

## I04 - Normal model

STAT 401 (Engineering) - Iowa State University

February 13, 2018

# Outline

- Normal model with known variance
- Normal model with known 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

- **Mean is unknown** while SD is known to be 20
- Mean is known to be 200 while **SD is unknown**
- **Both are mean and standard deviation are unknown**

# Normal model with known variance

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

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

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

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

```

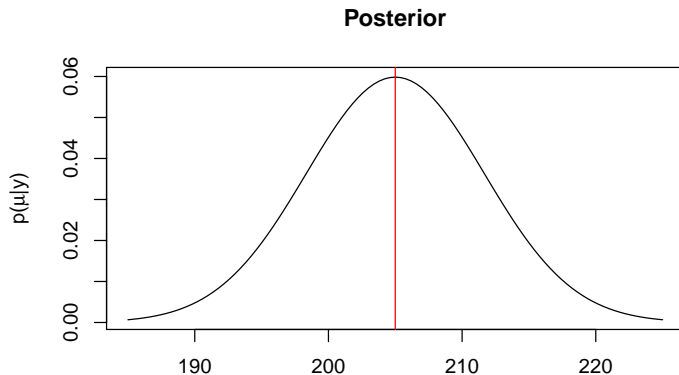
m <- 200
s <- 20
n <- 9
y <- rnorm(n, mean = m, sd = s)

```

```

curve(dnorm(x, mean = mean(y), sd = s/sqrt(n)), mean(y)-3*s/sqrt(n), mean(y)+3*s/sqrt(n),
      xlab = expression(mu),
      ylab = expression(paste("p(", mu, "| y)")),
      main = "Posterior")
abline(v=mean(y), col='red')

```



# Credible intervals

We can obtain credible intervals 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

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

to construct the interval using

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

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)*s/sqrt(n)
```

```
[1] 191.9342 218.0671
```

## Normal model with known mean

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

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

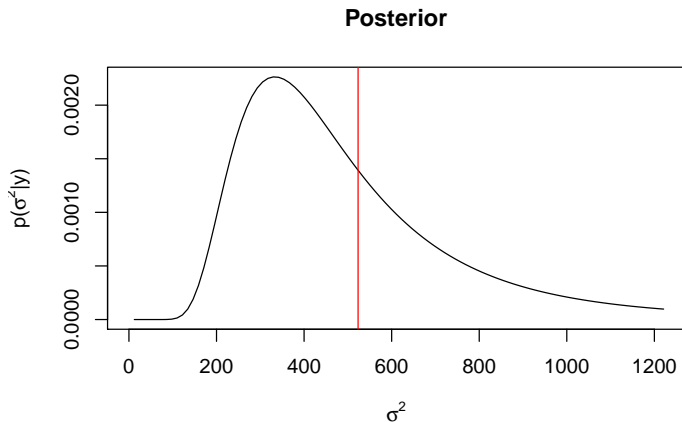
$$\sigma^2 | y \sim IG \left( \frac{n}{2}, \frac{\sum_{i=1}^n (y_i - m)^2}{2} \right)$$

where  $IG$  indicates an inverse gamma distribution.

The Bayes estimator is

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

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

```
      2.5%      97.5%
193.2657 1364.9407
```

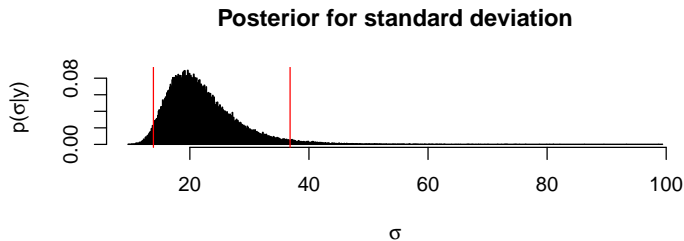
These are both close to the exact values.

# Posterior and credible intervals for standard deviation

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

```
[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 mean and variance)

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

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

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

The joint posterior is obtained using

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

The Bayes estimator is

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

## Focusing on $\mu$

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

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

If

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

then

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

that is,  $\mu|y$  has a  $t$  distribution with  $n-1$  degrees of freedom, location parameter  $\bar{y}$  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

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

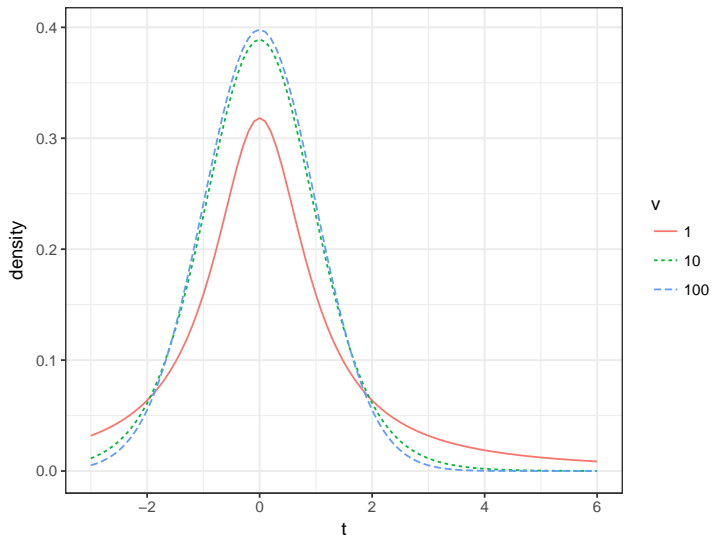
with degrees of freedom  $v$ , location  $m$ , and scale  $s^2$ . It has

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

In addition,

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

# $t$ 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

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

to construct the interval using

$$\bar{y} \pm t_{n-1, a/2} S/\sqrt{n}$$

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

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

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

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

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 variance is known

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

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

Then

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

```
m = 200
C = 33^2
Cp = 1/(1/C+n/s^2)
mp = Cp*(m/C+n*mean(y)/s^2)
```

So if we assume  $m = 200$  and  $C = 33^2$  and combine this with our observed data  $n = 9$  and  $\bar{y} = 205$ , then we have the posterior  $\mu|y \sim N(205, 7^2)$ .

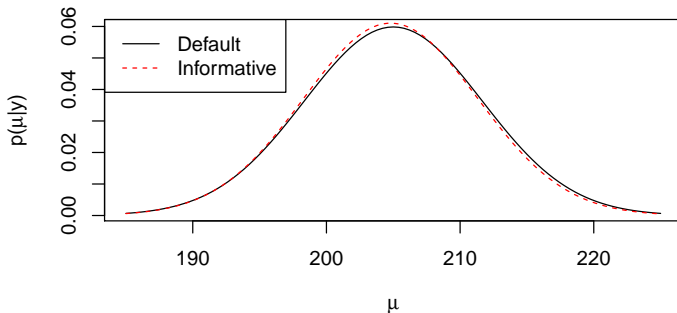
# Comparison of default vs informative Bayesian analysis

```

ybar = mean(y)
se = s/sqrt(n)
curve(dnorm(x, mean=ybar, sd=se), ybar-3*se, ybar+3*se,
      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)

```

**Default vs informative Bayesian analysis**



# Informative Bayesian analysis

The joint conjugate prior for  $\mu$  and  $\sigma^2$  is

$$\mu|\sigma^2 \sim N(m, \sigma^2/k) \quad \sigma^2 \sim \text{Inv-}\chi^2(v, s^2)$$

where  $s^2$  serves as a prior guess about  $\sigma^2$  and  $v$  controls how certain we are about that guess.

The posterior under this prior is

$$\mu|\sigma^2, y \sim N(m', \sigma^2/k') \quad \sigma^2|y \sim \text{Inv-}\chi^2(v', (s')^2)$$

where

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