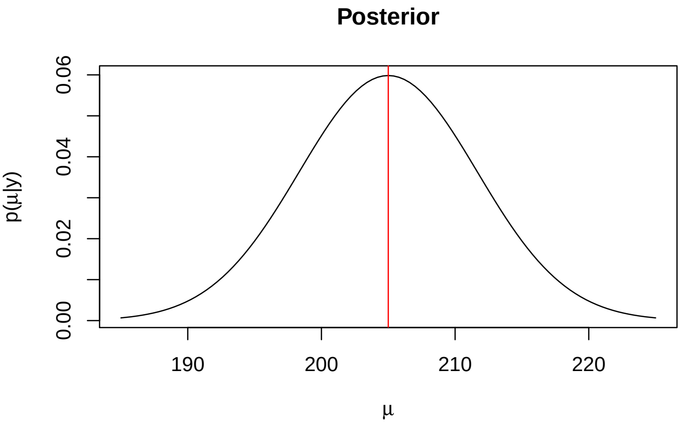
The Bayes estimator is

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

```
m <- 200
v <- 20
```

```
n <- 9
y <- rnorm(n, mean = m, sd = v); mean(y); sd(y)


[1] 205.0007
[1] 20.73131
```



Credible intervals

We can obtain credible intevals directly.

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

```
[1] 191.9342 218.0671
```

Or we can use the fact that

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

to construct the interval using

$$\overline{y} \pm z_{a/2} s/\sqrt{n}$$

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

```
mean(y) + c(-1,1)*qnorm(1-a/2)*v/sqrt(n) # equivalently mean(y) + qnorm(c(a/2,1-a/2))*v/sqrt(n)
```

```
[1] 191.9342 218.0671
```

Normal model with known population 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

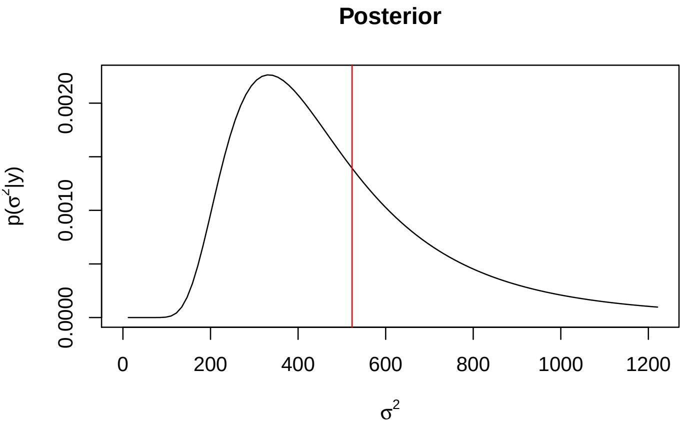
$$\sigma^2|y \sim IG\left(\frac{n}{2}, \frac{\sum_{i=1}^n (y_i - m)^2}{2}\right)$$

where IG indicates an inverse gamma distribution.

The Bayes estimator is

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

```
SS <- sum((y-m)^2)
curve(MCMCpack::dinvgamma(x, shape = n/2, scale = SS/2), 0, 3*SS/n,
      xlab = expression(sigma^2),
      ylab = expression(paste("p(",sigma^2,"|y)")),
      main = "Posterior")
abline(v = (SS/2)/((n/2)-1), col='red')
```



Credible intervals for variance - exact

For some reason, nobody has created a function to calculate the quantiles of an inverse gamma. So here is one

```
qinvgamma <- function(p, shape, scale = 1) {
  1/qgamma(1-p, shape = shape, rate = scale)
}
```

This function is slightly confusing because the ‘scale‘ parameter for the inverse gamma is the ‘rate‘ parameter for the gamma.
Now we can use this to calculate our credible intervals

```
(q <- qinvgamma(c(.025,.975), shape = n/2, scale = SS/2))

[1]  192.5775 1356.6034
```

Credible intervals for variance - simulation

We can also 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 = n/2, scale = SS/2)
quantile(draws, c(a/2, 1-a/2))


     2.5%     97.5%
 192.6423 1353.9686
```

If you don't have the MCMCpack library, you can draw from the gamma distribution and then invert the draws (which is the same trick that is used for the qinvgamma function).

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


     2.5%     97.5%
 193.2657 1364.9407
```

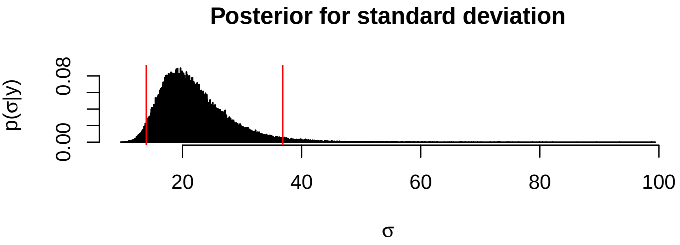
These are both Monte Carlo estimates of the true credible intervals and these estimates improve as the number of simultions increase.

Posterior and credible intervals for standard deviation

```
sqrt(q) # Take square root of end points of the CI for the variance to get the exact intervals

[1] 13.87723 36.83210

hist(sqrt(draws), 1001, xlab = expression(sigma), ylab = expression(paste("p(",sigma,"|y)")),
     main = "Posterior for standard deviation", probability = TRUE)
abline(v=sqrt(q), col="red")
```



There is actually a more sophisticated way to do this via transformations. You can learn this technique in STAT 447.

Normal model (unknown population mean and variance)

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 the 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

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

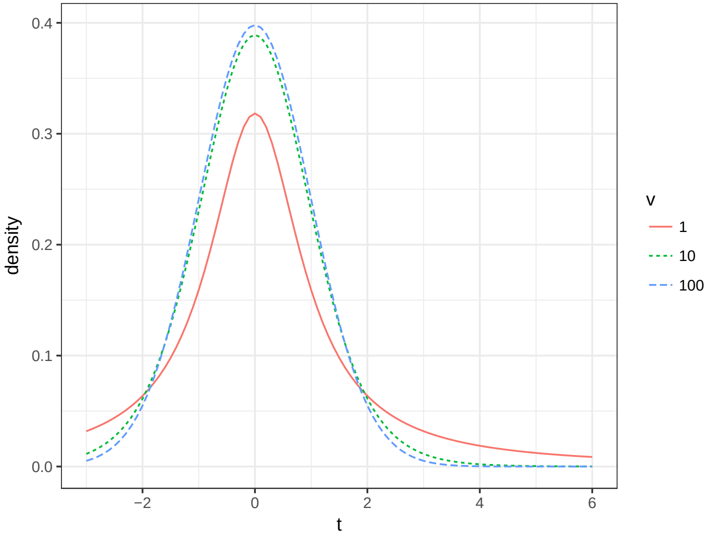
with degrees of freedom v, location m, and scale s^2. It has

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

In addition,

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

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

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

to construct the interval using

$$\overline{y} \pm t_{n-1,a/2} s/\sqrt{n}$$

where the area to the right of $t_{n-1,a/2}$ under the pdf of a standard t is $a/2$.

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

[1] 189.0652 220.9362
```

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 9 randomly selected fields in Iowa and find the sample average is 205 bushels per acre and the sample standard deviation is 21 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

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

A 90% interval is

```
a    <- 0.1
mean(y) +c(-1,1)*qt(1-a/2, df=n-1)*sd(y)/sqrt(n)
```

```
[1] 192.1504 217.8510
```

Informative Bayesian analysis when pop. variance is known

Let Y_i be the corn yield (in bushels/ac) from field i. Assume

$$Y_i \stackrel{ind}{\sim} N(\mu, v^2) \quad \text{and} \quad \mu \sim N(m, C)$$

where m provides your prior guess about the mean yield (not the population mean as was used previously in this slide set) and C provides your variance around that guess. Then

$$\begin{array}{rl} \mu|y & \sim N(m', C') \\ C' & = \left[\frac{1}{C} + \frac{n}{v^2}\right]^{-1} \\ m' & = C'\left[\frac{1}{C}m + \frac{n}{v^2}\overline{y}\right] = \frac{1/C}{1/C+n/v^2}m + \frac{n/v^2}{1/C+n/v^2}\overline{y} \end{array}$$

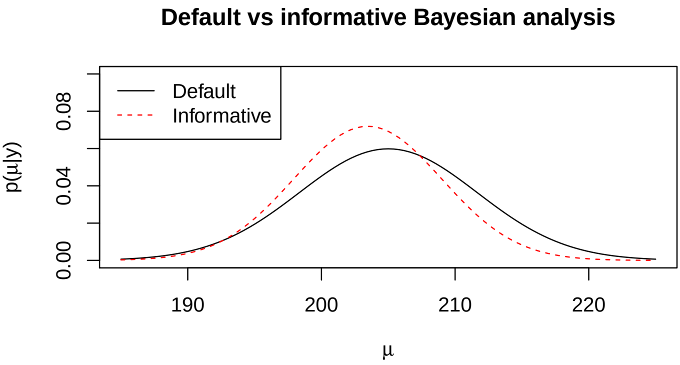
```
m = 200; C = 10^2
```

```
Cp = 1/(1/C+n/v^2); mp = Cp*(m/C+n*mean(y)/v^2)
```

So if we assume $m = 200$ and $C = 10^2$ and combine this with our observed data $n =$9 and $\overline{y} =$205 with population sd known to be $v =$20, then we have the posterior $\mu|y \sim N(203, 6^2)$.

Comparison of default vs informative Bayesian analysis

```
ybar = mean(y); se = v/sqrt(n)
curve(dnorm(x, mean=ybar, sd=se), ybar-3*se, ybar+3*se,
      ylim=c(0,.1),
      xlab=expression(mu),
      ylab=expression(paste("p(",mu,"|y)")),
      main="Default vs informative Bayesian analysis")
curve(dnorm(x, mean=mp, sd=sqrt(Cp)), col='red', lty=2, add=TRUE)
legend("topleft", c("Default","Informative"), col=c("black","red"),
       lty = 1:2)
```



Informative Bayesian analysis

The joint conjugate prior for μ and σ^2 is

$$\mu|\sigma^2 \quad \sim N(m, \sigma^2/k) \qquad \sigma^2 \quad \sim IG(d/2, dv^2/2)$$

where v^2 serves as a prior guess about σ^2 and d controls how certain we are about that guess.

The posterior under this prior is

$$\mu|\sigma^2, y \sim N(m', \sigma^2/k') \qquad \sigma^2|y \sim IG(d'/2, d'(v')^2/2)$$

where

$$\begin{aligned} k' &= k + n \\ m' &= [km + n\overline{y}]/k' \\ d' &= d + n \\ d'(v')^2 &= dv^2 + (n-1)s^2 + \tfrac{kn}{k'}(\overline{y} - m)^2 \end{aligned}$$