I4 - Bayesian parameter estimation in a normal model

STAT 587 (Engineering)
Iowa State University

September 17, 2020

Bayesian parameter estimation

Recall that Bayesian parameter estimation involves

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

with

- posterior $p(\theta|y)$
- prior $p(\theta)$
- model $p(y|\theta)$
- prior predictive $p(y)$

For this video, $\theta = (\mu, \sigma^2)$ and

$$y|\mu, \sigma^2 \sim N(\mu, \sigma^2).$$

Bayesian parameter estimation in a normal model

Let $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$ and the default prior

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

Note: This "prior" is not a distribution since its integral is not finite. Nonetheless, we can still derive the following posterior

$$\mu|y \sim t_{n-1}(\overline{y}, s^2/n) \qquad \text{and} \qquad \sigma^2|y \sim IG\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$$

where

- n is the sample size,
- $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean, and
- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$ is the sample variance.

Posterior for the mean

The posterior for the mean is

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

and from properties of the generalized Student's t distribution, we know

- $E[\mu|y] = \overline{y}$ for $n > 2$,
- $Var[\mu|y] = \frac{(n-1)s^2}{(n-3)} \Big/ n$ for $n > 3$,

and

$$\frac{\overline{y} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

Credible intervals for μ

Since

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

a $100(1-a)\%$ equal-tail credible interval is

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

where $t_{n-1,a/2}$ is a t critical value such that $P(T_{n-1} < t_{n-1,a/2}) = 1 - a/2$ when $T_{n-1} \sim t_{n-1}$.

For example, $t_{10-1,0.05/2}$ is

```
n = 10
a = 0.05 # 95\% CI
qt(1-a/2, df = n-1)
```

```
[1] 2.262157
```

Posterior for the variance

The posterior for the mean is

$$\sigma^2|y \sim IG\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$$

and from properties of the inverse Gamma distribution, we know

- $E[\sigma^2|y] = \frac{(n-1)s^2}{n-3}$ for $n > 3$,

and

$$\frac{1}{\sigma^2}\bigg| y \sim Ga\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$$

where $(n-1)s^2/2$ is the rate parameter.

Credible intervals for σ^2

For a $100(1-a)\%$ credible interval, we need

$$a/2 = P(\sigma^2 < L|y) = P(\sigma^2 > U|y).$$

To do this, we will find

$$a/2 = P\left(\frac{1}{\sigma^2} > \frac{1}{L}\middle| y\right) = P\left(\frac{1}{\sigma^2} < \frac{1}{U}\middle| y\right).$$

Here is a function that performs this computation

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

Yield data

Suppose we have a random sample of 9 Iowa farms and we obtain corn yield in bushels per acre on those farms. We are interested in making statements about the mean yield and the variability in yield for Iowa farms.

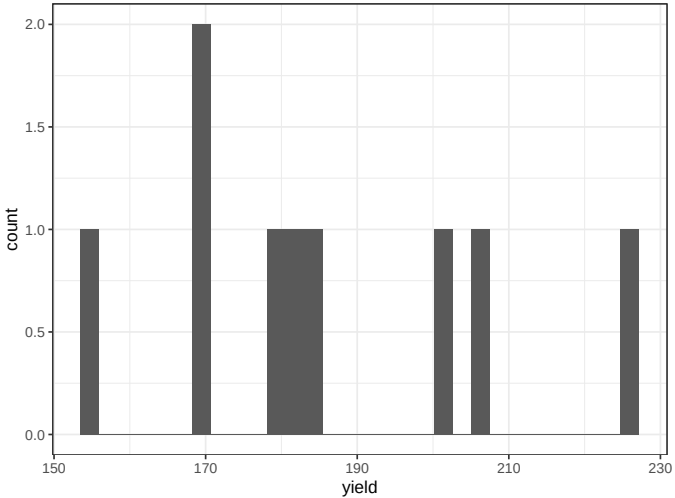
```
yield_data <- read.csv("yield.csv")
nrow(yield_data)
```

```
[1] 9
```

```
yield_data
```

```
   farm    yield
1 farm1 153.5451
2 farm2 205.6999
3 farm3 178.7548
4 farm4 170.1692
5 farm5 224.7723
6 farm6 184.0806
7 farm7 169.8615
8 farm8 201.2721
9 farm9 181.6356
```

Histogram of yield



Calculate sufficient statistics

```
n               = length(yield_data$yield); n
```

```
[1] 9
```

```
sample_mean     = mean(yield_data$yield);   sample_mean
```

```
[1] 185.5323
```

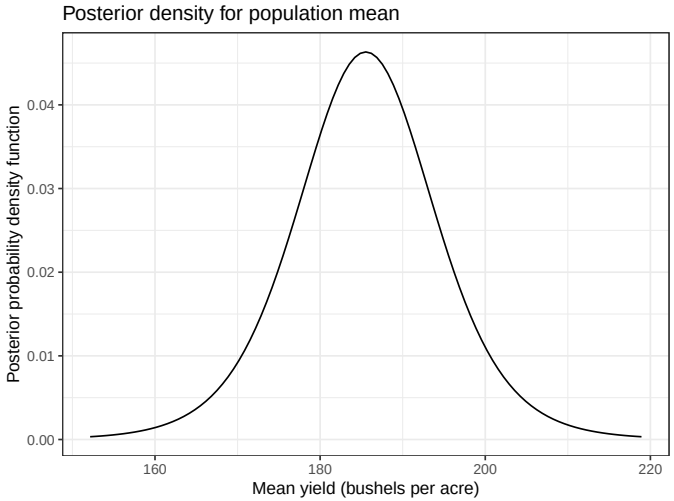
```
sample_variance = var(yield_data$yield);    sample_variance
```

```
[1] 470.2817
```

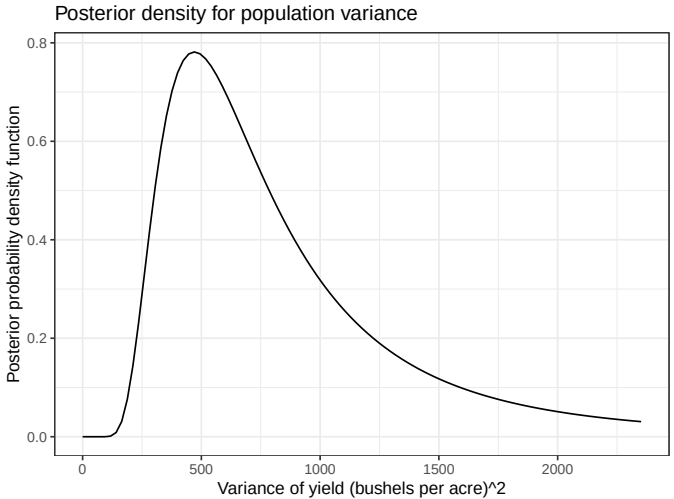
Use these sufficient statistics to calculate:

- posterior densities
- posterior means
- credible intervals

Posterior density for μ



Posterior density for σ^2



Posterior means

```
# Posterior mean of population yield mean, E[mu|y]
sample_mean
```

```
[1] 185.5323
```

Posterior mean for μ is $E[\mu|y] = 186$.

```
# Posterior mean of population yield variance
post_mean_var = (n-1)*sample_variance / (n-3)
post_mean_var
```

```
[1] 627.0422
```

Posterior mean for σ^2 is $E[\sigma^2|y] = 627$.

Credible intervals

```
# 95% credible interval for the population mean
a = 0.05
mean_ci = sample_mean + c(-1,1) *  qt(1-a/2, df = n-1) * sqrt(sample_variance/n)
mean_ci
```

```
[1] 168.8630 202.2017
```

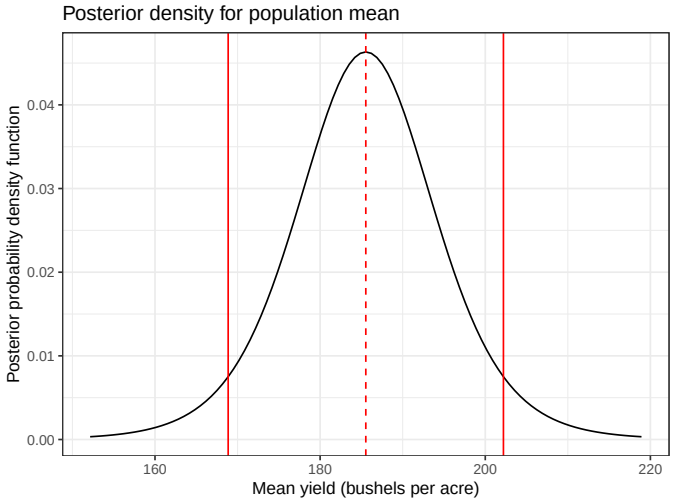
So a 95% credible interval for μ is (169,202).

```
# 95% credible interval for the population variance
var_ci = qinvgamma(c(a/2, 1-a/2),
                   shape = (n-1)/2,
                   scale = (n-1)*sample_variance/2)
var_ci
```

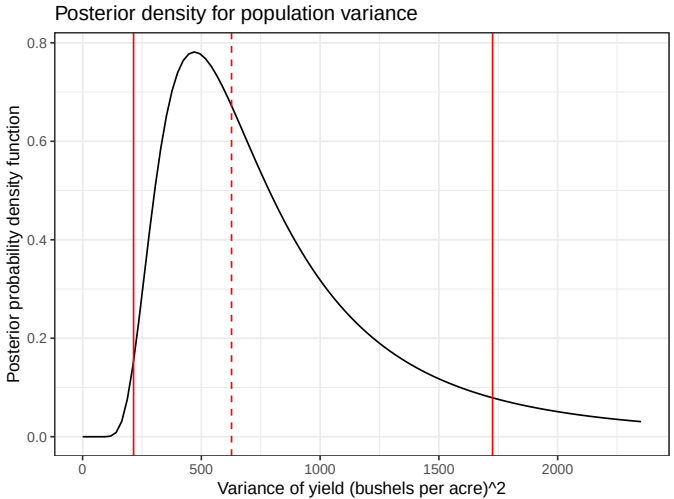
```
[1]  214.5623 1726.0175
```

So a 95% credible interval for σ^2 is (215,1726).

Posterior density for μ



Posterior density for σ^2



Posterior for the standard deviation, σ

We found

- $E[\sigma^2|y] = 627$
- 95% CI for σ^2 is $(215, 1726)$

but these values have units of (bushels/acre)^2.

The standard deviation $\sigma = \sqrt{\sigma^2}$ would have units (bushels/acre).

For credible intervals (or any quantile), we can compute the square root of the endpoints since

$$P(\sigma^2 < c^2) = P(\sigma < c).$$

Find the density through transformations of random variables. In R code,

```
dsqrtinvgamma = function(x, shape, scale) {
  dinvgamma(x^2, shape, scale)*2*x
}
```

Posterior for the standard deviation, σ

```
# Posterior median and 95% CI for population yield standard deviation
sd_median = sqrt(qinvgamma(.5, shape = (n-1)/2, scale = (n-1)*sample_variance/2))
sd_median
```

```
[1] 22.63362
```

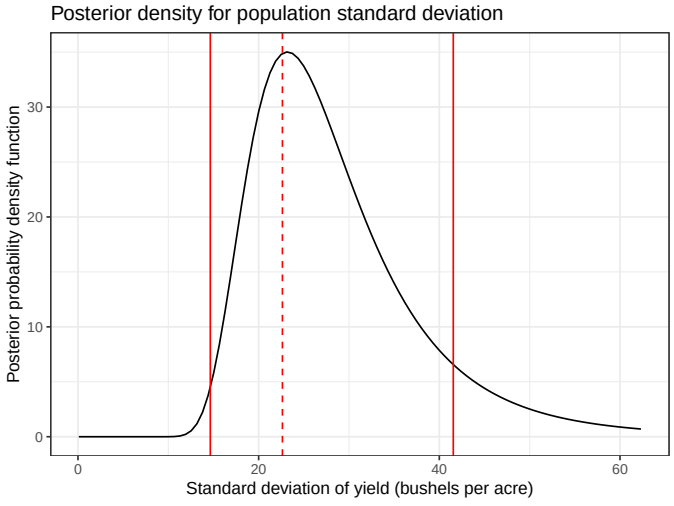
So the posterior median for σ is 23.

```
# Posterior 95% CI for the population yield standard deviation
sd_ci = sqrt(var_ci)
sd_ci
```

```
[1] 14.64795 41.54537
```

So a posterior 95% credible interval for σ is 15, 42.

Posterior for the standard deviation, σ



Bayesian inference in a normal model

- Prior: $p(\mu, \sigma^2) = 1/\sigma^2$
- Posterior:

$$\mu|y \sim t_{n-1}(\overline{y}, s^2/n) \qquad \text{and} \qquad \sigma^2|y \sim IG\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$$

```
# Sufficient statistics
n               = length(y)
sample_mean     = mean(y)
sample_variance = var(y)

# Posterior expectations
sample_mean                  # mu
(n-1)*sample_variance / (n-3) # sigma^2

# Posterior medians
var_median = qinvgamma(.5, shape = (n-1)/2, scale = (n-1)*sample_variance/2)
sd_median  = sqrt(median_var)

# Posterior credible intervals
sample_mean + c(-1,1) *  qt(1-a/2, df = n-1) * sqrt(sample_variance/n)
var_ci = qinvgamma(c(a/2,1-a/2), shape = (n-1)/2, scale = (n-1)*sample_variance/2)
sd_ci  = sqrt(var_ci)
```