I04 - Normal model

STAT 587 (Engineering) - Iowa State University

February 22, 2019

Example data set

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

```
d <- read.csv("yield.csv")
names(d)
```

```
[1] "farm"  "yield"
```

```
(n <- length(d$yield))
```

```
[1] 9
```

```
(ybar <- mean(d$yield))
```

```
[1] 185.5323
```

```
(s <- sd(d$yield))
```

```
[1] 21.68598
```

Normal model (unknown population mean and variance)

Let Y_i be the yield on farm i. Assume $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$ and the default prior $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2} \mathrm{I}(\sigma^2 > 0)$.

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

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

$$\begin{aligned} \mu|\sigma^2, y &\sim N(\overline{y}, \sigma^2/n) \\ \sigma^2|y &\sim IG\left(\tfrac{n-1}{2}, \tfrac{(n-1)s^2}{2}\right) \end{aligned}$$

where

- n is the number of observations,
- $\overline{y}$ is the sample mean, and
- s^2 is the sample variance.

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

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

then

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

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

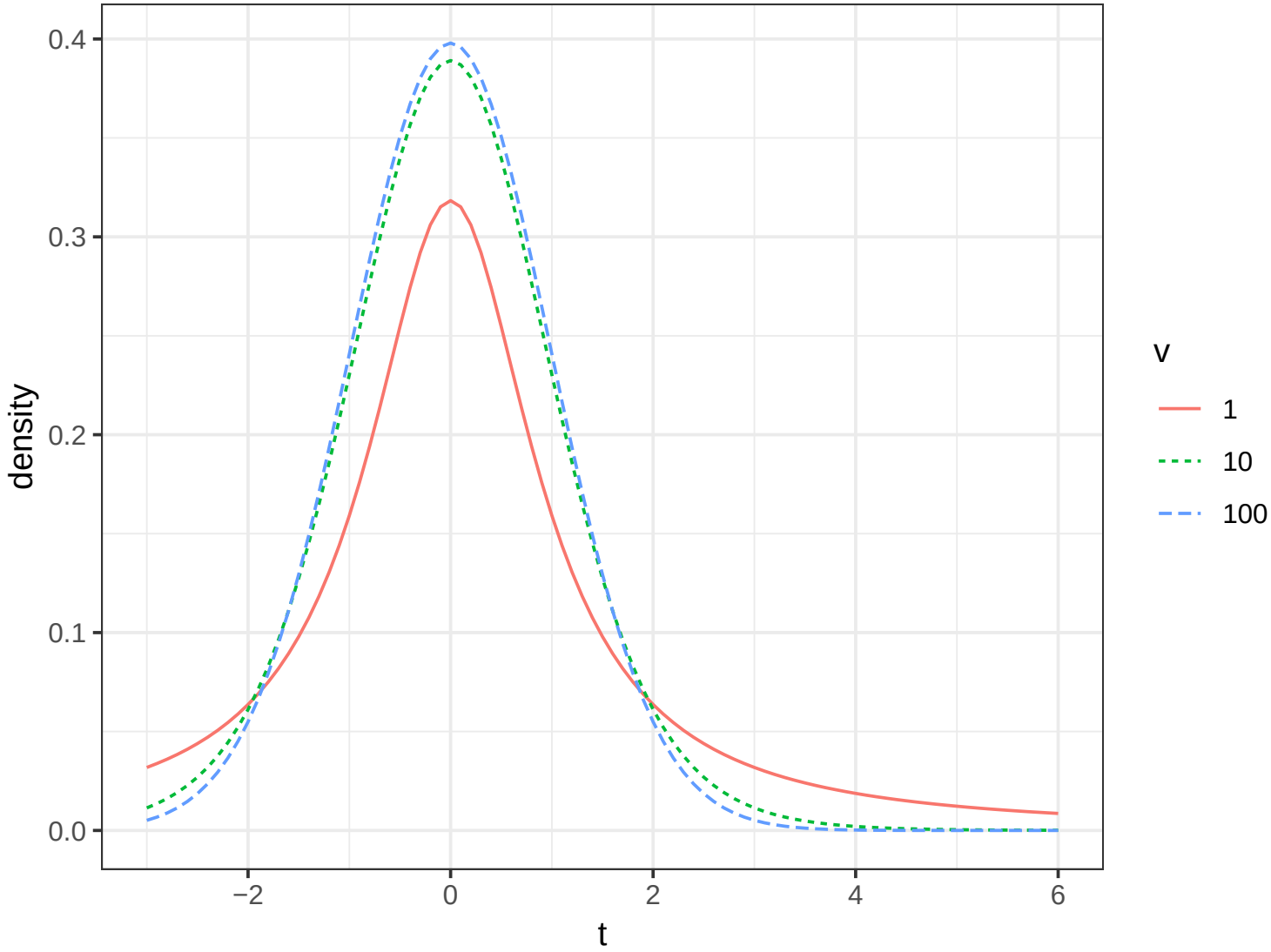
$$\begin{aligned} E[T] &= m & v > 1 \\ Var[T] &= s^2 \tfrac{v}{v-2} & v > 2. \end{aligned}$$

In addition,

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

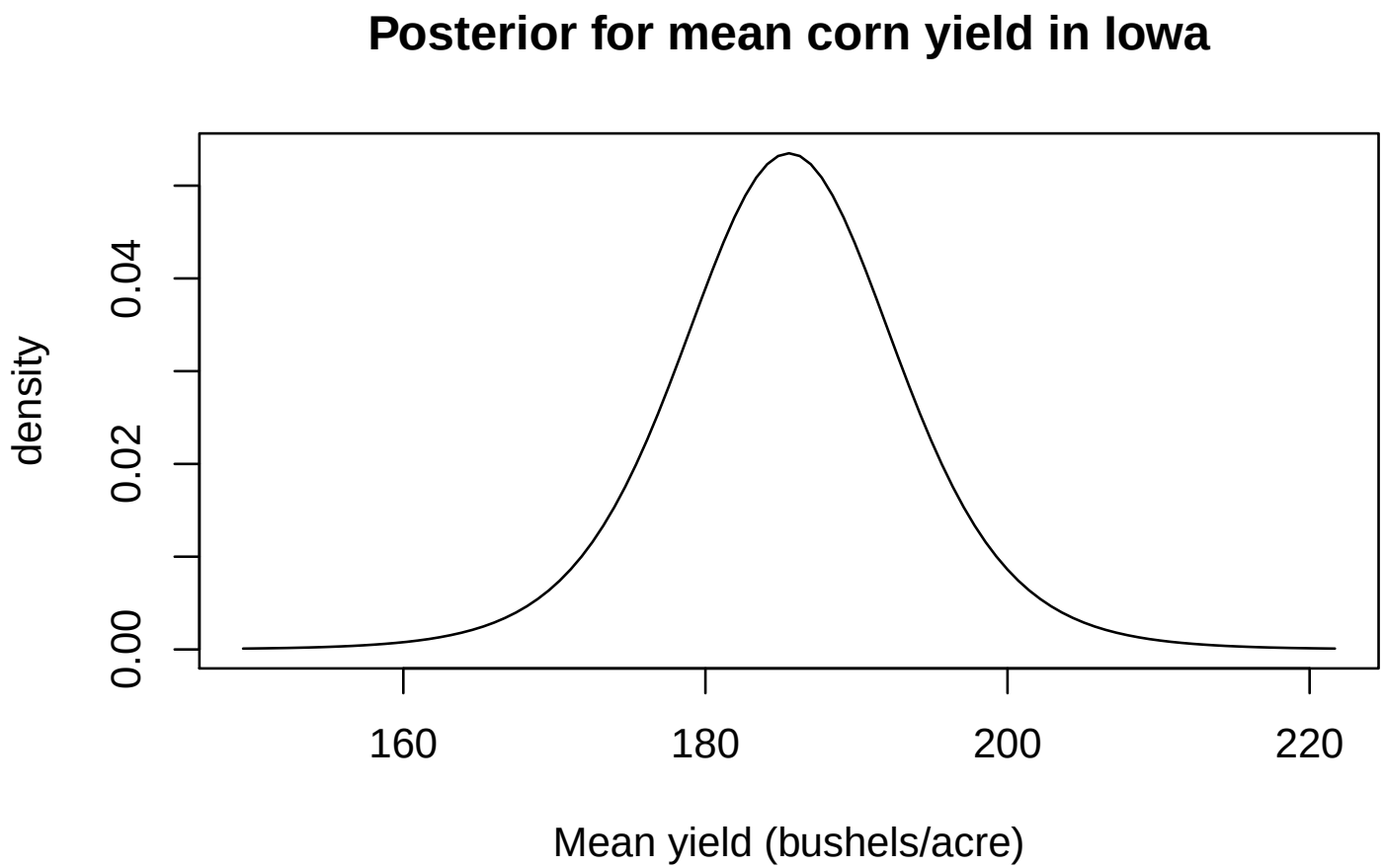
and a standard t has $m = 0$ and $s = 1$.

t distribution as v changes



$p(\mu|y)$

```
curve(dt((x-ybar)/(s/sqrt(n)), df = n-1)/(s/sqrt(n)),
      from = ybar - 5*s/sqrt(n), to = ybar + 5*s/sqrt(n),
      xlab = "Mean yield (bushels/acre)" , ylab = "density",
      main = "Posterior for mean corn yield in Iowa")
```



CIs for μ

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

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

to construct a 100(1-a)% equal-tail credible 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 $a/2$. A 90% equal-tail credible interval is

```
a <- 0.1
ybar + c(-1,1)*qt(1-a/2, df=n-1)*s/sqrt(n)
```

```
[1] 172.0903 198.9744
```

Probability (belief) statements about μ

In order to make probability (belief) statements about the mean, we need to standardize, e.g.

$$P(\mu < c|y) = P\left(\frac{\mu - \overline{y}}{s/\sqrt{n}} < \frac{c - \overline{y}}{s/\sqrt{n}}\middle| y\right) = P\left(T_{n-1} < \frac{c - \overline{y}}{s/\sqrt{n}}\middle| y\right)$$

were T_{n-1} is a standard t with $n-1$ degrees of freedom. Here are some probability statements and the calculations in R:
$P(\mu < 200|y)$:

```
pt((200-ybar)/(s/sqrt(n)), df = n-1)
```

```
[1] 0.959831
```

$P(180 < \mu < 200|y)$:

```
pt((200-ybar)/(s/sqrt(n)), df = n-1) - pt((180-ybar)/(s/sqrt(n)), df = n-1)
```

```
[1] 0.7268065
```

```
diff(pt((c(180,200)-ybar)/(s/sqrt(n)), df = n-1))
```

```
[1] 0.7268065
```

Posterior for σ^2

We already have the posterior distribution for σ^2, namely

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

Inverse-gamma distribution

Definition

An inverse-gamma random variable, $Y \sim IG(a, b)$, has probability density function

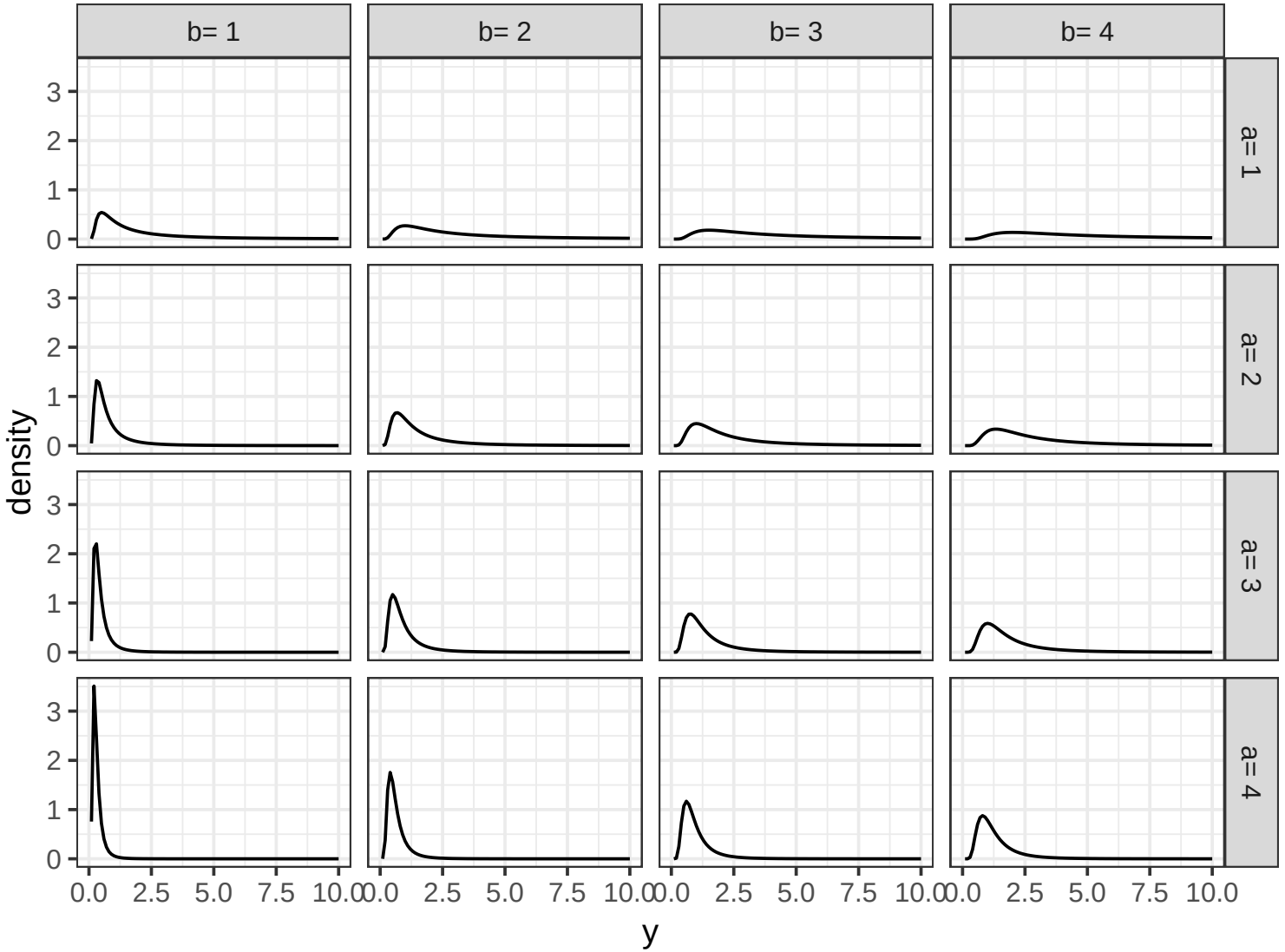
$$\frac{b^a}{\Gamma(a)} y^{-a-1} e^{-b/y}$$

with shape a and scale b. It has

$$\begin{aligned} E[Y] &= \frac{b}{a-1} & a > 1 \\ Var[Y] &= \frac{b^2}{(a-1)^2(a-2)} & a > 2. \end{aligned}$$

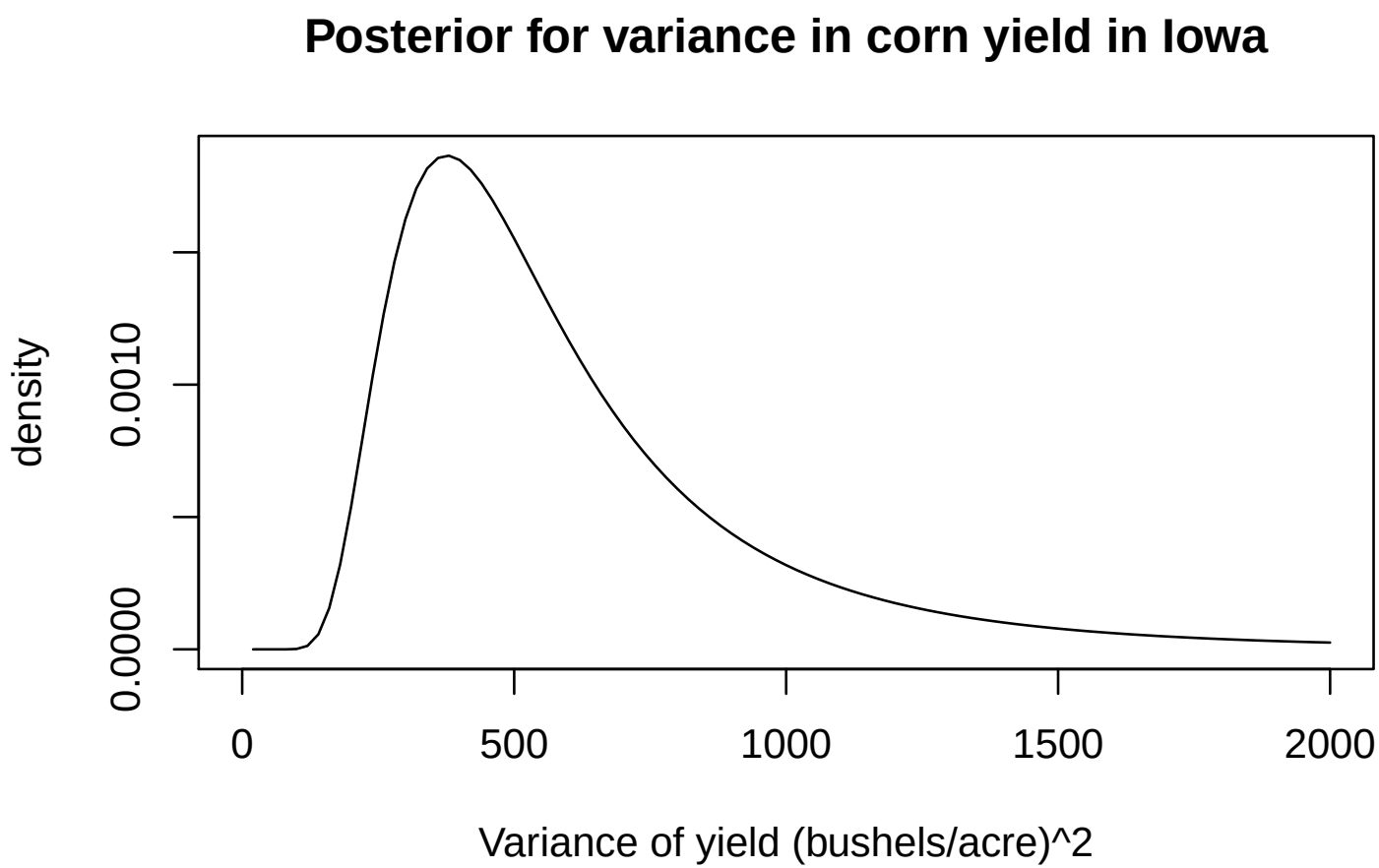
In addition, $1/Y$ is a gamma random variable with shape a and rate b.

Inverse-gamma distribution as the parameters vary



$p(\mu|y)$

```
curve(MCMCpack::dinvgamma(x, shape = (n-1)/2, scale = (n-1)*s^2/2),
      from = 0, to = 2000,
      xlab = "Variance of yield (bushels/acre)^2" , ylab = "density",
      main = "Posterior for variance in corn yield in Iowa")
```



$E[\sigma^2|y]$

Since the marginal posterior for σ^2 is

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

the

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

CIs for σ^2

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)
}
```

Using this function, we can calculate credible intervals. For example, an 80% equal-tail credible interval for σ^2 is

```
a <- 0.2
qinvgamma(c(a/2,1-a/2), shape = (n-1)/2, scale = (n-1)*s^2/2)
```

```
[1]  281.5728 1078.1519
```

Probability (belief) statements for σ^2

In addition to no quantile function, there is no cdf function, but here is one based on the fact that

$$P(\sigma^2 < c|y) = P\left(\frac{1}{\sigma^2} > \frac{1}{c}\middle| y\right)$$

```
pinvgamma <- function(c, shape, scale = 1) {
  1-pgamma(1/c, shape = shape, rate = scale)
}
```

Some example probability (belief) statements are:
$P(\sigma^2 < 500|y)$:

```
pinvgamma(500, shape = (n-1)/2, scale = (n-1)*s^2/2)
```

```
[1] 0.4812377
```

$P(100 < \sigma^2 < 500|y)$:

```
diff(pinvgamma(c(100,500), shape = (n-1)/2, scale = (n-1)*s^2/2))
```

```
[1] 0.4812289
```

CIs for σ

Since the standard deviation has the same units of the data, it is often easier to interpret. Thus, we would prefer to create credible intervals for the standard deviation. To do so, just calculate credible intervals for the for the variance and take the square root. For example, an 80% equal-tail credible interval for the standard deviation is

```
a <- 0.2
ci_variance = qinvgamma(c(a/2,1-a/2), shape = (n-1)/2, scale = (n-1)*s^2/2)
sqrt(ci_variance)


[1] 16.78013 32.83522
```

This trick works for any monotonic function and any quantile.

$P(\sigma < c|y)$

Since $\sigma > 0$, we have

$$P(\sigma < c|y) = P(\sigma^2 < c^2|y)$$

and we can use this to calculate probability statements for the standard deviation.

Some example probability (belief) statements are:
$P(\sigma < 20|y) = P(\sigma^2 < 20^2|y)$:

```
pinvgamma(20^2, shape = (n-1)/2, scale = (n-1)*s^2/2)
```

```
[1] 0.3092405
```

$P(20 < \sigma < 25|y) = P(20^2 < \sigma^2 < 25^2|y)$:

```
diff(pinvgamma(c(20,25)^2, shape = (n-1)/2, scale = (n-1)*s^2/2))
```

```
[1] 0.3357952
```

$E[\sigma|y]$

To calculate $E[\sigma|y]$ exactly you will need to learn how to take transformations of random variables which you would learn in STAT 588. Instead, we will use a Monte Carlo (simulation) approach which will provide us an estimate. Specifically for $m = 1, \ldots, M$ with M large, simulate

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

Then

$$\begin{aligned} E[\sigma^2|y] &\approx \frac{1}{M}\sum_{m=1}^{M} \sigma^{2(m)} \\ E[\sigma\ |y] &\approx \frac{1}{M}\sum_{m=1}^{M} \sqrt{\sigma^{2(m)}} \end{aligned}$$

```
sigma2 <- MCMCpack::rinvgamma(1e5, shape = (n-1)/2, scale = (n-1)*s^2/2)
c(mean(sigma2), (n-1)*s^2/(n-3)) # estimate first, truth second
```

```
[1] 628.7087 627.0422
```

```
mean(sqrt(sigma2))               # only have an estimate
```

```
[1] 24.04802
```

You can use the CLT to determine how good this estimate is.

Default analysis for normal model

Let $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$ with default prior $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2} \mathrm{I}(0 < \sigma^2)$. Then the posterior is

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

with a marginal posterior for μ of

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

The Bayes estimators are

$$\begin{aligned} E[\mu|y] &= \overline{y} \\ E[\sigma^2|y] &= \frac{(n-1)s^2}{n-3} \quad n > 3 \end{aligned}$$