Module 7, part I: Introduction to Bayesian Analysis

BIOS 526

Reading

- Chapters 1-2 in Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis. Chapman and Hall/CRC.

Concepts

- Review frequentist inference and CI.
- Prior and posterior distributions.
- Bayesian inference for univariate Normal data.
- Bayesian computation.

Road map

We will consider the simple case of estimating a mean and variance.

1. $[\mu|\mathbf{y}]$ for known σ^2.
2. $[\sigma^2|\mathbf{y}]$ for known μ.
3. $[\mu|\mathbf{y}]$ and $[\sigma^2|\mathbf{y}]$ when both μ and σ^2 are unknown.

Review the Frequentist Approach: MLE

Let $y_1, \ldots, y_n$ be a sample of independent and identically distributed values drawn from a Normal population $N(\mu, \sigma^2)$. We will first assume σ^2 to be known.

Let $\mathbf{y} = [y_1, y_2, \ldots, y_n]'$ denote the vector of data, and the data likelihood is given by

$$
\begin{aligned}
[\,\mathbf{y}\,|\,\theta\,] &= \prod_{i=1}^{n} [\,y_i\,|\,\theta\,] = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i-\theta)^2\right] \\
&= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right]. \qquad (1)
\end{aligned}
$$

One approach to obtain an estimate of μ is to maximize the above likelihood with respect to θ.

$$
\hat{\mu}_{mle} = \text{argmin}_\theta \sum_{i=1}^{n}(y_i-\theta)^2
$$

Estimating Population Mean - MLE

We first take the derivative with respect to θ.

$$\frac{d}{d\theta}\sum_{i=1}^{n}(y_i-\theta)^2 = -2\sum_{i=1}^{n}(y_i-\theta) = -2\left[\sum_{i=1}^{n} y_i - n\theta\right].$$

The minimum can then be found by setting the derivative to zero.

$$-2\left[\sum y_i - n\theta\right] = 0 \quad \rightarrow \quad \hat{\mu} = \frac{\sum y_i}{n}.$$

We often use the sampling standard error of $\hat{\mu}$ to quantify the uncertainty in our estimate.

For normally distributed data, the mean is normally distributed. More generally, we have a central limit theorem argument.

Parameter is Fixed, Estimator is random

Note $E\hat{\mu} = E\sum_{i=1}^{n} \frac{y_i}{n} = \frac{n\mu}{n}$. And

$$\begin{aligned} Var\hat{\mu} &= Var\frac{\sum_{i=1}^{n} y_i}{n} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Note our population parameter is fixed, but the estimator is random:

$$[\hat{\mu}] = N(\mu, \sigma^2/n)$$

Frequentist confidence intervals

Frequentist inference relies on the *repeated experiments* paradigm. Arguably, confidence intervals have an unintuitive definition:

- Each experiment generates a 95% confidence interval. Then the true parameter μ will lie in the set of 95% confidence intervals 95% of the time.
- For my experiment, I calculate a 95% confidence interval. Then it either contains μ or it doesn't.
- The interval is random, and covers the true parameter 95% of the time.
- E.g., $P(L(X) \leq \mu \leq U(X)) = 0.95$, where $[L(X), U(X)]$ is an interval estimator, then we obtain estimate $\hat{L}(x_1, \ldots, x_n)$ and $\hat{U}(x_1, \ldots, x_n)$.

One strength of the Bayesian approach is that we estimate a posterior distribution and can define a 95% credible interval where the probability (given our prior and data) that μ is contained in that interval is 95%.

The Bayesian Approach and Prior

- Frequentist inference assumes a fixed population parameter μ.
- Bayesian inference places a distributional assumption on μ and treats μ as a random variable.

We start with a prior distribution:

Here we assume our parameter of interest μ is normal with known mean μ_0 and known variance τ^2.

Prior parameters (e.g. μ_0 and τ^2) are known as hyper-parameters.

Priors may reflect additional information from

- previous results from pilot studies
- knowledge from expert opinions.

or they may be uninformative.

Bayesian Inference

Given the prior distribution $[\,\mu\,]$, Bayesian inference is based on the conditional distribution $[\,\mu \,|\, \mathbf{y}\,]$.

- $[\,\mu \,|\, \mathbf{y}\,]$ is known as the posterior distribution.
- $[\,\mu \,|\, \mathbf{y}\,]$ describes the probability distribution of μ based on the observed data.
- Uncertainty in μ is quantified by probability.
- We can make statements about the probability of μ: Posterior (credible) interval: find constants a and b such that $P(a < \mu < b) = 0.95$.

Bayes Theorem

Recall Bayes Theorem:

For random variables μ and $\mathbf{y}$, we have:

Therefore, $[\,\mu \,|\, \mathbf{y}\,]$ can be calculated using the data likelihood $[\,\mathbf{y} \,|\, \mu]$ and the prior $[\,\mu\,]$.

Estimating Population Mean with known σ^2 - Bayesian

Consider again the problem of estimating the population mean μ:

$$y_1, y_2, \ldots, y_n \sim N(\mu, \sigma^2), \qquad \mu \sim N(\mu_0, \tau^2).$$

We want to calculate

$$[\,\mu \,|\, \mathbf{y}\,] = \frac{[\,\mathbf{y} \,|\, \mu] \times [\,\mu\,]}{\int [\,\mathbf{y} \,|\, \mu] \times [\,\mu\,] d\mu}.$$

First note that the denominator is a constant that ensures the posterior distribution is a valid density (i.e. integrates to 1). Specifically:

$$\begin{aligned}\int [\,\mu \,|\, \mathbf{y}\,] d\mu &= \int \frac{[\,\mathbf{y} \,|\, \mu] \times [\,\mu\,]}{\int [\,\mathbf{y} \,|\, \mu] \times [\,\mu\,] d\mu}\, d\mu \\ &= \frac{1}{\int [\,\mathbf{y} \,|\, \mu] \times [\,\mu\,] d\mu} \int [\,\mathbf{y} \,|\, \mu] \times [\,\mu\,] d\mu = 1\end{aligned}$$

Therefore $[\,\mu \,|\, \mathbf{y}\,]$ is proportional to just the data likelihood times the prior.

Estimating Population Mean with known σ^2 - Bayesian

We will often use the proportional constant idea to simplify algebra:

Estimating Population Mean with known σ^2 - Bayesian

$$[\,\mu\,|\,\mathbf{y}\,] \propto \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2] \times \exp[-\frac{1}{2\tau^2}(\mu-\mu_0)^2]$$

Estimating Population Mean with known σ^2 - Bayesian

$$= \exp\left[-\frac{1}{2}\left(\frac{n}{\sigma^2}(\mu - \bar{Y})^2 + \frac{1}{\tau^2}(\mu - \mu_0)^2\right)\right] \ldots \textit{ then expand the quadratic}$$

Estimating Population Mean with known σ^2 - Bayesian

$$[\,\mu\,|\,\mathbf{y}\,] \propto \exp\left[-\frac{1}{2}\left(\left[\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right]\mu^2 - 2\left[\frac{n}{\sigma^2}\bar{Y}+\frac{1}{\tau^2}\mu_0\right]\mu\right)\right] \ldots \textit{ factor out } \ldots$$

Let's denote

$$\tilde{\sigma}^2 = \left[\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right]^{-1} \qquad \text{and} \qquad \tilde{\mu} = \left[\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right]^{-1}\left[\frac{n}{\sigma^2}\bar{Y}+\frac{1}{\tau^2}\mu_0\right].$$

Then

$$[\,\mu\,|\,\mathbf{y}\,] \propto \exp\left[-\frac{1}{2\tilde{\sigma}^2}\left(\mu^2 - 2\tilde{\mu}\mu\right)\right] \ldots \textit{ then complete the squares}$$

Posterior Distribution for Normal Mean with known σ^2

$$[\,\mu\,|\,\mathbf{y}\,] \propto \exp\left[-\frac{1}{2\tilde{\sigma}^2}(\mu-\tilde{\mu})^2\right]$$

- The posterior distribution of μ given data **y** has a distribution that is proportional to the above form.
- Since $\tilde{\mu}$ and $\tilde{\sigma}^2$ do not include μ, the above corresponds to a Guassian density.

Therefore,

$$[\,\mu\,|\,\mathbf{y}\,] \sim N(\,\tilde{\mu},\,\tilde{\sigma}^2),$$

where

$$\tilde{\mu} = \left[\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right]^{-1}\left[\frac{n}{\sigma^2}\bar{Y}+\frac{1}{\tau^2}\mu_0\right] \text{ and } \tilde{\sigma}^2 = \left[\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right]^{-1}.$$

Revisiting BLUPs

$$\tilde{\mu} = \left[\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right]^{-1}\left[\frac{n}{\sigma^2}\bar{Y} + \frac{1}{\tau^2}\mu_0\right] \text{ and } \tilde{\sigma}^2 = \left[\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right]^{-1}.$$

$\tilde{\mu}$ is the posterior mean.

This looks like a BLUP from frequentist mixed models!

We examined mixed models from the frequentist perspective, but there are some connections to posterior distributions. Recall slide 40 from M2:

$$\hat{\theta}_i = \frac{(1/\tau^2)\mu + (r_i/\sigma^2)\bar{Y}_i}{1/\tau^2 + (r_i/\sigma^2)}$$

where μ was the population mean, $\bar{Y}_i$ was the group mean, τ^2 was the inter-subject variance, and σ^2 the measurement error.

The parameters are μ, τ^2, σ^2 – these are fixed in frequentist mixed models. They are random in Bayesian.

Revisiting BLUPs

$$\hat{\theta}_i = \left[\frac{r_i}{\sigma^2} + \frac{1}{\tau^2}\right]^{-1} \left[\frac{r_i}{\sigma^2}\bar{Y}_i + \frac{1}{\tau^2}\mu\right]$$

The BLUPs are the conditional distribution of the random effects given the data.

Can be thought of as the posterior distribution of the random effect given the group mean.

The population mean and random effect variance are treated as given and determine the "prior" distribution for θ_i.

Then the prior is updated by the mean of the group, $\bar{Y}_i$, with less weight when the measurement error (treated as given) is large.

Posterior Distribution for Normal Mean with known σ^2

Back to our posterior mean in the Bayesian world:

$$\tilde{\mu} = \left[\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right]^{-1} \left[\frac{n}{\sigma^2}\bar{Y} + \frac{1}{\tau^2}\mu_0\right] \text{ and } \tilde{\sigma}^2 = \left[\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right]^{-1}.$$

Here, it is a weighted average of the sample mean $\bar{Y}$ and our prior mean μ_0, with inverse-variance as weights!

Note that when $\tau^2 \to \infty$,

Posterior Distribution for Normal Mean with known σ^2

Consider the posterior distribution:

$$[\,\mu \,|\, \mathbf{y}\,] \sim N(\,\tilde{\mu},\, \tilde{\sigma}^2)$$

We say that the unknown parameter μ is distributed normally with mean $\tilde{\mu}$ and variance $\tilde{\sigma}^2$.

This is different from Frequentist inference where μ is assumed to be non-random:

$$\hat{\mu} \sim N(\,\mu,\, \sigma^2/n\,).$$

Remarks on Bayesian Inference

- The approach we used to find $[\,\mu \,|\, \mathbf{y}\,]$ is based only on the data likelihood and the prior on μ.
- No asymptotic theory is involved.
- Therefore there is no confidence interval or p-value for Bayesian inference.
- $[\,\mu \,|\, \mathbf{y}\,]$ is a distribution from which we can make inference directly.

  For example, we can ask
  - 95% Posterior (credible) interval: find constants a and b such that $P(a < \mu < b) = 0.95$.
  - Posterior hypothesis test: calculate $P(\mu > c)$, e.g., $c = 0$.

Normal Mean Example

Let's assume we have the following data

$$4.2, 6.6, 5.1, 2.0, 2.8, 3.2, 4.7, 4.1, 7.3, 0.8$$

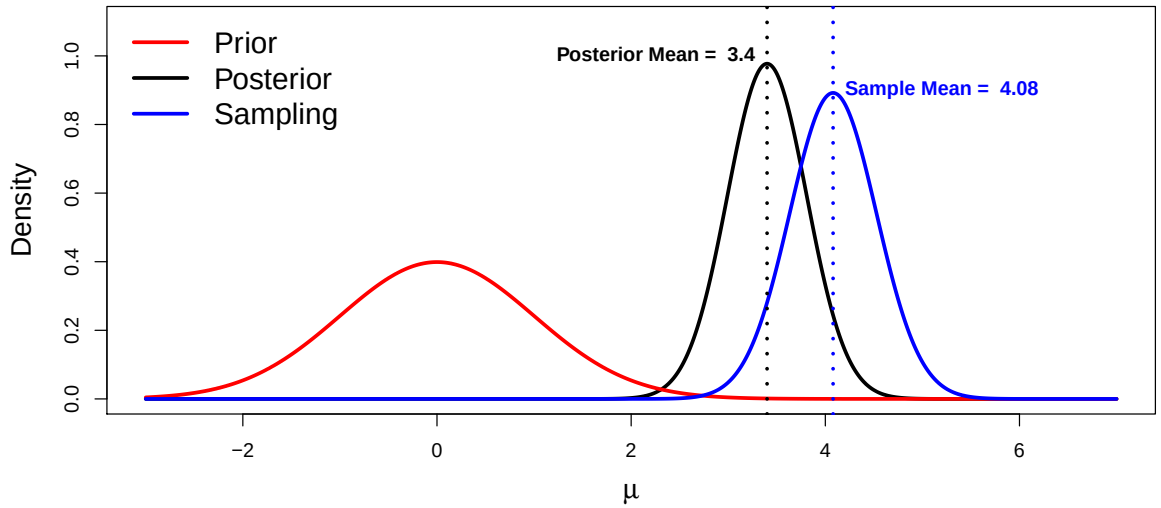
$$n = 10 \qquad \bar{Y} = 4.08 \qquad \sigma^2 = 2 \text{ (assume known)}$$

We will first assume prior $[\mu] = N(0, 1)$. Equation (5) gives the mean and variance of the Normal posterior distribution.

$$\tilde{\sigma}^2 = \left[\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right]^{-1} = \left[\frac{10}{2} + \frac{1}{1}\right]^{-1} = \frac{1}{6}.$$

$$\tilde{\mu} = \left[\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right]^{-1} \left[\frac{n}{\sigma^2}\bar{Y} + \frac{1}{\tau^2}\mu_0\right] = \frac{1}{6}\left[\frac{10}{2} \times 4.08 + \frac{1}{1} \times 0\right] = 3.4.$$

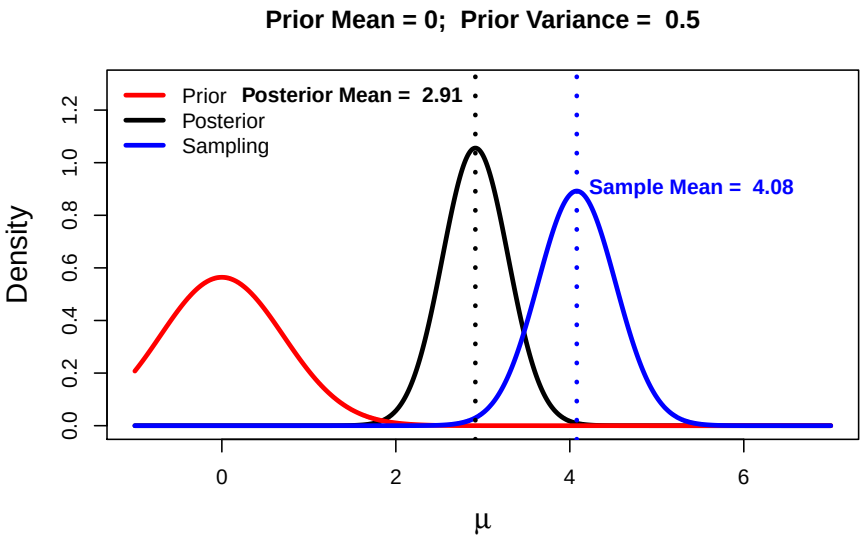
Normal Mean Example

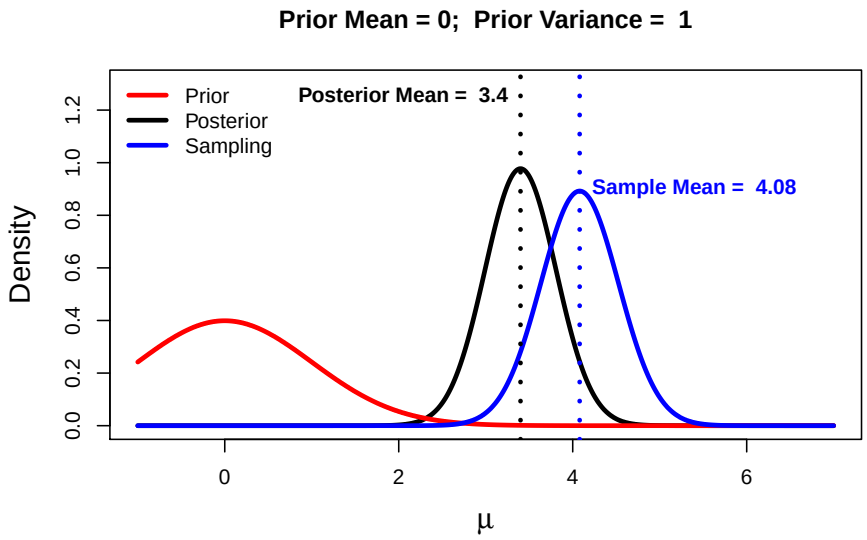


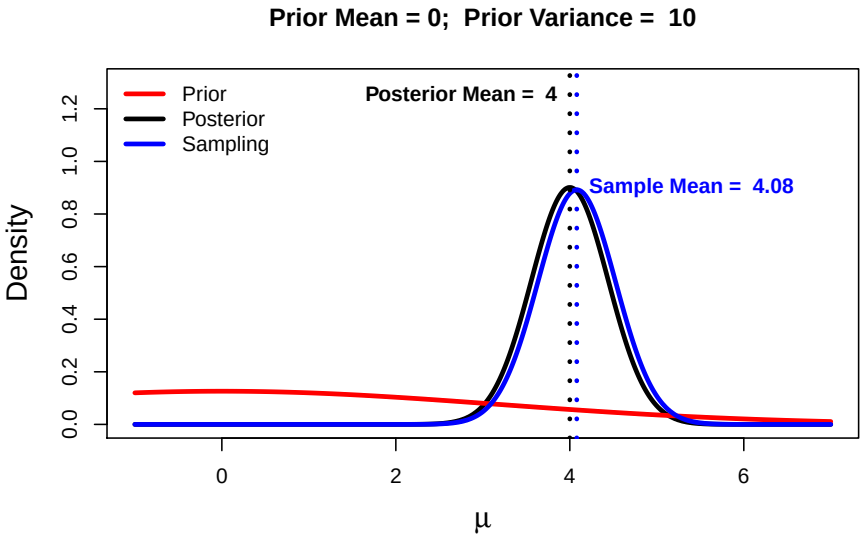
- Posterior mean is between the sample mean and the prior mean.
- Posterior standard deviation is smaller than the sampling variation.
- Posterior mean is *biased* according to frequentist view.
- A 95% posterior interval is $3.4 \pm 1.96 \times \frac{1}{\sqrt{6}}$.
- What is the probability $\mu > 2.5$?

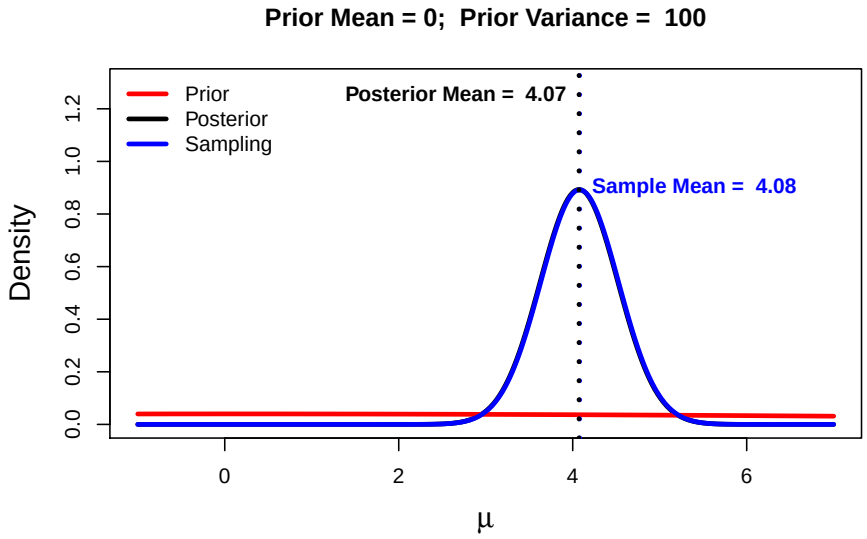
$$1 - \Phi\left(\frac{2.5 - 3.4}{\sqrt{1/6}}\right) = 0.986$$

Effects of Prior Variance (τ^2)

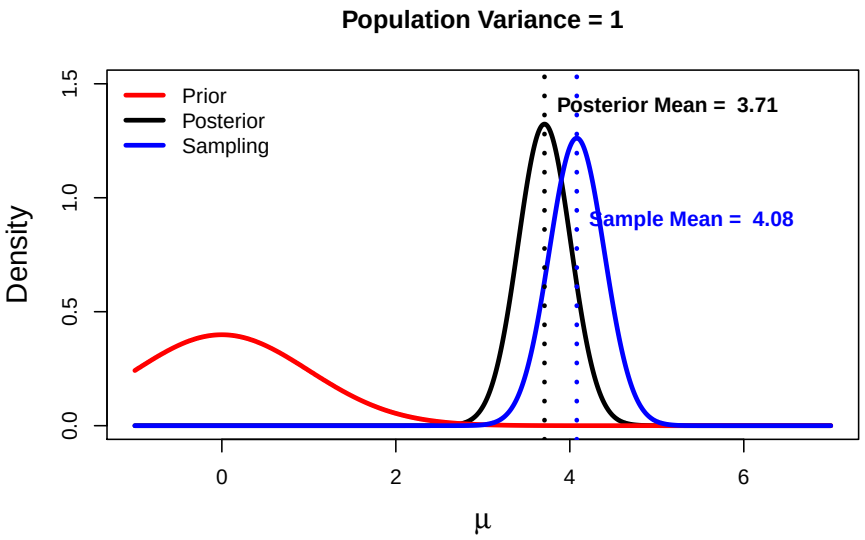


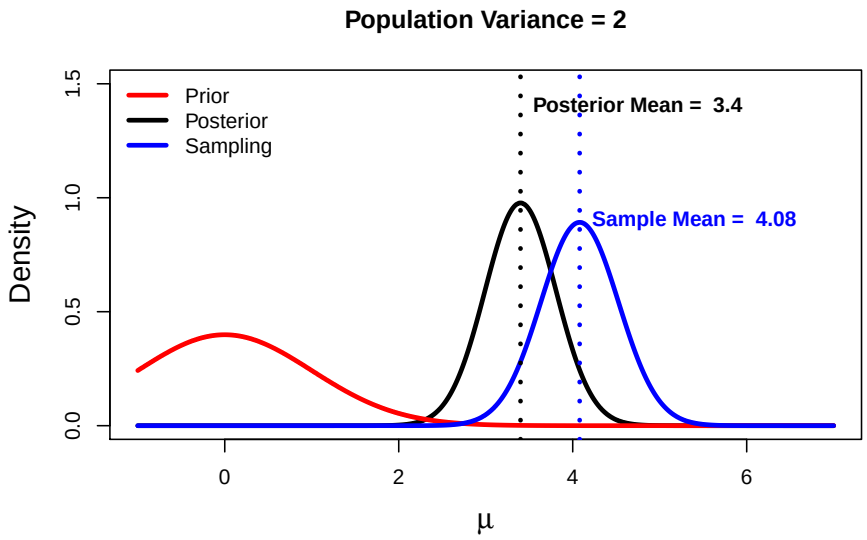


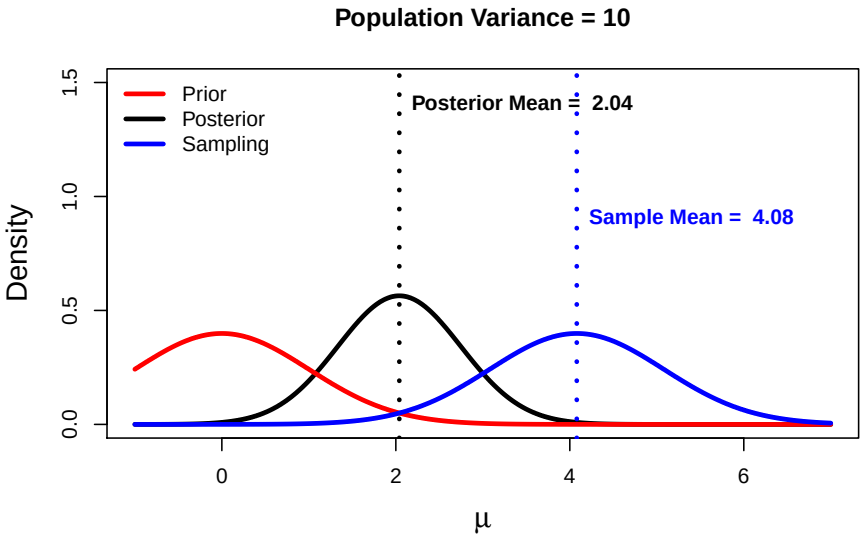


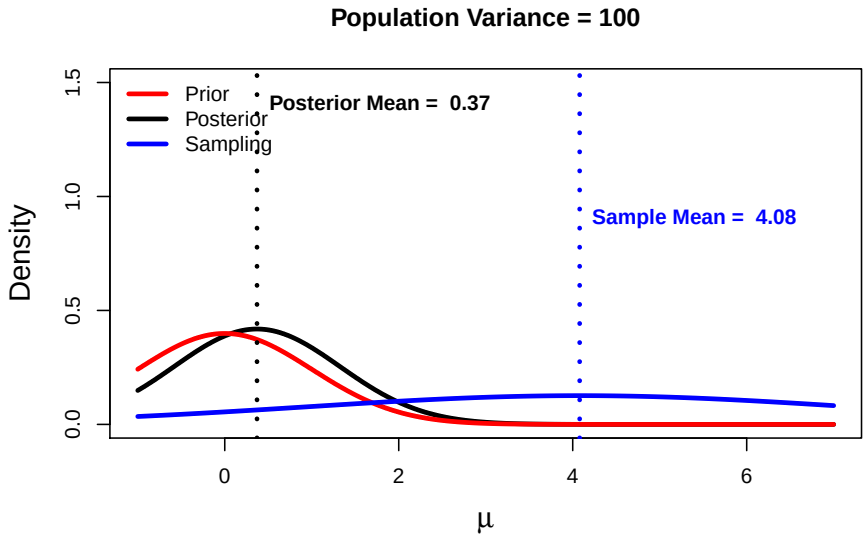


Effects of Population Variance (σ^2)

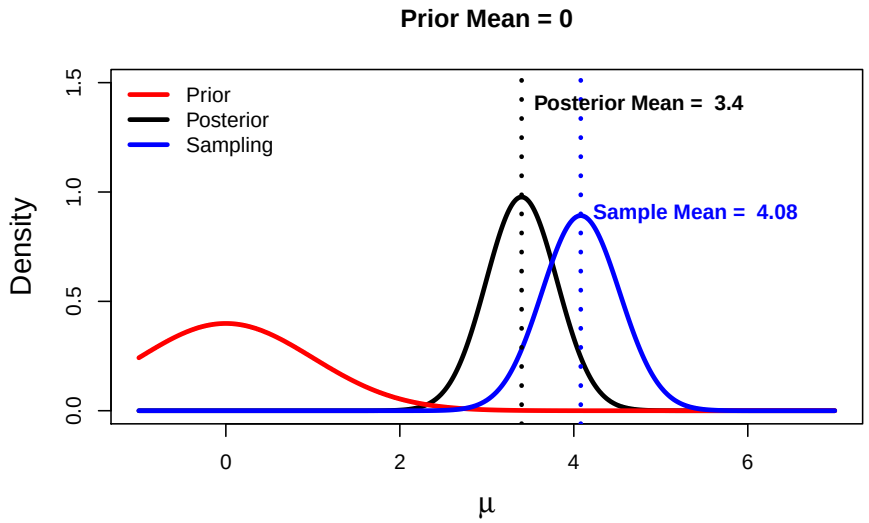


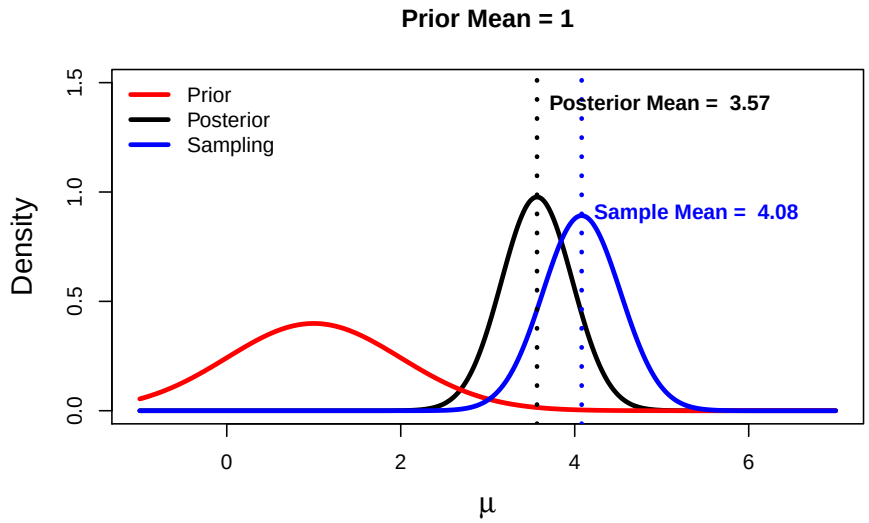


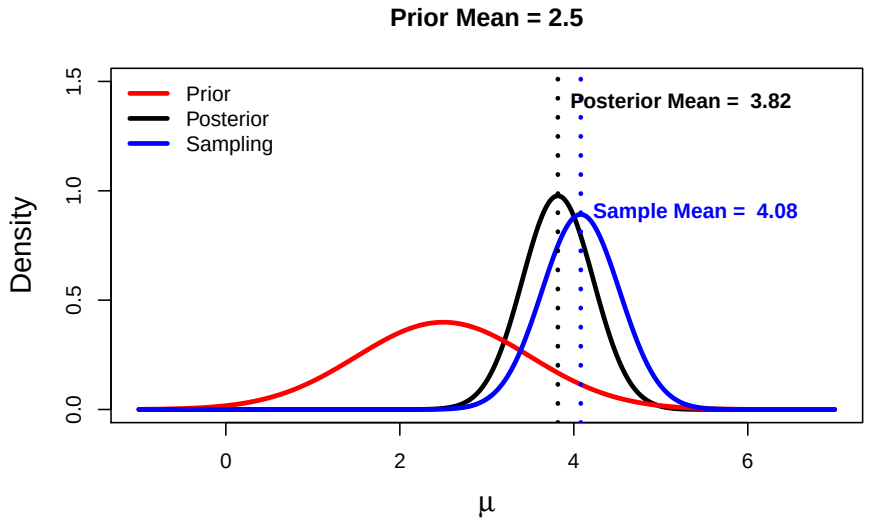


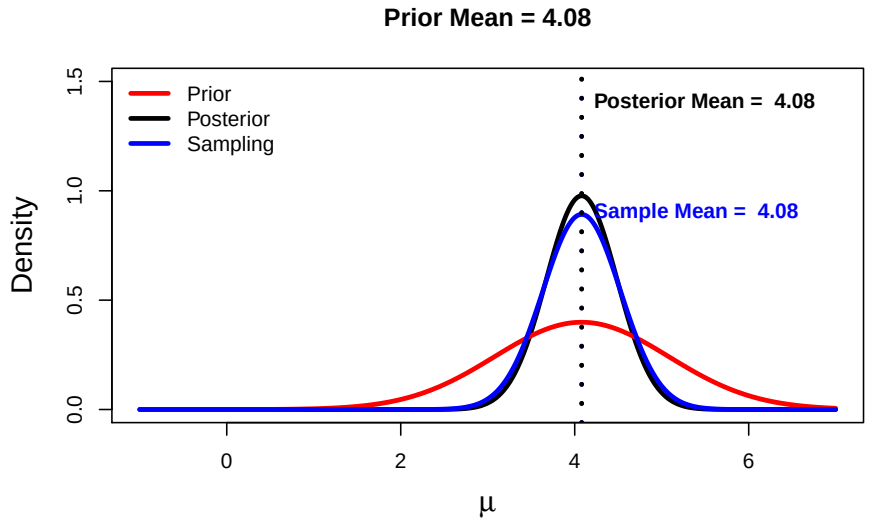


Effects of Prior Mean (μ_0)









Prior and Shrinkage

Effects of the prior distribution $\mu \sim N(\mu_0, \tau^2)$ on the posterior estimate of the population average μ depend on three quantities:

1. prior mean μ_0
2. prior variance τ^2
3. sample variance σ^2/n.

A prior's effect on the maximum likelihood estimate decreases with:

- large prior variance $\tau^2 \rightarrow$ high uncertainty in the prior information.
- small sampling variance $\sigma^2/n \rightarrow$ small uncertainty in the sample average.
- prior mean is close to the MLE.

In practice, priors with large variances are often used when the sample size is sufficient. For example $\mu \sim N(0, 10000^2)$.

Priors of this type are often called vague or uninformative prior.

Improper Priors

Consider the normal prior $\mu \sim N(\mu_0, \tau^2)$ which has density

Note that in the extremely uninformative case, $[\,\mu\,] \propto 1$ is not a valid density because it does not integrate to 1:

$$\int_{-\infty}^{\infty} 1 d\mu = \infty.$$

Let's assume that $[\,\mu\,] \propto 1$ to reflect our complete lack of prior knowledge on μ (as is often the case). These priors are referred to as being improper.

What would the posterior distribution be?

Improper Priors

$$
\begin{aligned}
[\,\mu \,|\, \mathbf{y}\,] &\propto [\,\mathbf{y} \,|\, \mu] \times [\,\mu\,] \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \times 1 = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\mu + \mu^2)\right] \\
&= \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\mu\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\mu^2\right)\right] \text{ ... } y_i^2 \textit{ does not involve } \mu \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\left(-2n\mu\bar{Y} + n\mu^2\right)\right] \text{ ... } \textit{factor } n \textit{ out} \\
&= \exp\left[-\frac{n}{2\sigma^2}\left(-2\mu\bar{Y} + \mu^2\right)\right] \textit{ ...complete the squares} \\
&\propto \exp\left[-\frac{n}{2\sigma^2}\left(-2\mu\bar{Y} + \mu^2 + \bar{Y}^2\right)\right] \\
&= \exp\left[-\frac{n}{2\sigma^2}(\mu-\bar{Y})^2\right]
\end{aligned}
$$

Therefore, the posterior is a valid density and it coincides with the MLE:

$$[\,\mu \,|\, \mathbf{y}\,] \sim N(\bar{Y},\, \sigma^2/n).$$

Road map

We will consider the simple case of estimating a mean and variance.

1. $[\mu|\mathbf{y}]$ for known σ^2.
2. $[\sigma^2|\mathbf{y}]$ for known μ.
3. $[\mu|\mathbf{y}]$ and $[\sigma^2|\mathbf{y}]$ when both μ and σ^2 are unknown.

Inference for Population Variance - MLE

Now we consider the case where we have known μ and unknown σ^2.

Let $y_1, \ldots, y_n$ be a sample of independent and identically distributed values drawn from a Normal population $N(\mu, \sigma^2)$.
Let θ now be the variance. Then the data likelihood is:

Maximize the log likelihood:

$$\ell(\theta; \mathbf{y}) = -\frac{n}{2}\log(2\pi\theta) - \frac{1}{2\theta}\sum_{i=1}^{n}(y_i - \mu)^2$$

$$= -\frac{n}{2}\log(\theta) - \frac{1}{2\theta}\sum_{i=1}^{n}(y_i - \mu)^2 + \text{constant.}$$

Inference for Population Variance with known μ - MLE

$$\frac{d}{d\theta}\ell(\theta;\mathbf{y}) = \frac{d}{d\theta}\left[-\frac{n}{2}\log(\theta) - \frac{1}{2\theta}\sum_{i=1}^{n}(y_i-\mu)^2\right]$$
$$= -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(y_i-\mu)^2.$$

Setting to 0,

$$-\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(y_i-\mu)^2 = 0$$
$$-n\theta + \sum_{i=1}^{n}(y_i-\mu)^2 = 0$$

Finally,

$$\hat{\sigma}^2_{mle} = \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{n}.$$

Inference for Population Variance with known μ - MLE

We now need to derive the sampling distribution of $\hat{\sigma}^2_{mle}$. Recall that if $y \sim N(\mu, \sigma^2)$,

$$\left(\frac{y_i - \mu}{\sigma}\right) \sim N(0,1) \qquad \rightarrow \qquad \left(\frac{y_i - \mu}{\sigma}\right)^2 \sim \chi^2_1 \, .$$

Then

$$\sum_{i=1}^{n} \left(\frac{y_i - \mu}{\sigma}\right)^2 \sim \chi^2_n \qquad \rightarrow \qquad \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \sim \chi^2_n \, .$$

Therefore

$$\hat{\sigma}^2_{mle} = \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{n} \sim \frac{\sigma^2}{n} X, \qquad \text{where } X \sim \chi^2_n \, .$$

So the sampling distribution of $\hat{\sigma}^2_{mle}$ is a scaled chi-squared distribution.

Bayesian case: The Inverse-Gamma Distribution

The most common distribution to model variance components is the inverse-gamma distribution, which has two parameters $\alpha > 0$ (shape) and $\beta > 0$ (scale). For a random variable y with support $y > 0$, it has density

$$f(\theta; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\frac{\beta}{\theta}}.$$

Inv-gamma is a skewed distribution for positive random variables. It has

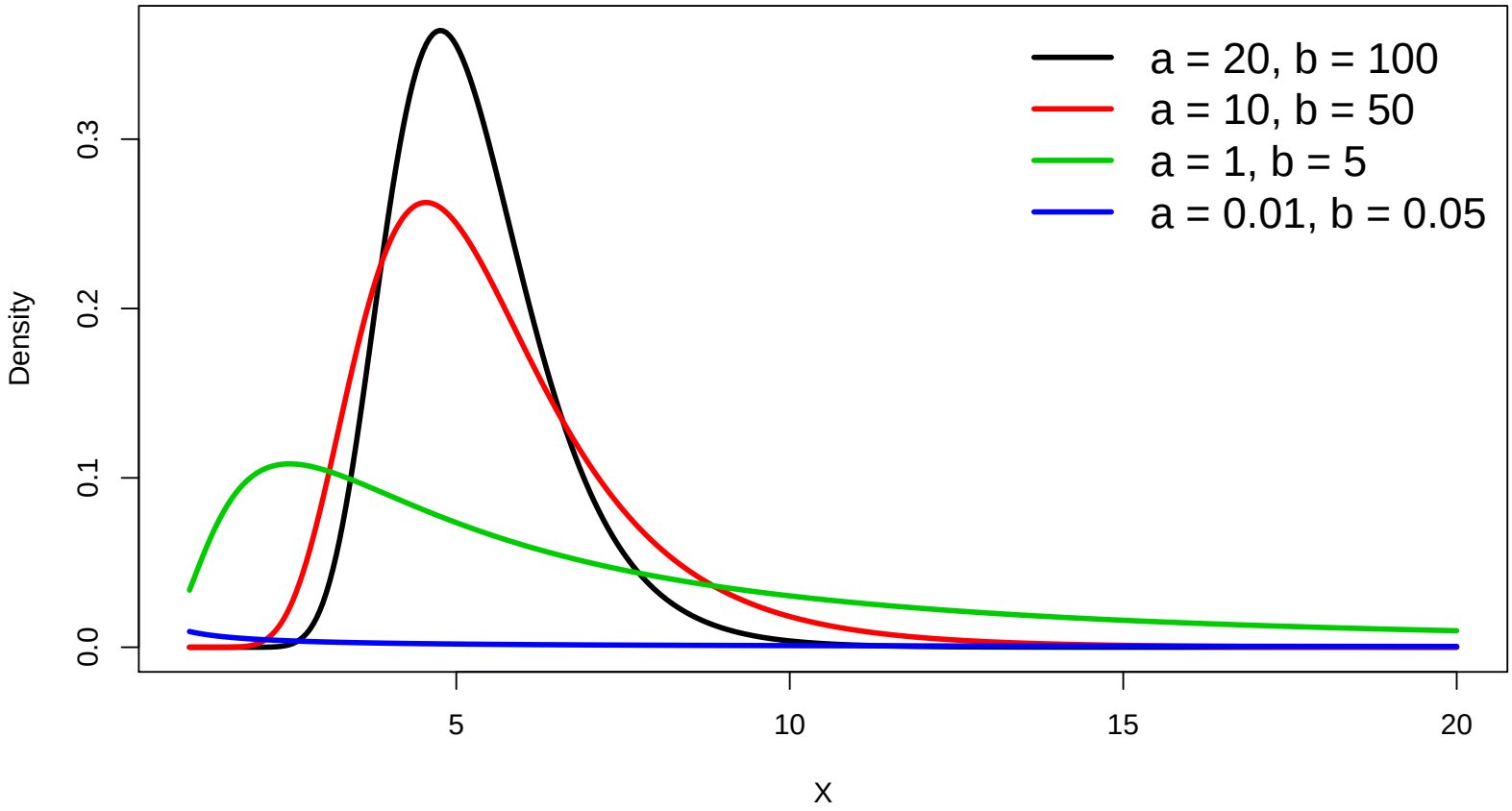
$$E[y] = \frac{\beta}{\alpha - 1} \qquad \text{and} \qquad Var[y] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$$

Also, if $\theta \sim$Inv-gamma (α, β), then $\tilde{\theta} = 1/\theta \sim$ Gamma (α, β) with density:

$$f(\tilde{\theta}; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tilde{\theta}^{(\alpha-1)} e^{-\beta\tilde{\theta}}.$$

Note $\chi^2_{df} = \text{Gamma}(df/2, 2)$.

Inverse-Gamma (a, b) Density



Bayesian Inference for Population Variance with known μ

Let's assume an inverse-gamma prior on $\sigma^2 \sim$ inv-gamma(α_0, β_0). Then the posterior distribution $[\sigma^2|\mathbf{y}]$ is given by

$$
\begin{aligned}
[\sigma^2|\mathbf{y}] &\propto [\mathbf{y}|\sigma^2] \times [\sigma^2] \\
&= \prod_{i=1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right] \times \frac{\beta_0^\alpha}{\Gamma(\alpha_0)}(\sigma^2)^{-(\alpha_0+1)} e^{-\frac{\beta_0}{\sigma^2}} \\
&\propto (\sigma^2)^{-n/2} \exp\left[\sum_{i=1}^{n} -\frac{1}{2\sigma^2}(y_i-\mu)^2\right](\sigma^2)^{-(\alpha_0+1)} e^{-\frac{\beta_0}{\sigma^2}} \\
&= (\sigma^2)^{-n/2-(\alpha_0+1)} \exp\left[\sum_{i=1}^{n} -\frac{1}{2\sigma^2}(y_i-\mu)^2 - \frac{\beta_0}{\sigma^2}\right] \\
&= (\sigma^2)^{-[(n/2+\alpha_0)+1]} \exp\left[-\frac{1}{\sigma^2}\left(\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2 + \beta_0\right)\right]
\end{aligned}
$$

Treating σ^2 as the random variable, the above density form is distributed:

$$
\text{Inv-gamma}\left(n/2+\alpha_0,\ \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2 + \beta_0\right).
$$

Bayesian Inference for Population Variance with known μ

The inverse-gamma prior distribution on σ^2 has mean $\frac{\beta_0}{\alpha_0 - 1}$. The posterior inverse-gamma distribution has mean

Intuitive view of hyperparameters:

α_0 and β_0 correspond to a pilot study that estimated σ^2 with $2\alpha_0$ sample size, and sum of squares $2\beta_0$.

Smaller values of $\alpha_0 =$ less informative. $\alpha_0 = 1$ is a pilot study with 2 observations.

Bayesian Inference for Population Variance with known μ

$$E[\sigma^2|\mathbf{y}] = \frac{\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2 + \beta_0}{n/2 + \alpha_0 - 1}.$$

Note when $\alpha_0 \to 0$ and $\beta_0 \to 0$ (less prior information), posterior expectation approaches MLE.

Also

$$E[\sigma^2|\mathbf{y}] \overset{n}{\to} \hat{\sigma}^2_{mle}$$

Prior and posterior distributions

Again assume we have the following data

$$4.2, 6.6, 5.1, 2.0, 2.8, 3.2, 4.7, 4.1, 7.3, 0.8$$

$$n = 10 \qquad \mu = 4 \text{ (assume known)} \qquad \sum_{i=1}^{10}(y_i - \mu) = 35.72$$

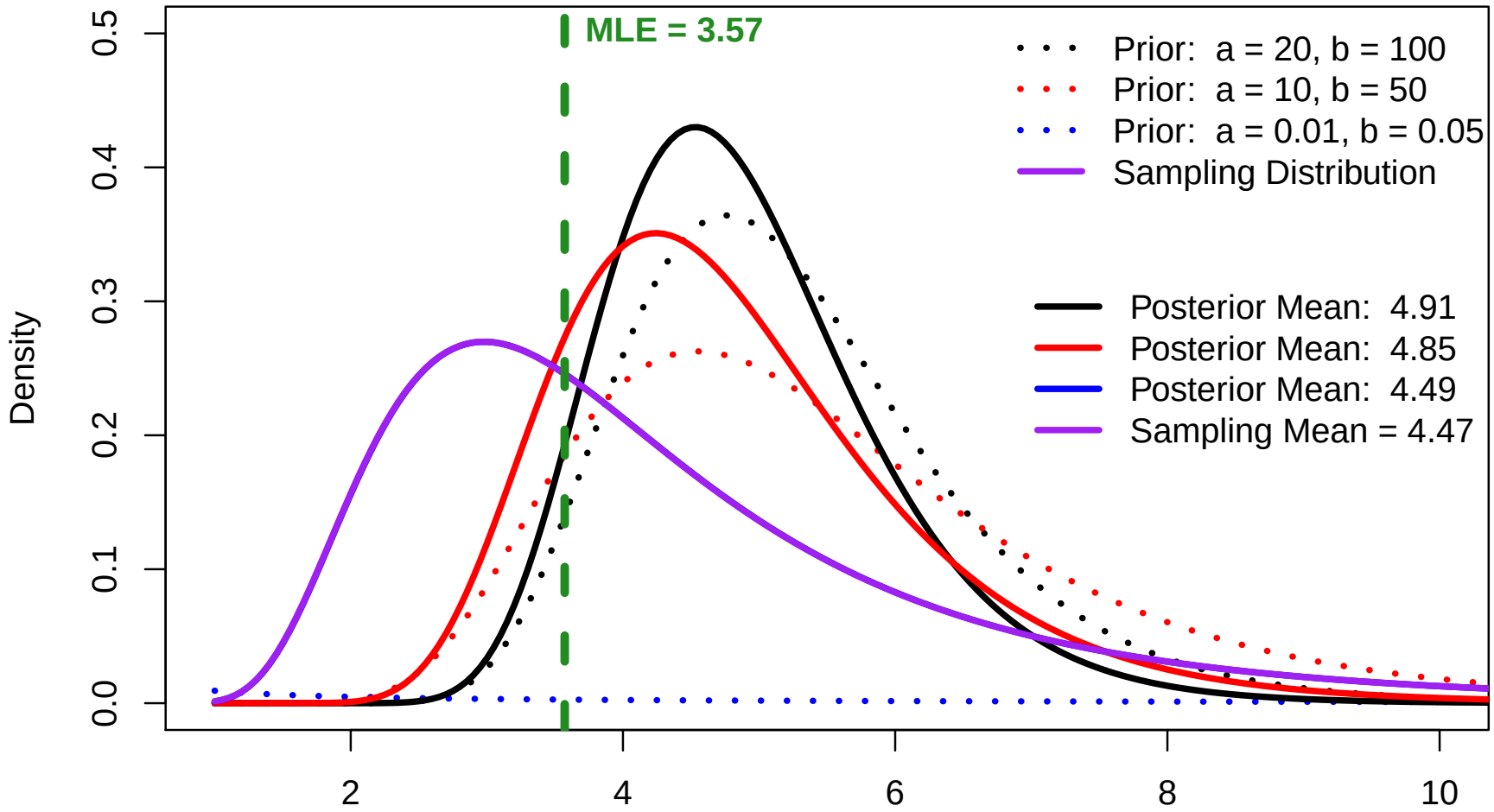
MLE: $\frac{1}{n}\sum_{i=1}^{n}(y_i - 4)^2$

Sampling distribution: Inv-gamma$(10/2, \sum(y_i - 4)^2/2)$

Mean (expected value) of the sampling distribution $= \frac{\sum(y_i - 4)^2/2}{10/2-1}$.

Note here an uninformative prior converges to mean of the sampling distribution $= \frac{\sum(y_i - 4)^2/2}{10/2-1}$.

Effects of Inverse Gamma Prior



Prior and Posterior Distributions

- For estimating population mean and variance, we used prior distributions that are the same as their corresponding posterior distributions.
- These priors are special and are known as conjugate priors. They are convenient because the posterior density is available in closed-form.
- We can often find uninformative/vague priors such that the posterior distributions are very close to that from a Frequentist inference.
- In practice, data analyses usually utilize uninformative (proper or improper) priors.
- Prior information is particularly useful when the sample size is small.
- Posterior distributions can be viewed as the prior distributions being updated by the data.

Road map

We will consider the simple case of estimating a mean and variance.

1. $[\mu|\mathbf{y}]$ for known σ^2.
2. $[\sigma^2|\mathbf{y}]$ for known μ.
3. $[\mu|\mathbf{y}]$ and $[\sigma^2|\mathbf{y}]$ when both μ and σ^2 are unknown.

Inference for Multiple Parameters

Let $y_1, y_2, \ldots, y_n$ be a sample of independent and identically distributed values drawn from a Normal population $N(\mu, \sigma^2)$. We now assume both μ and σ^2 are unknown. We can write the model as:

This gives us a joint posterior distribution:

Here we assume the prior distributions for μ and σ^2 are independent. We are often interested in making inference for each parameter separately. This is accomplished by determining the marginal posterior distribution

$$[\mu|\mathbf{y}] \quad \text{and} \quad [\sigma^2|\mathbf{y}].$$

Big Picture: Joint, marginals, and full conditionals

- In most cases, we will not have a closed form for the joint posterior. We saw a few exceptions for single-parameter models.
- In most cases, we will not have a closed form for marginal posteriors. We will look at an example next where we can calculate the marginal posteriors for certain choices of improper priors.
- In the absence of the joint and marginal posteriors, we can sometimes derive the full conditional distributions. In such cases, we can use an algorithm called Gibbs Sampling to generate samples from the joint posterior. We will see an example of this later in this module.
- We will often not be able to calculate the full conditionals, either, but rather the conditionals up to a constant. Then we can pursue the Metropolis-Hastings Algorithm. We will not go into this, but will instead rely upon software.

Inference for Population Mean with Variance Unknown

The marginal posterior distribution of μ can be written as

Note that

$$\int [\mathbf{y}|\mu, \sigma^2] \times [\sigma^2]\, d\sigma^2$$

can be interpreted as an adjusted data likelihood function that is integrated (averaged) over σ^2 for a given μ. In order words, the marginal posterior distribution $[\mu|\mathbf{y}]$ incorporates the uncertainty in σ^2.

The greater ability to account for parameter uncertainty is one of the advantages of Bayesian inference.

Inference for Population Mean with Variance Unknown

We will assign the following uninformative priors:

$$[\mu] \propto 1 \qquad \text{and} \qquad [\sigma^2] \propto (\sigma^2)^{-1}.$$

The prior on σ^2 corresponds to $\log(\sigma^2) \propto 1$. Then

$$\begin{aligned}
[\mu, \sigma^2|\mathbf{y}] &\propto \prod_{i=1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right] \times (\sigma^2)^{-1} \\
&\propto (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \\
&= (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2\right)\right] \\
&= (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y}-\mu)^2\right]\right]
\end{aligned}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$, the sample variance.

Inference for Population Mean with Variance Unknown

Now,

$$[\mu|\mathbf{y}] \propto \int_0^\infty (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y}-\mu)^2\right]\right] d\sigma^2.$$

Let $Z = (n-1)s^2 + n(\bar{y}-\mu)^2$ and $u = \frac{Z}{2\sigma^2}$. Towards a change of variable, $d\sigma^2/du = -Z/(2u^2)$, we obtain

$$= \int_0^\infty \left(\frac{Z}{2u}\right)^{-n/2-1} \exp(-u) \mid -\frac{Z}{2u^2} \mid du$$

$$\propto Z^{-n/2} \int_0^\infty u^{n/2-1} \exp(-u)\, du \ldots \text{ proportional to gamma(n/2, 1), integrates to constant}$$

$$\propto [(n-1)s^2 + n(\bar{y}-\mu)^2]^{-n/2}$$

$$\propto \left[1 + \frac{n(\bar{y}-\mu)^2}{(n-1)s^2}\right]^{-n/2}.$$

Inference for Population Mean with Variance Unknown

We have

$$[\mu|\mathbf{y}] \propto \left[1 + \frac{n(\bar{y}-\mu)^2}{(n-1)s^2}\right]^{-n/2}$$
$$= \left[1 + \left(\frac{\bar{y}-\mu}{s/\sqrt{n}}\right)^2 /(n-1)\right]^{-\{(n-1)+1\}/2}$$

The above is a t distribution with $(n-1)$ degrees of freedom:

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

Note this is for the special improper priors we chose.

Bayesian Computation

Consider a model with multiple parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$. Here $\boldsymbol{\theta}$ may include regression coefficients and residual variance. It becomes increasingly difficult to perform integration. For example, to make inference on θ_1 we need

$$[\theta_1|\mathbf{y}] = \int \cdots \int [\theta_1, \theta_2, \ldots, \theta_p \,|\, \mathbf{y}]\, d\theta_2 d\theta_3 \ldots d\theta_p.$$

Often, the desired marginal distribution is also not a standard probability density.

One approach is to perform numerical integration. Specifically, we will generate samples/realizations from the joint distribution.

$$\boldsymbol{\theta}^{(k)} \sim [\theta_1, \theta_2, \ldots, \theta_p \,|\, \mathbf{y}].$$

We can then make inference by examining the marginal distribution of the sampled values.

Bayesian Inference via Monte Carlo Simulations

Given K samples of $\mu^{(k)}$, we can compute several useful statistics for inference.

Posterior mean:

Posterior variance:

Exceedance probability:

Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is an algorithm to obtain realizations from a high-dimensional probability density. When full conditionals are available, we can use the Gibbs Sampler, described below. Assume we have three parameters θ_1, θ_2, and θ_3.

1. Set initial values of θ_1, θ_2, and θ_3.
2. For the $k+1$ iteration
   2.1 Draw (update) a new $\theta_1^{(k+1)}$ from it's full conditional distribution:
   $$\theta_1^{(k+1)} \sim [\theta_1 | \theta_2^{(k)}, \theta_3^{(k)}, \mathbf{y}].$$
   2.2 Draw (update)
   $$\theta_2^{(k+1)} \sim [\theta_2 | \theta_1^{(k+1)}, \theta_3^{(k)}, \mathbf{y}].$$
   2.3 Draw (update)
   $$\theta_3^{(k+1)} \sim [\theta_3 | \theta_1^{(k+1)}, \theta_2^{(k+1)}, \mathbf{y}].$$
3. Repeat step 2 until convergence.
4. Discard the first B-many samples as burn-in (pre-convergence) samples.

Note that the generated samples $\theta_1^{(1)}, \theta_1^{(2)}, \theta_1^{(3)}, \ldots$ are dependent.

Gibbs Sampler

Assume the following priors:

The joint posterior is proportional to

$$
\begin{aligned}
[\mu, \sigma^2] &\propto [\mathbf{y}|\mu, \sigma^2][\mu][\sigma^2] \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right] \\
&\times (2\pi\tau^2)^{-1/2} \exp[-\frac{1}{2\tau^2}(\mu - \mu_0)^2] \times \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-(\alpha+1)} e^{-\frac{\beta}{\sigma^2}}
\end{aligned}
$$

Goal: marginal posterior

$$
\begin{aligned}
[\mu|\mathbf{y}] &= \int_0^\infty [\mu|\mathbf{y}, \sigma^2][\sigma^2] d\sigma^2 \\
&\propto \int_0^\infty \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(y_i - \mu)^2] \\
&\quad \times (2\pi\tau^2)^{-1/2} \exp[-\frac{1}{2\tau^2}(\mu - \mu_0)^2] \times \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-(\alpha+1)} e^{-\frac{\beta}{\sigma^2}} \, d\sigma^2
\end{aligned}
$$

How do we calculate this?

We will sample from the full conditionals, which here have a simple form. This is called Gibbs sampling.

This will generate a sample from the joint posterior. Then we can calculate statistics, e.g., the mean, from the values corresponding to μ, which implicitly are weighted according to the distribution of σ^2. Then the statistics are an estimate of the marginal posterior statistic, e.g., posterior mean.

Full conditionals

The conditional distribution of μ given σ^2 is the same as the distribution of μ for known σ^2. We already derived it!

We found $[\mu|\sigma^2, \mathbf{y}]$ is Normal with mean $\tilde{\mu}$ and variance $\tilde{\sigma}^2$:

$$\tilde{\mu} = \left[\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right]^{-1} \left[\frac{n}{\sigma^2}\bar{Y} + \frac{1}{\tau^2}\mu_0\right] \text{ and } \tilde{\sigma}^2 = \left[\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right]^{-1}.$$

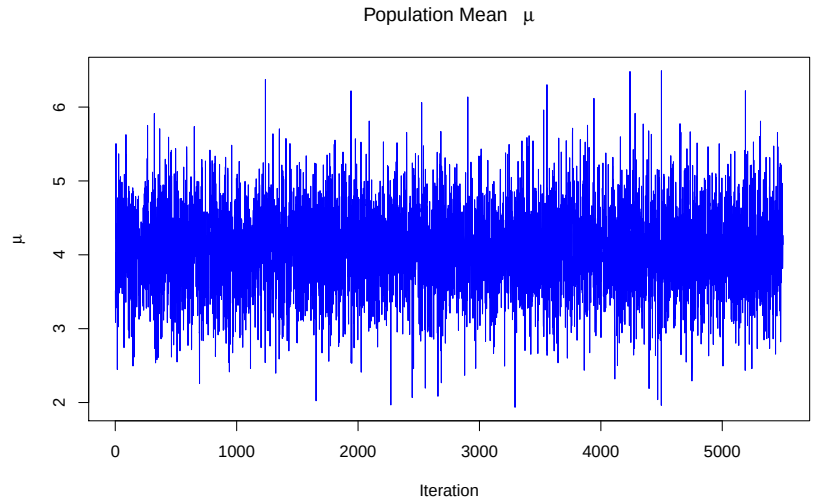
The conditional distribution of σ^2 given μ is the same as the distribution of σ^2 for known μ. We already derived it!!

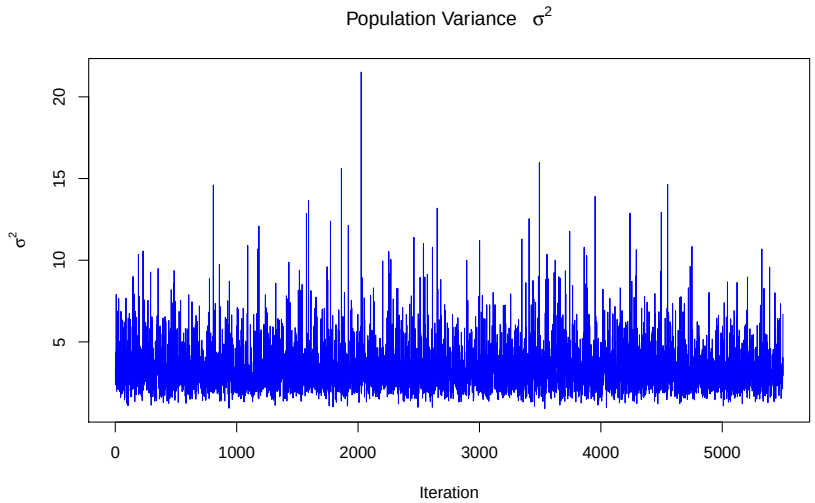
$$[\sigma^2|\mu, \mathbf{y}] \sim \text{Inv-gamma}\left(n/2 + \alpha,\ \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2 + \beta\right).$$

Example R Code

```
y = c(4.2, 6.6, 5.1, 2.0, 2.8, 3.2, 4.7, 4.1, 7.3, 0.8) ## Data

n.iter = 2000
sigmasq.save = rep (NA, n.iter)
mu.save = rep(NA, n.iter)

mu = 0         ### Initial values for mu
sigmasq = 1      ### Initial values for sigma2
mu0=0          ### Prior mean for mu
tausq= 5^2       ### Prior variance for mu
n= length(y)  ### Sample size
a = 2         ### Prior alpha value for sigma2
b = 1         ### Prior beta value for sigma2

for (i in 1:n.iter){
  # calculate conditional dist of mu given sigmasq and data
  # update variance:
  sigmasqk = 1 / (n/sigmasq+1/tausq)
  # update mean:
  meank = (mean(y)/(sigmasq/n) + mu0/tausq)*sigmasqk
mu = rnorm (1, meank, sqrt(sigmasqk))
mu.save[i] = mu

# calculate conditional dist of sigmasq given mu and data
RRR = y - mu
sigmasq = rinvgamma(1,shape=n/2+a,scale=sum(RRR^2)/2 + b)
sigmasq.save[i] = sigmasq
}
```

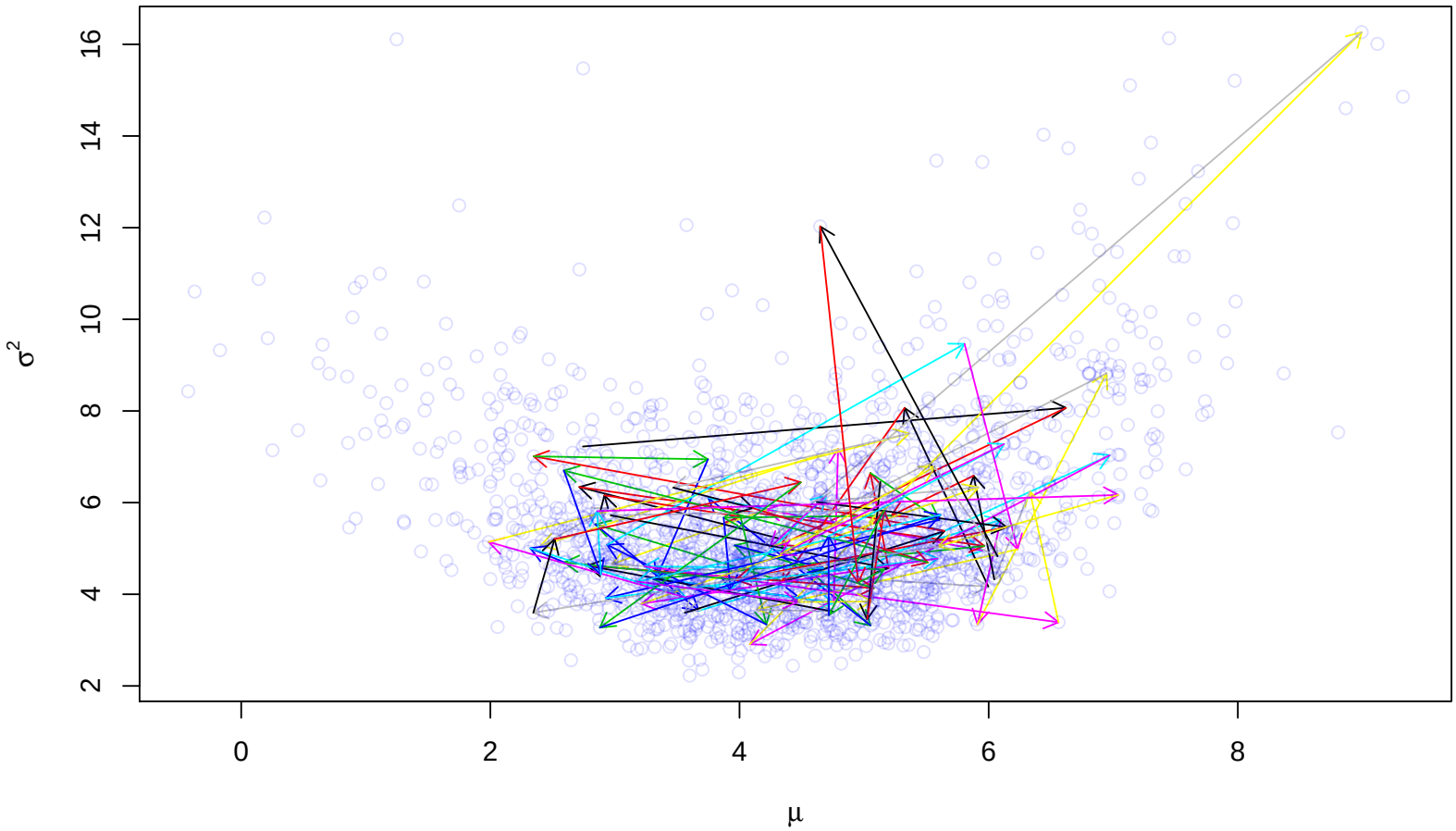
Trace Plots


Population Mean μ
μ
2
3
4
5
6
0
1000
2000
3000
4000
5000
Iteration


Population Variance σ2
σ2
5
10
15
20
0
1000
2000
3000
4000
5000
Iteration

Gibbs Sampler Paths

The first 100 updates are shown.



Joint Distribution of σ^2 and μ