Part 4:
Multi-parameter and normal models

The normal model

Perhaps the most useful (or utilized) probability model for data analysis is the normal distribution

There are several reasons for this, e.g.,

- the CLT
- it is a simple model with separate parameters for the population mean and variance - two quantities that are often of primary interest

The normal PDF

A RV Y is said to be normally distributed with mean μ and variance σ^2 if the density of Y is given by

$$f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

for $-\infty < y < \infty$

We shall write $Y \sim \mathcal{N}(\mu, \sigma^2)$ and call the pair of parameters $\theta \equiv (\mu, \sigma^2)$

Normal properties

Some important things to remember about the normal distribution include

- the distribution is symmetric about μ: the mode, the median and the mean are all equal to μ
- about 95% of the population lies within two (more precisely 1.96) standard deviations of the mean
- if $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ and X and Y are independent, then

$$aX + bY \sim \mathcal{N}(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$$

Joint sampling density

Suppose our sampling model is

$$\{Y_1, \ldots, Y_n | \mu, \sigma^2\} \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$$

Then the joint sampling density is given by

$$p(y_1, \ldots, y_n | \mu, \sigma^2) = \prod_{i=1}^{n} p(y_i | \mu, \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_i - \mu)^2}{\sigma^2} \right\}$$

$$= (\sqrt{2\pi\sigma^2})^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^2} \right\}$$

Joint sampling density

By expanding the quadratic term in the exponent

$$\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} y_i^2 - 2\frac{\mu}{\sigma^2} \sum_{i=1}^{n} y_i + n\frac{\mu^2}{\sigma^2}$$

we see that $p(y_1, \ldots, y_n | \mu, \sigma^2)$ depends upon $y_1, \ldots, y_n$ through $\{\sum y_i^2, \sum y_i\}$, a two-dimensional sufficient statistic

Knowing these quantities is equivalent to knowing

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \text{and} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

and so $\{\bar{y}, s^2\}$ are also a sufficient statistic

Inference by conditioning

Inference for this two-parameter model can be broken down into two one-parameter problems

We begin with the problem of making inference for μ when σ^2 is known, and use a conjugate prior for μ

For any (conditional) prior distribution $p(\mu|\sigma^2)$ the posterior distribution will satisfy

$$p(\mu|y_1, \ldots, y_n, \sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\} \times p(\mu|\sigma^2)$$

$$\propto e^{c_1(\mu-c_2)^2} \times p(\mu|\sigma^2)$$

Conjugate prior

So we see that if $p(\mu|\sigma)$ is to be conjugate, then it must include quadratic terms like $\exp\{c_1(\mu - c_2)^2\}$

The simplest such class of probability densities is the normal family, suggesting that if $p(\mu|\sigma)$ is normal and

$$\{y_1, \ldots, y_n\} \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$$

then $p(\mu|y_1, \ldots, y_n, \sigma^2)$ is also a normal density

Posterior derivation

If $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$ then

$$p(\mu|y_1, \ldots, y_n, \sigma^2) \propto \exp\left\{-\frac{(\mu - b/a)^2}{2/a}\right\}$$

where

$$a = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \quad \text{and} \quad b = \frac{\mu_0}{\tau_0^2} + \frac{\sum y_i}{\sigma^2}$$

This function has exactly the same form as a normal density curve with $1/a$ playing the role of the variance and b/a playing the role of the mean

Normal posterior

Since probability distributions are determined by their shape, this means that $p(\mu|y_1,\ldots,y_n,\sigma^2)$ is indeed a normal density

i.e., $\{\mu|y_1,\ldots,y_n,\sigma^2\} \sim \mathcal{N}(\mu_n,\tau_n^2)$ where

$$\tau_n^2 = \frac{1}{a} = \frac{1}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}} \quad \text{and} \quad \mu_n = \frac{b}{a} = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}$$

So: normal prior + normal sampling model = normal posterior

Combining information

The (conditional) posterior parameters τ_n^2 and μ_n combine the prior parameters τ_0^2 and μ_0 with the terms from the data

First consider the posterior inverse variance, a.k.a. the precision:

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

$\tilde{\tau}_n^2$

$\tilde{\tau}_0^2$

$$\tilde{\sigma}^2 = \frac{1}{\sigma^2}$$

So the posterior precision combines the sampling precision and the prior precision

$$\tilde{\tau}_n^2 = \tilde{\tau}_0^2 + n\tilde{\sigma}^2 \quad \text{and} \quad \tilde{\tau}_n^2 \longrightarrow \infty, \quad \text{as} \quad n \longrightarrow \infty$$

Combining means

Notice that the posterior mean decomposes as

$$\mu_n = \frac{\tilde{\tau}_0^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\mu_0 + \frac{n\tilde{\sigma}^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\bar{y}$$

so it is a weighted average of the prior mean and the data mean

The weight on the prior mean is $1/\tau_0^2$, the prior precision

Prior sample size

If the prior mean were based on κ_0 prior observations from the same (or similar) population as $Y_1, \ldots, Y_n$ then we might want to set

$$\tau_0^2 = \frac{\sigma^2}{\kappa_0}$$

which may be interpreted as the variance of the *mean* of the prior observations

In this case the formula for the posterior mean reduces to

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$

Finally, $\mu_n \to \bar{y}, \text{ as } n \to \infty$

Prediction

Consider predicting a new observation $\tilde{Y}$ from the population after having observed

$$\{Y_1 = y_1, \ldots, Y_n = y_n\}$$

To find the predictive distribution we may take advantage of the following fact

$$\{\tilde{Y}|\mu, \sigma^2\} \sim \mathcal{N}(\mu, \sigma^2) \Leftrightarrow \tilde{Y} = \mu + \tilde{\varepsilon}, \;\; \tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$$

In other words, saying that $\tilde{Y}$ is normal with mean μ is the same as saying that $\tilde{Y}$ is equal to μ plus some mean-zero normally distributed noise

Predictive moments

Using this result, let's first compute the posterior mean and variance of $\tilde{Y}$

$$\begin{aligned}
&\mathbb{E}\{\tilde{Y}|y_1,\ldots,y_n,\sigma^2\}\\
&\qquad= \mathbb{E}\{\mu+\tilde{\varepsilon}|y_1,\ldots,y_n,\sigma^2\}\\
&\qquad= \mathbb{E}\{\mu|y_1,\ldots,y_n,\sigma^2\}+\mathbb{E}\{\tilde{\varepsilon}|y_1,\ldots,y_n,\sigma^2\}\\
&\qquad= \mu_n+0=\mu_n
\end{aligned}$$

$$\begin{aligned}
&\mathrm{Var}[\tilde{Y}|y_1,\ldots,y_n,\sigma^2]\\
&\qquad= \mathrm{Var}[\mu+\tilde{\varepsilon}|y_1,\ldots,y_n,\sigma^2]\\
&\qquad= \mathrm{Var}[\mu|y_1,\ldots,y_n,\sigma^2]+\mathrm{Var}[\tilde{\varepsilon}|y_1,\ldots,y_n,\sigma^2]\\
&\qquad= \tau_n^2+\sigma^2
\end{aligned}$$

Moments to distribution

Recall that the sum of independent normal random variables is normal

Therefore, since μ and $\tilde{\varepsilon}$, conditional on $y_1, \ldots, y_n$ and σ^2, are normally distributed, so is $\tilde{Y} = \mu + \tilde{\varepsilon}$

So the predictive distribution is

$$\tilde{Y} | y_1, \ldots, y_n, \sigma^2 \sim \mathcal{N}(\mu_n, \tau_n^2 + \sigma^2)$$

Observe that, as $n \to \infty$,

$$\mathrm{Var}[\tilde{Y} | y_1, \ldots, y_n, \sigma^2] \to \sigma^2 > 0$$

i.e., certainty in μ does not translate into certainty about $\tilde{Y}$

Example: Midge wing length

Grogan and Wirth (1981) provide data on the wing length in millimeters of nine members of a species of midge (small, two-winged flies)

From these measurements we wish to make inference about the population mean μ

Example: prior information

Studies from other populations suggest that wing lengths are usually around 1.9mm, so we set $\mu_0 = 1.9$

We also know that lengths must be positive $(\mu > 0)$

We can approximate this restriction with a normal prior distribution for μ as follows:

Since most of the normal density is within two standard deviations of the mean we choose τ_0^2 so that

$$\mu_0 - 2\tau_0 > 0 \Rightarrow \tau_0 < 1.9/2 = 0.95$$

For now we take $\tau_0^2 = 0.95^2$ which somewhat overstates our prior uncertainty

Example: data and posterior

The observations in order of increasing magnitude are

$$(1.64, 1.70, 1.72, 1.74, 1.82, 1.82, 1.82, 1.9, 2.08)$$

Therefore the posterior is

$$\{\mu | y_1, \ldots, y_9, \sigma^2\} \sim \mathcal{N}(\mu_n, \tau_n^2)$$

where

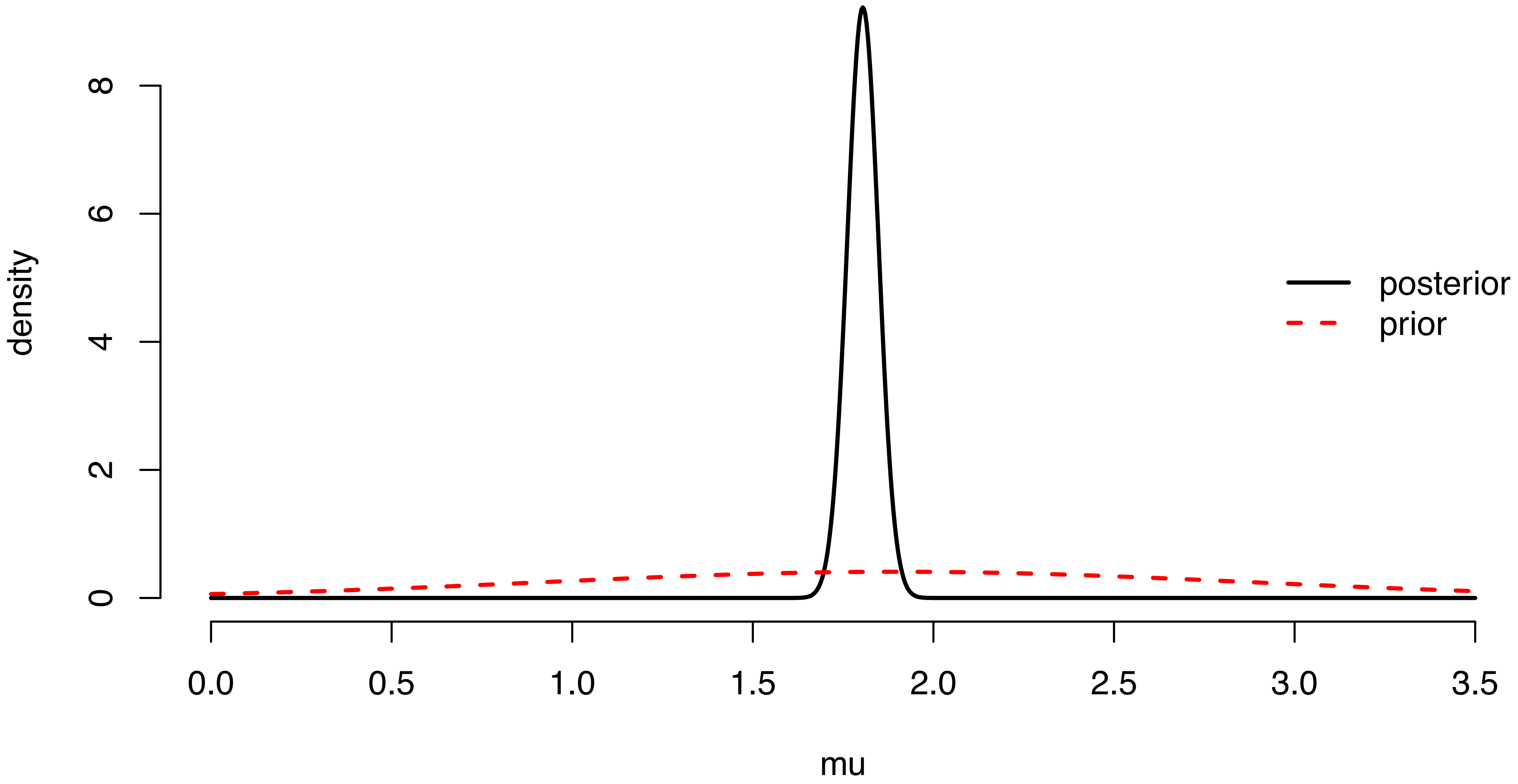
$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{1.11 \times 1.9 + \frac{9}{\sigma^2} \times 1.804}{1.11 + \frac{9}{\sigma^2}}$$

$$\tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{1}{1.11 + \frac{9}{\sigma^2}}$$

Example: a choice of σ^2

If we take $\sigma^2 = s^2 = 0.017$, then

$$\{\mu | y_1, \ldots, y_9, \sigma^2 = 0.017\} \sim \mathcal{N}(1.805, 0.002)$$



A 95% CI for μ is (1.72, 1.89)

Example: full accounting of uncertainty

However, these results assume that we are certain that $\sigma^2 = s^2$ when in fact s^2 is only a rough estimate of σ^2 based upon only nine observations

To get a more accurate representation of our information (and in particular of our posterior uncertainty) we need to account for the fact that σ^2 is not known

Joint Bayesian inference

Bayesian inference for two or more parameters is not conceptually different from the one-parameter case

For any joint prior distribution $p(\mu, \sigma^2)$ for μ and σ^2, posterior inference proceeds using Bayes' rule:

$$p(\mu, \sigma^2 | y_1, \ldots, y_n) = \frac{p(y_1, \ldots, y_n | \mu, \sigma^2) p(\mu, \sigma^2)}{p(y_1, \ldots, y_n)}$$

As before, we will begin by developing a simple conjugate class of prior distributions which will make posterior calculations easy

Prior decomposition

A joint distribution can always be expressed as the product of a conditional probability and a marginal probability:

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

We just saw that if σ^2 were known, then a conjugate prior distribution for μ was $\mathcal{N}(\mu_0, \tau_0^2)$

Prior decomposition

Lets consider the particular case where $\tau_0^2 = \sigma^2/\kappa_0$:

$$\begin{aligned} p(\mu, \sigma^2) &= p(\mu|\sigma^2)p(\sigma^2) \\ &= \mathcal{N}(\mu; \mu_0, \tau_0^2 = \sigma^2/\kappa_0) \times p(\sigma^2) \end{aligned}$$

The parameters μ_0 and κ_0 can be interpreted as the mean and sample size from a set of prior observations

For σ^2 we need a family of prior distributions that has support on $(0, \infty)$

Conjugate prior for σ^2

One such family of distributions is the gamma family, as we used for the Poisson sampling model

Unfortunately, this family is not conjugate for the normal variance

However, the gamma family does turn out to be a conjugate class of densities for the precision $1/\sigma^2$

When using such a prior we say that σ^2 has an inverse-gamma distribution

$$\begin{aligned} & 1/\sigma^2 \sim G(a,b) \\ \Rightarrow \qquad & \sigma^2 \sim \mathrm{IG}(a,b) \end{aligned}$$

Inverse-gamma prior

For interpretability, instead of using a and b we will use

$$\sigma^2 \sim \mathrm{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$

Under this parameterization:

$$\mathbb{E}\{\sigma^2\} = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2 - 1}$$

$$\mathrm{mode}(\sigma^2) = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2 + 1}, \quad \text{so mode}(\sigma^2) < \sigma_0^2 < \mathbb{E}\{\sigma^2\}$$

$$\mathrm{Var}[\sigma^2] \quad \text{is decreasing in } v_0$$

We can interpret the prior parameters (σ_0^2, ν_0) as the prior sample variance and the prior sample size

Posterior inference

Suppose our prior model and sampling model are

$$\sigma^2 \sim \mathrm{IG}(\nu_0/2, \nu_0\sigma_0^2/2)$$
$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)$$
$$Y_1, \ldots Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$$

Just as the prior distribution for μ and σ^2 can be decomposed as $p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2)$, the posterior distribution can be similarly decomposed as

$$p(\mu, \sigma^2|y_1, \ldots, y_n) = p(\mu|\sigma^2, y_1, \ldots, y_n)p(\sigma^2|y_1, \ldots, y_n)$$

Posterior conditional(s)

We have already derived the posterior conditional distribution of μ given σ^2

Plugging in σ^2/κ_0 for τ_0^2 we have

$$\{\mu|y_1,\ldots,y_n,\sigma^2\} \sim \mathcal{N}(\mu_n,\sigma^2/\kappa_n), \quad \text{where}$$

$$\kappa_n = \kappa_0 + n \text{ and } \mu_n = \frac{(\kappa_0/\sigma^2)\mu_0 + (n/\sigma^2)\bar{y}}{\kappa_0/\sigma^2 + n/\sigma^2} = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n}$$

(posterior sample size and mean)

Posterior conditional(s)

The posterior distribution of σ^2 can be obtained by integrating over the unknown value of μ

$$
\begin{aligned}
p(\sigma^2|y_1,\ldots,y_n) &\propto p(y_1,\ldots,y_n|\sigma^2)p(\sigma^2) \\
&= p(\sigma^2)\int p(y_1,\ldots,y_n|\mu,\sigma^2)p(\mu|\sigma^2)\,d\mu
\end{aligned}
$$

This integral can be done without much knowledge of calculus, but it is somewhat tedious

Posterior conditional(s)

Check that this leads to

$$\{\sigma^2 | y_1, \ldots, y_n\} \sim \mathrm{IG}(\nu_n/2, \nu_n\sigma_n^2/2), \quad \text{where}$$

$$\nu_n = \nu_0 + n$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left[\boxed{\nu_0\sigma_0^2} + \boxed{(n-1)s^2} + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2\right]$$

These formulae suggest that the interpretation of ν_0 as a prior sample size, from which a prior sample variance σ_0^2 has been obtained, is reasonable

Similarly, we may think of $\nu_0\sigma_0^2$ and $\nu_n\sigma_n^2$ as a prior and posterior sum of squares, which is the sum of the prior and data sum of squares

Posterior conditional(s)

$$\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y}-\mu_0)^2\right]$$

However, the third term is a bit harder to understand

It says that a large value of $(\bar{y}-\mu_0)^2$ increases the posterior probability of a large σ^2

This makes sense for our joint prior $p(\mu, \sigma^2)$:

- If we want to think of μ_0 as the sample mean of prior observations with variance σ^2, then this term is also an estimate of σ^2

Example: the midge data

The studies of other populations suggest that the true mean and standard deviation of our population under study should not be far from 1.9mm and 0.1mm respectively, suggesting

$$\mu_0 = 1.9 \quad \text{and} \quad \sigma_0^2 = 0.01$$

However, this population may be different from others in terms of wing length, and so we choose $\kappa_0 = \nu_0 = 1$ so that our prior distributions are only weakly centered around the estimates from other populations

Example: the midge data

The sample mean and variance of our observed data are

$$\bar{y} = 1.804 \quad \text{and} \quad s^2 = 0.0169$$

From these values and the prior parameters, we compute

$$\mu_n = \frac{\kappa_0 \mu_0 + n\bar{y}}{\kappa_n} = \frac{1.9 + 9 \times 1.804}{1 + 9} = 1.814$$

$$\begin{aligned} \sigma_n^2 &= \frac{1}{\nu_n} \left[\nu_0 \mu_0 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n} (\bar{y} - \mu_0)^2 \right] \\ &= \frac{0.010 + 0.135 + 0.008}{10} = 0.015 \end{aligned}$$

Example: the full posterior

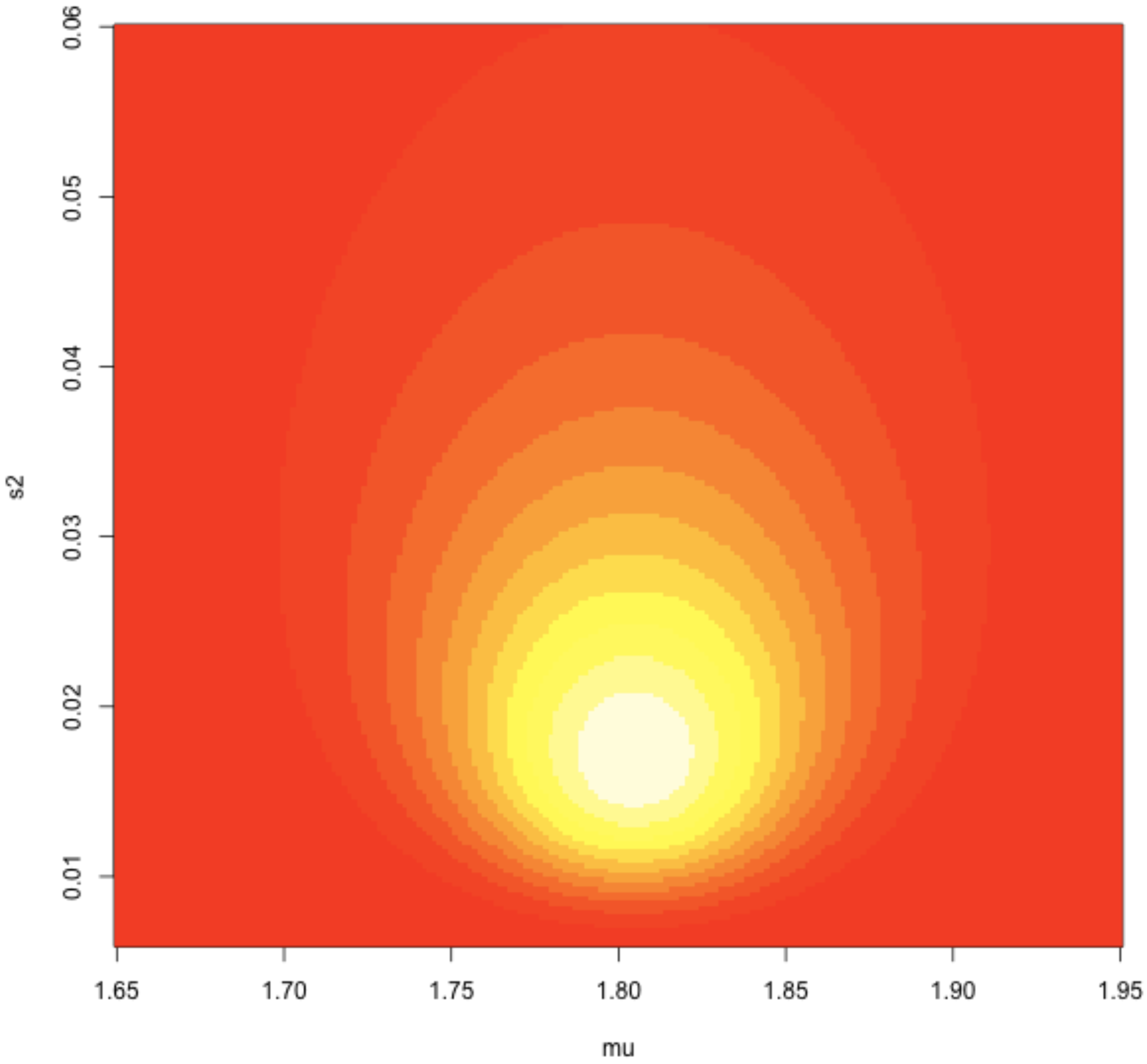
Our joint posterior distribution is completely determined by these values

$$\mu_n = 1.814,\ \kappa_n = 10,\ \sigma_n^2 = 0.015,\ \nu_n = 10$$

and can be expressed as

$$\begin{aligned} \{\mu | y_1, \ldots, y_n, \sigma^2\} &\sim \mathcal{N}(1.814, \sigma^2/10) \\ \{\sigma^2 | y_1, \ldots, y_n\} &\sim \mathrm{IG}(10/2, 10 \times 0.015/2) \end{aligned}$$

Example: the full posterior image



Observe that the contours are more peaked as a function of μ for low values of σ^2 than high values

Marginal interests

For many data analyses, interest primarily lies in estimating the population mean μ, and so we would like to calculate quantities like

- $\mathbb{E}\{\mu|y_1, \ldots, y_n\}$
- $\mathrm{Var}[\mu|y_1, \ldots, y_n]$
- $P(\mu_1 < \mu_2|y_1, \ldots, y_n)$

These quantities are all determined by the *marginal* posterior distribution of μ given the data

But all we know so far is that the *conditional* distribution of μ given the data and σ^2 is normal, and that σ^2 given the data is inverse-gamma

Joint sampling

If we could generate *marginal* samples of μ, then we could use the MC method to approximate the marginal quantities of interest

It turns out that this is easy to do by generating samples of μ and σ^2 from their joint posterior by MC

$$\begin{aligned}
\sigma^{2(1)} &\sim \mathrm{IG}(\nu_n/2, \sigma_n^2\nu_n/2), & \mu^{(1)} &\sim \mathcal{N}(\mu_n, \sigma^{2(1)}/\kappa_n) \\
\sigma^{2(2)} &\sim \mathrm{IG}(\nu_n/2, \sigma_n^2\nu_n/2), & \mu^{(2)} &\sim \mathcal{N}(\mu_n, \sigma^{2(2)}/\kappa_n) \\
&\vdots & &\vdots \\
\sigma^{2(S)} &\sim \mathrm{IG}(\nu_n/2, \sigma_n^2\nu_n/2), & \mu^{(S)} &\sim \mathcal{N}(\mu_n, \sigma^{2(S)}/\kappa_n)
\end{aligned}$$

Monte Carlo marginals

The sequence of pairs

$$\{(\sigma^{2(1)}, \mu^{(1)}), \ldots, (\sigma^{2(S)}, \mu^{(S)})\}$$

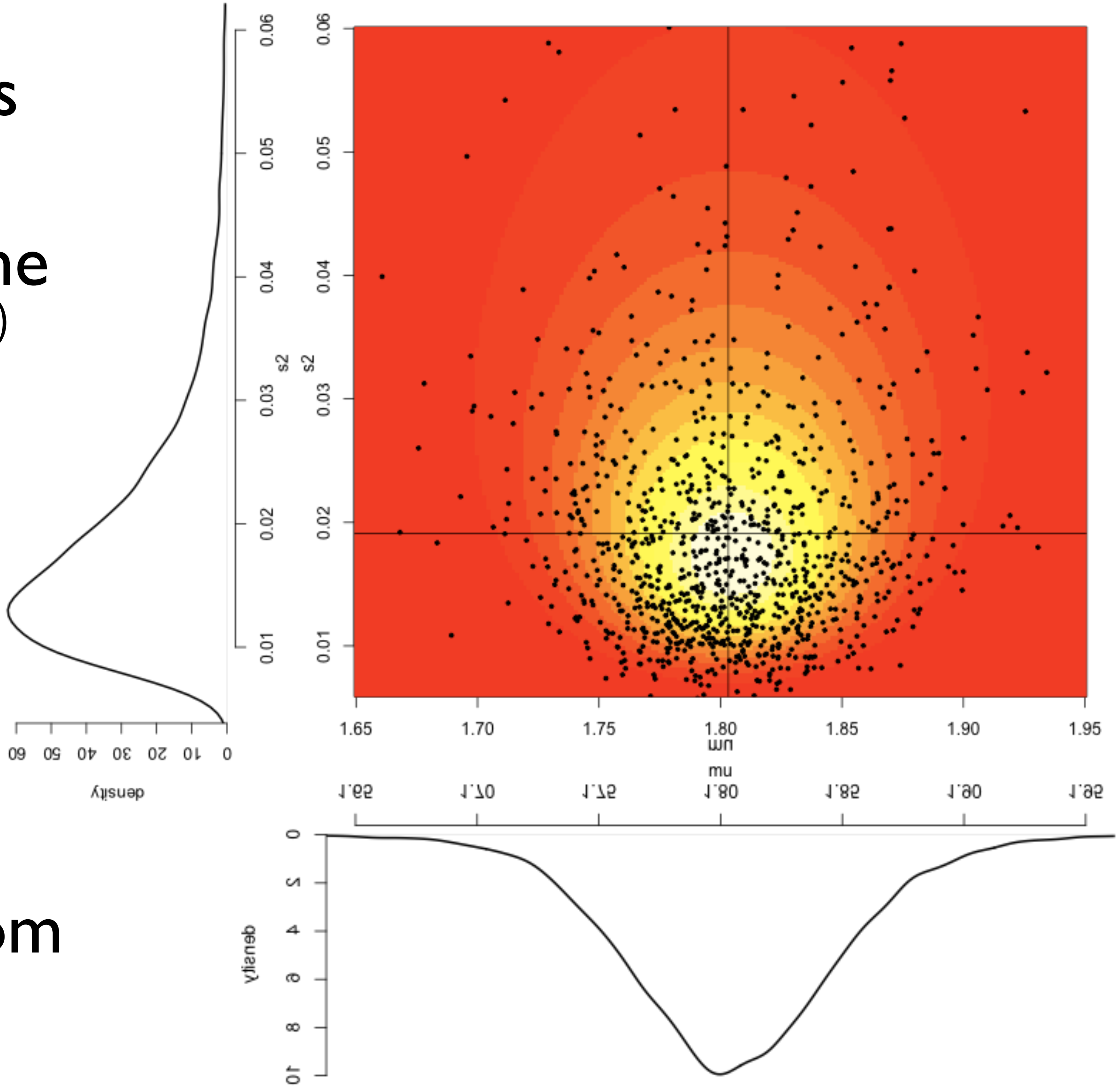
simulated with this procedure are independent samples from the joint posterior $p(\mu, \sigma^2 | y_1, \ldots, y_n)$

Additionally, the simulated sequence $\{\mu^{(1)}, \ldots, \mu^{(S)}\}$ can be seen as independent samples from the *marginal* posterior distribution $p(\mu | y_1, \ldots, y_n)$

So these may be used to make MC approximations to posterior expectations of (functions of) μ

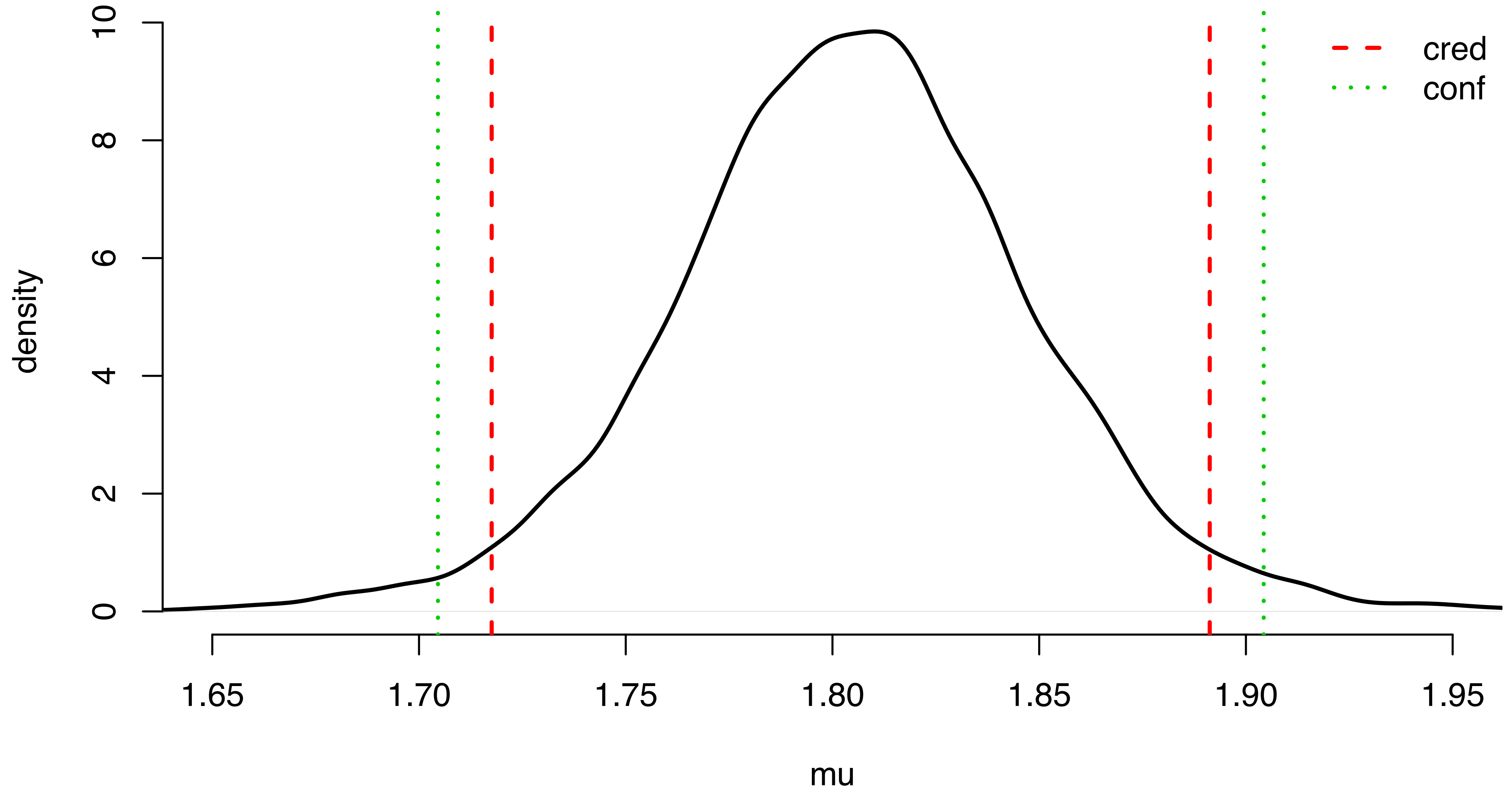
Example: joint & marginal samples

The $\mu^{(s)}$ samples are obtained conditional on the samples of $\sigma^{2(s)}$ leading to joint samples



We may extract the marginals from the joint

Example: credible v. confidence intervals



The Bayesian CI is very close to the frequentist CI, based on t-statistics since

$$p(\mu|y_1, \ldots, y_n) \sim \mathrm{St}_{\nu_0+n}(\mu_n, \sigma_n^2/\kappa_n)$$

If κ_0 and ν_0 are small, this will be very close to the sampling distribution of the MLE

“Objective” priors?

How small can κ_0 and ν_0, the “prior sample sizes”, be?

Consider

$$\mu_n = \frac{\kappa_0 \mu_0 + n\bar{y}}{\kappa_n}$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\mu_0 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2\right]$$

So as $\kappa_0, \nu_0 \to 0$

$$\mu_n \to \bar{y}$$

$$\sigma_n^2 \to \frac{n}{n-1}s^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$$

Improper prior

This leads to the following "posterior":

$$\{\sigma^2 | y_1, \ldots, y_n\} \sim \mathrm{IG}\left(\frac{n}{2}, \frac{n}{2}\frac{1}{n}\sum (y_i - \bar{y})^2\right)$$

$$\{\mu | y_1, \ldots, y_n, \sigma^2\} \sim \mathcal{N}\left(\bar{y}, \frac{\sigma^2}{n}\right)$$

More formally, if we take the improper prior $p(\mu, \sigma^2) \propto 1/\sigma^2$ the posterior is proper, and we get the same posterior conditional for μ as above, but

$$\{\sigma^2 | y_1, \ldots, y_n\} \sim \mathrm{IG}\left(\frac{n-1}{2}, \frac{1}{2}\sum (y_i - \bar{y})^2\right)$$

Classical comparison

The marginal posterior for μ may then be obtained as

$$
\begin{aligned}
& p(\mu|y_1,\dots,y_n) \\
\text{(LTP)} \quad & = \int \underset{\text{(normal conditional)}}{p(\mu|\sigma^2, y_1,\dots,\overset{\text{(cond. prob)}}{y_n})}\, \underset{\text{(IG conditional)}}{p(\sigma^2|y_1,\dots,y_n)}\, d\sigma^2 \\
& \cdots \propto \mathrm{St}_{n-1}(\mu; \bar{y}, s^2/n)
\end{aligned}
$$

It is interesting to compare this result to the sampling distribution of the t-statistic, i.e., of the MLE

$$
t = \frac{\bar{Y} - \mu}{s/\sqrt{n}}|\mu \sim t_{n-1} \quad \Leftrightarrow \quad \hat{\mu} = \bar{Y} \sim \mathrm{St}_{n-1}(\mu, s^2/n)
$$

Point estimator

A point estimator of an unknown parameter θ is a function that converts your data into a single element of the parameter space Θ

In the case of the normal sampling model and conjugate prior distribution, the posterior mean estimator of μ is

$$\begin{aligned}\hat{\mu}_b(y_1, \ldots, y_n) &= \mathbb{E}\{\mu | y_1, \ldots, y_n\} \\ &= \frac{n}{\kappa_0 + n}\bar{y} + \frac{\kappa_0}{\kappa_0 + n}\mu_0 \\ &= w\bar{y} + (1 - w)\mu_0\end{aligned}$$

Sampling properties

The sampling properties of an estimator such as $\hat{\mu}_b$ refers to its behavior under hypothetically repeatable surveys or experiments

Lets compare the sampling properties of $\hat{\mu}_b$ to $\hat{\mu}_e(y_1, \ldots, y_n) = \bar{y}$, the sample mean, when the true population mean is μ_{true}

$$\mathbb{E}\{\hat{\mu}_e | \mu = \mu_{\text{true}}\} = \mu_{\text{true}}$$

$$\mathbb{E}\{\hat{\mu}_b | \mu = \mu_{\text{true}}\} = w\mu_{\text{true}} + (1 - w)\mu_0$$

So we say that $\hat{\mu}_e$ is unbiased and, unless $\mu_0 = \mu_{\text{true}}$, we say that $\hat{\mu}_b$ is biased

Bias

- Bias refers to how close the center of mass of the sampling distribution of the estimator is to the true value
- An unbiased estimator is an estimator with zero bias, which sounds desirable
- However, bias does not tell us how far away an estimate might be from the true value

For example, $\hat{\mu} = y_1$ is an unbiased estimator of the population mean μ_{true}, but will generally be farther away from μ_{true} than $\bar{y}$

Mean squared error

To evaluate how close an estimator $\hat{\theta}$ is likely to be to the true value θ_{true}, we might use the mean squared error (MSE)

Letting $m = \mathbb{E}\{\hat{\theta}|\theta_{\text{true}}\}$, the MSE is

$$\begin{aligned}
\text{MSE}[\hat{\theta}|\theta_{\text{true}}] &= \mathbb{E}\{(\theta - \theta_{\text{true}})^2|\theta = \theta_{\text{true}}\} \\
&= \mathbb{E}\{(\hat{\theta} - m)^2|\theta = \theta_{\text{true}}\} + (m - \theta_{\text{true}})^2 \\
&= \text{Var}[\hat{\theta}|\theta = \theta_{\text{true}}] + \text{Bias}^2[\hat{\theta}|\theta = \theta_{\text{true}}]
\end{aligned}$$

Mean squared error

This means that, before the data are gathered, the expected distance from the estimator to the true value depends on

- how close θ_{true} is to the center of the distribution of $\hat{\theta}$ (the bias), *and*
- how spread out the distribution is (the variance)

Comparing estimators

Getting back to our comparison of $\hat{\mu}_b$ to $\hat{\mu}_e$, the bias of $\hat{\mu}_e$ is zero, but

$$\mathrm{Var}[\hat{\mu}_e|\mu = \mu_{\mathrm{true}}, \sigma^2] = \frac{\sigma^2}{n}, \quad \text{whereas}$$

$$\mathrm{Var}[\hat{\mu}_b|\mu = \mu_{\mathrm{true}}, \sigma^2] = w^2 \times \frac{\sigma^2}{n} < \frac{\sigma^2}{n}$$

and so $\hat{\mu}_b$ has lower variability

Comparing estimators

Which one is better in terms of MSE?

$$\mathrm{MSE}[\hat{\mu}_e|\mu = \mu_{\mathrm{true}}] = \mathrm{Var}[\hat{\mu}_e|\mu = \mu_{\mathrm{true}}] = \frac{\sigma^2}{n}$$

$$\begin{aligned}\mathrm{MSE}[\hat{\mu}_b|\mu = \mu_{\mathrm{true}}] &= \mathrm{Var}[\hat{\mu}_b|\mu = \mu_{\mathrm{true}}] + \mathrm{Bias}^2[\hat{\mu}_b|\mu = \mu_{\mathrm{true}}] \\ &= w^2\frac{\sigma^2}{n} + (1-w)^2(\mu_0 - \mu_{\mathrm{truth}})^2\end{aligned}$$

Therefore

$$\mathrm{MSE}[\hat{\mu}_b|\mu = \mu_{\mathrm{true}}] < \mathrm{MSE}[\hat{\mu}_e|\mu = \mu_{\mathrm{true}}]$$

if

$$(\mu_0 - \mu_{\mathrm{true}})^2 < \frac{\sigma^2}{n}\frac{1+w}{1-w} = \sigma^2\left(\frac{1}{n} + \frac{2}{\kappa_0}\right)$$

Low MSE Bayesian estimators

So if you know just a little about the population you are about to sample from, you should be able to find values of μ_0 and κ_0 such that the Bayesian estimator has lower average distance to the truth (MSE) than the MLE

E.g., if you are pretty sure that your prior guess μ_0 is within two standard deviations of the true population mean, then if you pick $\kappa_0 = 1$ you can be pretty sure that the Bayes estimator has a lower MSE since

$$(\mu_0 - \mu_{\text{true}})^2 < 4\sigma^2$$

Example: IQ scores

Scoring on IQ tests is designed to produce a normal distribution with a mean of 100 and a variance of 225 when applied to the general population

Suppose that we were to sample n individuals from a particular town in the USA and then use it to estimate μ, the town-specific mean IQ score

For a Bayesian estimation, if we lack much information about the town in question, a natural choice of would be $\mu_0 = 100$

Example: IQ scores, MSE

Suppose that, unknown to us, the people in this town are extremely exceptional and the true mean and variance of IQ scores in the town are

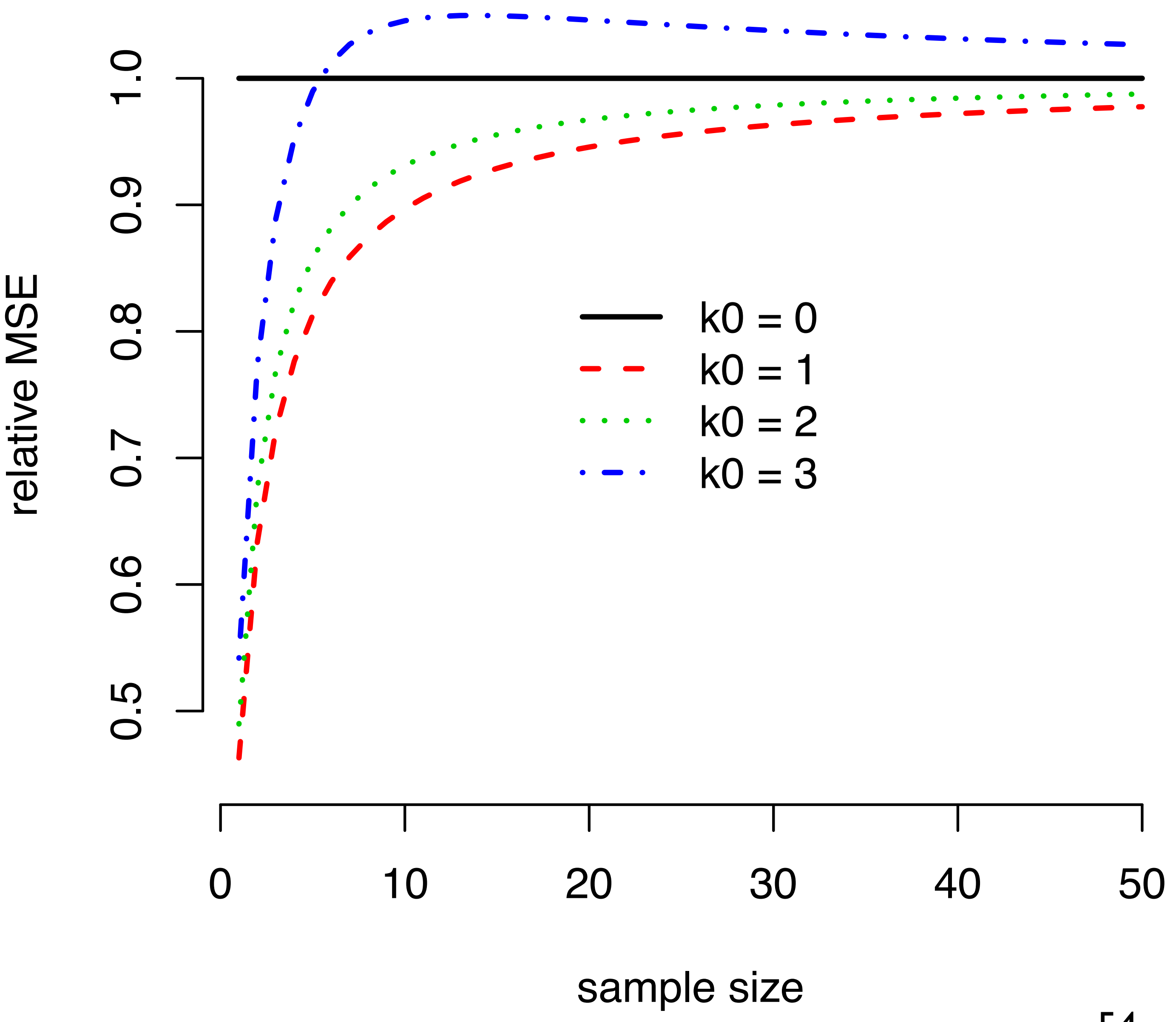
$$(\mu_{\text{true}} = 112, \sigma^2 = 169)$$

The MSEs of the estimators $\hat{\mu}_e$ and $\hat{\mu}_b$ are

$$\begin{aligned}
\text{MSE}[\hat{\mu}_e|\mu = 112, \sigma^2 = 169] &= \text{Var}[\hat{\mu}_e|\mu = 112, \sigma^2 = 169] \\
&= \frac{169}{n} \\
\text{MSE}[\hat{\mu}_b|\mu = 112, \sigma^2 = 169] &= \text{Var}[\hat{\mu}_b|\mu = 112, \sigma^2 = 169] \\
&\quad + \text{Bias}^2[\hat{\mu}_b|\mu = 112, \sigma^2 = 169] \\
&= w^2\frac{169}{n} + (1-w)^2 144
\end{aligned}$$

Example: Relative MSE

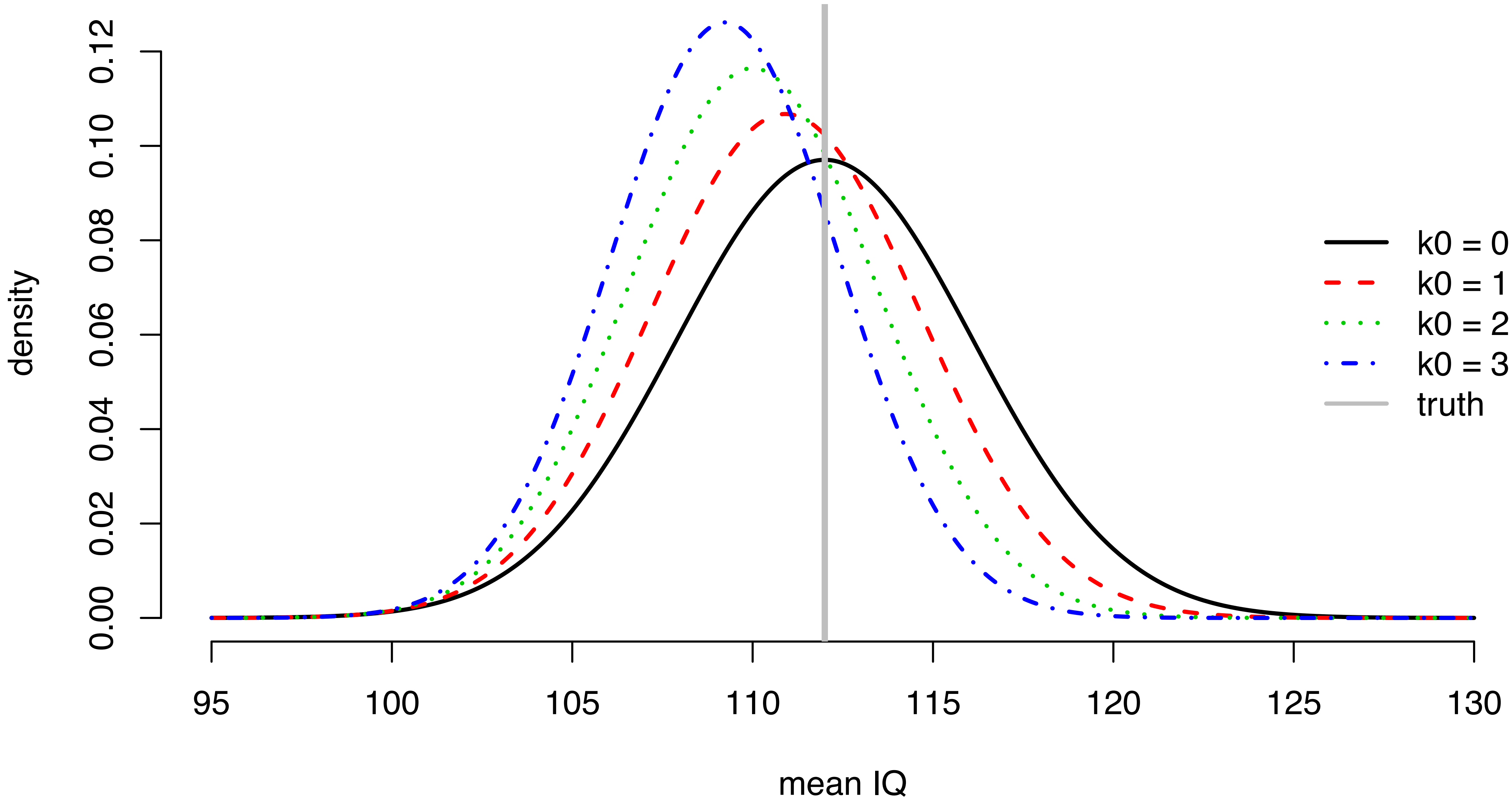
One way compare MSEs is through their ratio. Consider $\text{MSE}[\hat{\mu}_b|\mu = 112]/\text{MSE}[\hat{\mu}_e|\mu = 112]$ plotted as a function of n, for $k_0 \in \{1, 2, 3\}$



- when $k_0 \in \{1, 2\}$, the Bayes estimate has lower MSE than the MLE
- the $\mu_0 = 100$ setting is only bad for $\kappa_0 \geq 3$
- as n increases, the bias of the estimators shrinks to zero

Example: Sampling densities

Consider the sampling densities when $n = 10$ which highlight the relative contributions of the bias and variance to the MSE


density
mean IQ
k0 = 0
k0 = 1
k0 = 2
k0 = 3
truth