GPH-GU2372/3372
Applied Bayesian Analysis in Public Health

Lecture 3: The normal model

Hai Shu, PhD

09/26/2022

Normal model: Topics

- Normal model with known variance
- Normal model with known mean
- Normal model with unknown mean and variance

Normal distribution

- Let Y be a continuous random variable with sample space $\mathcal{Y} = \mathbb{R}$. We say that Y has a normal distribution $N(\mu, \sigma^2)$ if, conditional on two parameters μ and σ^2 representing, respectively, the center and the spread of the density, it has a density of the form:

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}), \qquad y \in \mathbb{R}$$

where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

Normal distribution

- The normal distribution plays an important role because of the Central Limit Theorem that states that, regardless of their distribution, the average of a sequence of n i.i.d. random variables has a normal distribution as n gets large.

- Approximately 95% of the population lies within 2 standard deviations (σ) of the population mean, μ, and approximately 99% lies within 3 standard deviations.

- If Y_1 and Y_2 are two independent random variables that, conditionally on parameters μ_1, μ_2, σ_1^2 and σ_2^2, follow a normal distribution, then:

$$\left.\begin{array}{l} Y_1 \sim N(\mu_1, \sigma_1^2) \\ Y_2 \sim N(\mu_2, \sigma_2^2) \\ Y_1 \perp\!\!\!\perp Y_2 \end{array}\right\} \Longrightarrow Y_1 + Y_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Normal distribution

- Example: A study of 1,100 English families from 1893 to 1898 gathered height data on $n = 1375$ women over the age of 18. The sample mean for these data is $\bar{y} = 63.75$ inches and the sample standard deviation is $s = 2.62$ inches.

- Let Y_i denote the height (in inches) of woman i.

- Since we can think of a person's height as the result of the additive action of several factors, such as genetics, diet, disease, and so on, we can model $Y_1, ..., Y_n$ as random variables that **conditionally** on the two parameters μ and σ^2, follow a $N(\mu, \sigma^2)$ distribution.

Normal distribution

- Let's suppose that the following sampling model holds for our data: $Y_1, ..., Y_n$ are conditionally independent and identically distributed given μ and σ^2 with $p(y|\mu, \sigma^2) = N(\mu, \sigma^2)$. Then:

$$
\begin{aligned}
p(y_1, ..., y_n|\mu, \sigma^2) &= \prod_{i=1}^n p(y_i|\mu, \sigma^2) = \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\tfrac{1}{2} \tfrac{(y_i-\mu)^2}{\sigma^2}) \right\} \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp(-\tfrac{1}{2} \sum_{i=1}^n \tfrac{(y_i-\mu)^2}{\sigma^2}) \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\tfrac{1}{2\sigma^2} [\sum_{i=1}^n y_i^2 + n\mu^2 - 2\mu \sum_{i=1}^n y_i]\right)
\end{aligned}
$$

- The density $p(y_1, ..., y_n|\mu, \sigma^2)$ depends on the data $y_1, ..., y_n$ only through two statistics: $\sum_{i=1}^n y_i^2$ and $\sum_{i=1}^n y_i \implies$ they are sufficient statistics.

- Since they are functions of $\sum_{i=1}^n y_i^2$ and $\sum_{i=1}^n y_i$, the sample variance s^2 and the sample mean $\bar{y}$ are also sufficient statistics.

Normal distribution: with known variance

- Consider the situation where $Y_1, ..., Y_n$ are conditionally i.i.d. with $p(y|\mu, \sigma^2) = N(\mu, \sigma^2)$ with σ^2 **known**.
- We want to infer upon μ assuming that we know σ^2. If we place a prior $p(\mu)$ on μ, the posterior distribution $p(\mu|y_1, ..., y_n, \sigma^2)$ for μ is given by:

$$\begin{aligned} p(\mu|y_1, ..., y_n, \sigma^2) &\propto p(y_1, ..., y_n|\mu, \sigma^2) \cdot p(\mu) \\ &= \{\textstyle\prod_{i=1}^n p(y_i|\mu, \sigma^2)\} \cdot p(\mu) \\ &= \{\textstyle\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2})\} \cdot p(\mu) \\ &\propto \exp(-\frac{1}{2}\textstyle\sum_{i=1}^n \frac{(y_i-\mu)^2}{\sigma^2}) \cdot p(\mu) \end{aligned}$$

Normal distribution: with known variance

- Since
$$p(\mu|y_1,...,y_n,\sigma^2) \propto \exp(-\tfrac{1}{2}\textstyle\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2})\cdot p(\mu)$$
a prior $p(\mu)$ is a conjugate prior if it contains terms of the form $\exp(-c_2(\mu-c_1)^2)$.

- The normal distribution is such a density. Let's verify that it provides a conjugate prior for μ.
Let's assume $p(\mu)=N(\mu_0,\tau_0^2)$, then:
$$\begin{aligned} p(\mu|y_1,...,y_n,\sigma^2) &\propto \exp(-\tfrac{1}{2}\textstyle\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2})\cdot\exp(-\tfrac{1}{2}\frac{(\mu-\mu_0)^2}{\tau_0^2}) \\ &= \exp(-\tfrac{1}{2}[\textstyle\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2}+\frac{(\mu-\mu_0)^2}{\tau_0^2}]) \end{aligned}$$

- Let's focus on the term inside the square bracket ignoring the factor $\frac{1}{2}$ and let's look at it as a function of μ only.

Normal distribution: with known variance

- We have:

$$\begin{aligned}\sum_{i=1}^{n} \frac{(y_i-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\tau_0^2} &= \mu^2\left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}\right) - 2\mu\left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2} + \frac{\mu_0}{\tau_0^2}\right) \\ &\quad+\left(\frac{\sum_{i=1}^{n} y_i^2}{\sigma^2} + \frac{\mu_0^2}{\tau_0^2}\right) \\ &= A\mu^2 - 2B\mu + C\end{aligned}$$

- where $A = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}$, $B = \frac{\sum_{i=1}^{n} y_i}{\sigma^2} + \frac{\mu_0}{\tau_0^2}$ and $C = \frac{\sum_{i=1}^{n} y_i^2}{\sigma^2} + \frac{\mu_0^2}{\tau_0^2}$.

Normal distribution: with known variance

$$
\begin{aligned}
p(\mu|y_1,...,y_n,\sigma^2) &\propto \exp(-\tfrac{1}{2}[\textstyle\sum_{i=1}^{n} \frac{(y_i-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\tau_0^2}]) \\
&= \exp(-\tfrac{1}{2}[A\mu^2 - 2B\mu + C]) \\
&\propto \exp(-\tfrac{1}{2}[A\mu^2 - 2B\mu]) \\
&= \exp(-\tfrac{1}{2}A[\mu^2 - 2\tfrac{B}{A}\mu + \tfrac{B^2}{A^2} - \tfrac{B^2}{A^2}]) \\
&\propto \exp(-\tfrac{1}{2}A[\mu^2 - 2\tfrac{B}{A}\mu + \tfrac{B^2}{A^2}]) \\
&= \exp\left(-\tfrac{1}{2}\frac{(\mu-\frac{B}{A})^2}{(\sqrt{\frac{1}{A}})^2}\right)
\end{aligned}
$$

$$\implies p(\mu|y_1,...,y_n,\sigma^2) = N(\tfrac{B}{A}, \tfrac{1}{A}).$$

Normal distribution: with known variance

- So, we have seen that:

$$\begin{aligned} Y_1, ..., Y_n|\mu, \sigma^2 &\overset{i.i.d.}{\sim} p(y|\mu, \sigma^2) = N(\mu, \sigma^2) \\ \mu &\sim p(\mu) = N(\mu_0, \tau_0^2) \end{aligned}$$

$$\implies p(\mu|y_1, ..., y_n, \sigma^2) = N(\tfrac{B}{A}, \tfrac{1}{A}).$$

  where

  - $\tau_n^2 = \text{Var}(\mu|y_1, ..., y_n, \sigma^2) = \frac{1}{A} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}$
  - $\mu_n = \text{E}(\mu|y_1, ..., y_n, \sigma^2) = \frac{B}{A} = \frac{\frac{\sum_{i=1}^n y_i}{\sigma^2} + \frac{\mu_0}{\tau_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}$

Normal distribution: with known variance

- The inverse of the variance is called the precision and it gives a measure of the spread and uncertainty in the distribution.
  The larger the precision (thus, the smaller the variance), the less uncertainty, less spread there is in the density.

- Under our model:
  - $p(\mu) = N(\mu_0, \tau_0^2) \Longrightarrow$ prior precision $= \tilde{\tau}_0^2 = \frac{1}{\tau_0^2}$
  - $p(y|\mu, \sigma^2) = N(\mu, \sigma^2) \Longrightarrow$ sampling (model) precision $= \tilde{\sigma}^2 = \frac{1}{\sigma^2}$

    Thus, if $y_1, ..., y_n$ is the data, $n\tilde{\sigma}^2 = \frac{n}{\sigma^2}$ is the precision in the data.

  $$\Longrightarrow \text{posterior precision} = \tilde{\tau}_n^2 = \frac{1}{\mathsf{Var}(\mu|y_1,...,y_n,\sigma^2)} = \frac{1}{\tau_n^2} = \frac{1}{\frac{1}{A}} = A$$
  $$= \frac{n}{\sigma^2} + \frac{1}{\tau_0^2} = n\tilde{\sigma}^2 + \tilde{\tau}_0^2$$

  Thus: posterior precision = precision in the data + prior precision or posterior information = information in the data + prior information.

Normal distribution: with known variance

- prior precision $= \tilde{\tau}_0^2 = \frac{1}{\tau_0^2}$
  data precision $= n\tilde{\sigma}^2 = \frac{n}{\sigma^2}$
  $\Longrightarrow$ posterior precision $= \tilde{\tau}_n^2 = n\tilde{\sigma}^2 + \tilde{\tau}_0^2 = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}$
- The posterior mean, μ_n is equal to:

$$\begin{aligned}
\mu_n = \mathsf{E}(\mu|y_1, ..., y_n, \sigma^2) &= \frac{\frac{\sum_{i=1}^n y_i}{\sigma^2} + \frac{\mu_0}{\tau_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}} \\
&= \frac{1}{\tilde{\tau}_n^2} \cdot \frac{n}{\sigma^2} \cdot \frac{\sum_{i=1}^n y_i}{n} + \frac{1}{\tilde{\tau}_n^2} \cdot \frac{1}{\tau_0^2} \cdot \mu_0 \\
&= \frac{n\tilde{\sigma}^2}{\tilde{\tau}_n^2} \cdot \bar{y} + \frac{\tilde{\tau}_0^2}{\tilde{\tau}_n^2} \cdot \mu_0 \\
&= \frac{n\tilde{\sigma}^2}{n\tilde{\sigma}^2 + \tilde{\tau}_0^2} \cdot \text{sample mean} + \frac{\tilde{\tau}_0^2}{n\tilde{\sigma}^2 + \tilde{\tau}_0^2} \cdot \text{prior mean}
\end{aligned}$$

Normal distribution: with known variance

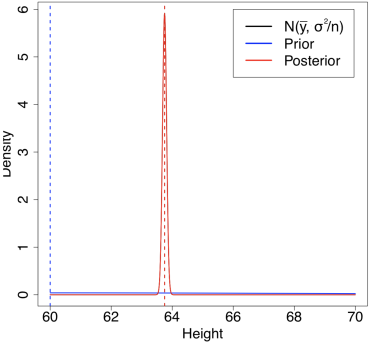
- $\mu_n = \mathsf{E}(\mu|y_1,...,y_n,\sigma^2) = \frac{n\tilde{\sigma}^2}{n\tilde{\sigma}^2+\tilde{\tau}_0^2} \cdot \bar{y} + \frac{\tilde{\tau}_0^2}{n\tilde{\sigma}^2+\tilde{\tau}_0^2} \cdot \mu_0$
- The posterior mean μ_n is a weighted average of the sample mean and the prior mean.
- As n gets large, the weight of the sample mean becomes close to 1 and the weight of the prior mean goes to 0, i.e., the data overwhelms the prior.
- The weight of the prior mean is proportional to the prior precision $\tilde{\tau}_0^2$. The weight of the sample mean is proportional to the data precision $n\tilde{\sigma}^2 = \frac{n}{\sigma^2}$.
- If the prior distribution $p(\mu)$ for the mean parameter, μ, was built based on k_0 observations from the same population, then the prior precision would be: $\tilde{\tau}_0^2 = \frac{k_0}{\sigma^2} = k_0\tilde{\sigma}^2$ and thus:

$$\mu_n = \frac{n\tilde{\sigma}^2}{n\tilde{\sigma}^2 + k_0\tilde{\sigma}^2} \cdot \bar{y} + \frac{k_0\tilde{\sigma}^2}{n\tilde{\sigma}^2 + k_0\tilde{\sigma}^2} \cdot \mu_0 = \frac{n}{n+k_0} \cdot \bar{y} + \frac{k_0}{n+k_0} \cdot \mu_0$$

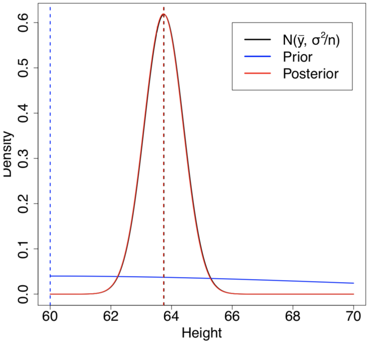
Normal distribution: with known variance

- Example: Consider again the example on the height of women of age over 18. We have collected data on n women. We model the data as follows: conditional on μ and σ^2, $Y_1, ..., Y_n$ are conditionally i.i.d. with $p(y|\mu, \sigma^2) = N(\mu, \sigma^2 = 6.25)$. We place a $p(\mu) = N(\mu_0, \tau_0^2)$ prior on μ, and want to infer upon μ.

Normal distribution: with known variance



$n = 1375$; $\bar{y} = 63.75$; $n\tilde{\sigma}^2 = \frac{n}{\sigma^2} = 220$
$\mu_0 = 60$; $\tau_0^2 = 100$; $\tilde{\tau}_0^2 = 0.01$
$\mu_n = 63.75$; $\tau_n^2 = 0.0045$; $\tilde{\tau}_n^2 = 220.01$



$n = 15$; $\bar{y} = 63.75$; $n\tilde{\sigma}^2 = \frac{n}{\sigma^2} = 2.4$
$\mu_0 = 60$; $\tau_0^2 = 100$; $\tilde{\tau}_0^2 = 0.01$
$\mu_n = 63.73$; $\tau_n^2 = 0.414$; $\tilde{\tau}_n^2 = 2.41$

Normal distribution: with known variance

- Suppose now that after having observed the data for the $n = 1375$ women, you want to predict the height y_{n+1} of a woman age over 18.

- You can predict y_{n+1} using the posterior predictive distribution $p(y_{n+1}|y_1, ..., y_n, \sigma^2)$. This entails computing the integral:

$$\begin{aligned} p(y_{n+1}|y_1, ..., y_n, \sigma^2) &= \int_{-\infty}^{\infty} p(y_{n+1}|\mu, \sigma^2) p(\mu|y_1, ..., y_n, \sigma^2) d\mu \\ &= \int_{-\infty}^{\infty} N(y_{n+1}|\mu, \sigma^2) N(\mu|\mu_n, \tau_n^2) d\mu \\ &\propto \int_{-\infty}^{\infty} \exp(-\tfrac{1}{2} \tfrac{(y_{n+1}-\mu)^2}{\sigma^2}) \exp(-\tfrac{1}{2} \tfrac{(\mu-\mu_n)^2}{\tau_n^2}) d\mu \\ &= \int_{-\infty}^{\infty} \exp(-\tfrac{1}{2} [\tfrac{(y_{n+1}-\mu)^2}{\sigma^2} + \tfrac{(\mu-\mu_n)^2}{\tau_n^2}]) d\mu \end{aligned}$$

Normal distribution: with known variance

- In fact, the posterior predictive distribution

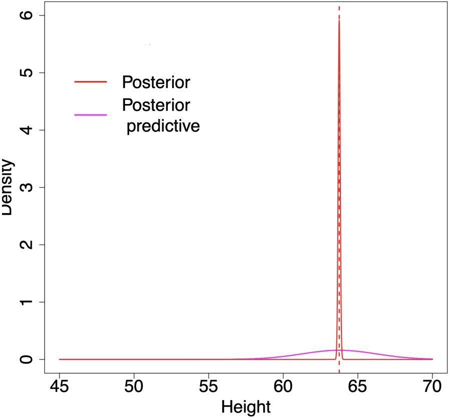
  $$p(y_{n+1}|y_1, ..., y_n, \sigma^2) \propto \int_{-\infty}^{\infty} \exp(-\frac{1}{2}[\frac{(y_{n+1}-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_n)^2}{\tau_n^2}])d\mu$$

  is the normal distribution $N(\mu_n, \tau_n^2 + \sigma^2)$.

- The uncertainty in a new prediction y_{n+1} comes from variability in the population (quantified by the variance σ^2) and uncertainty about μ (quantified by τ_n^2).
  If n is large, τ_n^2 is small and the uncertainty in the new prediction gets close to σ^2.

Normal distribution: with known variance

- Example: Consider again the example on the height of women of age over 18. We have collected data on $n = 1375$ women, for which $\bar{y} = 63.75$ inches and we assume to know $\sigma^2 = 6.25$ inches^2. Using a $p(\mu) = N(\mu_0 = 60, \tau_0^2 = 100)$ prior on μ, we want to predict the height for a new woman over 18.



- The posterior distribution for μ is $N(\mu_n, \tau_n^2) = N(63.75, 0.0045)$.

- Thus, the posterior predictive distribution is $N(\mu_n, \tau_n^2 + \sigma^2) = N(63.75, 6.2545)$.

Normal distribution: with known mean

- The normal density $p(y|\mu, \sigma^2)$ depends on two parameters, the mean μ and the variance σ^2.

- We have seen how to carry out Bayesian inference for the mean μ when we assume that the variance σ^2 is known.

- Now, we will see how to perform Bayesian inference on the variance σ^2 when the mean μ is known.

- Considering still the example on the women height, our data consist of $n = 1375$ observations on women over the age of 18, whose height was recorded.
  We model the observations $y_1, ..., y_n$ as realizations of n random variables $Y_1, ..., Y_n$, conditionally i.i.d., given μ and σ^2, with density $p(y|\mu, \sigma^2) = N(\mu, \sigma^2)$.

- We assume that we know the average height, μ, for the population: $\mu = 63.75$ inches. We want to infer upon σ^2.

Normal distribution: with known mean

- The likelihood in this case is:

$$p(y_1, ..., y_n|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^{n}(y_i - \mu)^2)$$

- If $p(\sigma^2)$ is a prior distribution for σ^2, then the posterior distribution $p(\sigma^2|y_1, ..., y_n, \mu)$ for σ^2 is given by:

$$\begin{aligned} p(\sigma^2|y_1, ..., y_n, \mu) &\propto p(y_1, ..., y_n|\mu, \sigma^2) \cdot p(\sigma^2) \\ &\propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp(-\frac{1}{\sigma^2}(\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2})) \cdot p(\sigma^2) \end{aligned}$$

Normal distribution: with known mean

- Since

$$p(\sigma^2|y_1,...,y_n,\mu) \propto \tfrac{1}{(\sigma^2)^{\frac{n}{2}}} \exp(-\tfrac{1}{\sigma^2}(\tfrac{\sum_{i=1}^n (y_i-\mu)^2}{2})) \cdot p(\sigma^2),$$

  a prior $p(\sigma^2)$ for σ^2 is a conjugate prior if it contains terms of the form $(\frac{1}{\sigma^2})^{c_1} \cdot \exp(-\frac{c_2}{\sigma^2})$.

- A random variable X with sample space $\mathcal{X} = (0,\infty)$ is said to have an Inverse Gamma(a,b) distribution if its density $p(x;a,b)$ is given by:

$$p(x;a,b) = \frac{b^a}{\Gamma(a)} \cdot \frac{1}{x^{a+1}} \cdot \exp(-\frac{b}{x}), \qquad x > 0$$

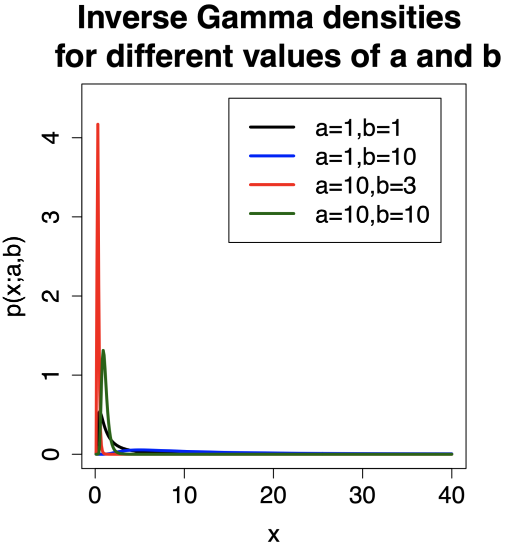
  with $a, b > 0$

- For such a random variable: Mode$(X) = \frac{b}{a+1}$;

$$\mathrm{E}(X) = \tfrac{b}{a-1} \text{ for } a > 1; \qquad \mathrm{Var}(X) = \tfrac{b^2}{(a-1)^2(a-2)} \text{ for } a > 2.$$

Normal distribution: with known mean

The Inverse Gamma(a, b) density $p(x; a, b)$ for different values of a and b.



Normal distribution: with known mean

- The Inverse Gamma distribution arises from a Gamma distribution: If X is a random variable with a Gamma(a, b) distribution, and $\widetilde{X} = \frac{1}{X}$, then $\widetilde{X}$ has an Inverse Gamma(a, b) distribution.

- Remember that the inverse $\tilde{\sigma}^2 = \frac{1}{\sigma^2}$ of the variance σ^2 is called precision. Specifying an Inverse Gamma(a, b) prior for the variance σ^2 is equivalent to specifying a Gamma(a, b) prior for the precision $\tilde{\sigma}^2$.

- If $p(\sigma^2) = $ InverseGamma(a, b), then:

$$\begin{aligned}
p(\sigma^2|y_1, ..., y_n, \mu) &\propto p(y_1, ..., y_n|\mu, \sigma^2) \cdot p(\sigma^2) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp(-\frac{1}{\sigma^2}(\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2})) \\
&\quad \cdot \frac{b^a}{\Gamma(a)} \frac{1}{(\sigma^2)^{a+1}} \exp(-\frac{b}{\sigma^2}) \\
&\propto \frac{1}{(\sigma^2)^{\frac{n}{2}+a+1}} \exp(-\frac{1}{\sigma^2}(b + \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2}))
\end{aligned}$$

Normal distribution: with known mean

- Therefore:

$$\begin{aligned} Y_1, ..., Y_n|\mu, \sigma^2 &\overset{i.i.d.}{\sim} p(y|\mu, \sigma^2) = N(\mu, \sigma^2) \\ \sigma^2 &\sim p(\sigma^2) = \text{Inverse Gamma}(a, b) \end{aligned}$$

$$\Longrightarrow p(\sigma^2|y_1, \ldots, y_n, \mu) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}+a+1}} \exp(-\frac{1}{\sigma^2}(b + \frac{\sum_{i=1}^n (y_i-\mu)^2}{2}))$$

that is, $p(\sigma^2|y_1, \ldots, y_n, \mu) = \text{Inverse Gamma}(\tilde{a}, \tilde{b})$ where:

$$\tilde{a} = a + \frac{n}{2}, \qquad \tilde{b} = b + \frac{\sum_{i=1}^n (y_i - \mu)^2}{2}$$

Normal distribution: with known mean

- Example: Consider again the example on women's height. We have observations on $n = 1375$ women. We assume that, conditional on the mean μ and on the variance σ^2, the variables $Y_1, ..., Y_n$ are conditionally i.i.d. with density $p(y|\mu, \sigma^2) = N(\mu, \sigma^2)$.
  Assume that we know $\mu = 63.75$ inches.
  If we take $p(\sigma^2) = \text{Inverse Gamma}(a = 2.5, b = 13.5)$ and the observations are such that $\sum_{i=1}^{n}(y_i - \mu)^2 = 8684.9$, then:

$$\tilde{a} = a + \frac{n}{2} = 2.5 + \frac{1375}{2} = 690$$

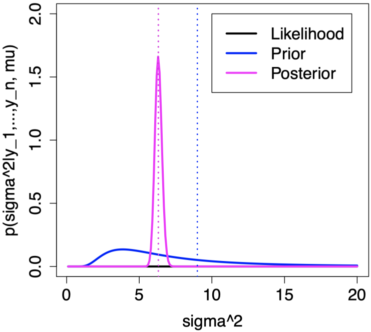
$$\tilde{b} = b + \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2} = 13.5 + \frac{8684.9}{2} = 4355.95$$

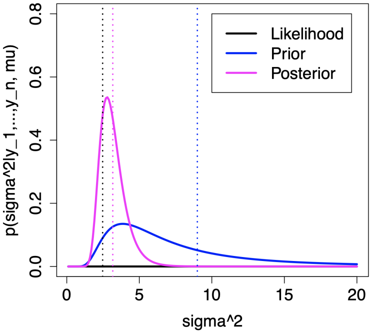
$$\Longrightarrow p(\sigma^2|y_1, ..., y_n, \mu) = \text{Inverse Gamma}(690, 4355.95)$$

Thus: $\text{E}(\sigma^2|y_1, ..., y_n, \mu) = \frac{4355.95}{690-1} = 6.32$

and $\text{Mode}(\sigma^2|y_1, ..., y_n, \mu) = \frac{4355.95}{690+1} = 6.30.$

Normal distribution: with known mean





$n = 1375$; $\mu = 63.75$

$\sum_{i=1}^{n}(y_i - \mu)^2 = 8684.9$

$\sigma^2_{\text{MLE}} = 6.32$

$a = 2.5$; $b = 13.5$

$\tilde{a} = 690$; $\tilde{b} = 4355.95$

$E(\sigma^2 | y_1, ..., y_n, \mu) = 6.32$

$n = 25$; $\mu = 63.75$

$\sum_{i=1}^{n}(y_i - \mu)^2 = 61.83$

$\sigma^2_{\text{MLE}} = 2.47$

$a = 2.5$; $b = 13.5$

$\tilde{a} = 15$; $\tilde{b} = 44.42$

$E(\sigma^2 | y_1, ..., y_n, \mu) = 3.17$

Normal distribution: with known mean

- In the previous example, we have used a prior for σ^2 that was proper but had a large variance.

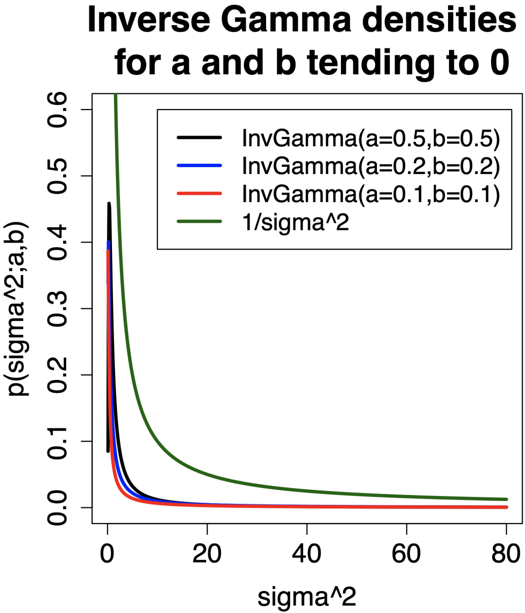
- How can we determine an improper prior for σ^2 that still leads to a proper posterior distribution?

- We could take as prior $p(\sigma^2)$ on σ^2 an Inverse Gamma(a, b) distribution with $a \to 0$ and $b \to 0$. Then:

$$p(\sigma^2) = \frac{b^a}{\Gamma(a)} \cdot \frac{1}{(\sigma^2)^{a+1}} \cdot \exp(-\frac{b}{\sigma^2}) \to \frac{1}{\sigma^2}$$

  but the posterior for σ^2 is proper since it is $p(\sigma^2|y_1, ..., y_n, \mu) =$ Inverse Gamma$(a + \frac{n}{2}, b + \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2})$.

Normal distribution: with known mean



Normal distribution: with known mean

- We have seen how to perform Bayesian inference when the sampling model is a normal density where either the mean, μ, or the variance, σ^2, are known.

- How can we perform Bayesian inference if both are unknown?

- More generally: we have seen how to perform Bayesian inference for particular class of models that depend only on one parameter. How can we perform Bayesian inference if there are multiple parameters?

Multiparameter models

- Most problems in statistics involve inferring upon more than just one parameter. However, even if a problem can include several parameters, the conclusions will often be drawn about one or few parameters at a time.

- In a Bayesian setting, our interest is in the marginal posterior distribution of one or few parameters of interest.

- To derive this distribution, we first need to derive the joint posterior distribution of **all** parameters and then integrate out all the parameters that are not of interest to obtain the marginal posterior distribution.

- If we use Monte Carlo methods (the next class), this means that we will sample from the joint posterior distribution and then we will look only at samples for those parameters of interest.

- The parameters that are needed to derive the joint posterior distribution but are not of interest are called nuisance parameters.

Nuisance parameters

- Suppose we are in the usual context of random variables that are conditionally independent and identically distributed.
- Suppose that the sampling model we are adopting for the data is: $y \sim p(y|\theta)$ where θ is a vector that can be divided into two parts: $\theta = (\theta_1, \theta_2)$.
- We are only interested in θ_1, so θ_2 is a nuisance parameter.
- We want to derive the marginal posterior distribution $p(\theta_1|\text{data})$. We will derive this from the joint posterior distribution $p(\theta_1, \theta_2|\text{data})$.
- From Bayes' theorem:

$$\begin{aligned} p(\theta_1, \theta_2|\text{data}) &\propto p(\text{data}|\theta_1, \theta_2) \cdot p(\theta_1, \theta_2) \\ &= \left[\prod_{i=1}^{n} p(y_i|\theta_1, \theta_2)\right] \cdot p(\theta_1, \theta_2) \end{aligned}$$

Nuisance parameters

- We know from Bayes' theorem $p(\theta_1, \theta_2|\text{data})$. We are interested in deriving $p(\theta_1|\text{data})$.

$$p(\theta_1|\text{data}) = \int p(\theta_1, \theta_2|\text{data}) d\theta_2 = \int p(\theta_1|\theta_2, \text{data}) p(\theta_2|\text{data}) d\theta_2$$

- Usually the integral above is not computed explicitly. We will use Monte Carlo methods (the next class) and proceed as follows:

  Repeat the following steps B times:

  1. Sample $\theta_2^{(i)}$ from the marginal posterior distribution $p(\theta_2|\text{data})$
  2. Sample $\theta_1^{(i)}$ from the conditional posterior distribution $p(\theta_1|\theta_2^{(i)}, \text{data})$

Normal distribution: unknown (μ, σ^2): noninformative prior

- We will start our analysis of multiparameter models with the case of normal data, that is: $y_1, ..., y_n|\mu, \sigma^2 \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$.
- We will assume that the two parameters are independent: $p(\mu, \sigma^2) = p(\mu) \cdot p(\sigma^2)$.
- We will start considering a noninformative prior on (μ, σ^2).
  - A noninformative improper prior on σ^2: $p(\sigma^2) \propto \frac{1}{\sigma^2}$ (Jeffreys', here)
  - A noninformative improper prior on μ: $p(\mu) \propto 1$.
- The prior that we consider is then: $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$.
- Therefore the joint posterior distribution is:

$$
\begin{aligned}
p(\mu, \sigma^2|y_1, \ldots, y_n) &\propto [\textstyle\prod_{i=1}^n \{\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(y_i - \mu)^2)\}] \cdot \frac{1}{\sigma^2} \\
&\propto \frac{1}{\sigma^{n+2}} \exp(-\frac{1}{2\sigma^2}[\textstyle\sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2]) \\
&= \frac{1}{\sigma^{n+2}} \exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2])
\end{aligned}
$$

unknown (μ, σ^2): noninformative prior

- We factor the joint posterior distribution $p(\mu, \sigma^2)$ of μ, σ^2 in the product of a conditional posterior distribution and a marginal posterior distribution:

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

- We have seen previously what is the (conditional) posterior distribution (page 11) $p(\mu | y_1, ..., y_n, \sigma^2)$ under a different prior on μ:

$$\begin{cases} Y_1, ..., Y_n | \mu, \sigma^2 & \overset{i.i.d.}{\sim} & N(\mu, \sigma^2) \\ \mu & \sim & N(\mu_0, \tau_0^2) \end{cases}$$

$$\Longrightarrow p(\mu | y_1, ..., y_n, \sigma^2) = N(\mu_n, \tau_n^2), \text{ where}$$

$$\begin{cases} \tau_n^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}} \\ \mu_n = \frac{n\tau_n^2}{\sigma^2} \cdot \bar{y} + \frac{\tau_n^2}{\tau_0^2} \cdot \mu_0 \end{cases}$$

unknown (μ, σ^2): noninformative prior

- The conditional posterior distribution $p(\mu|y_1, ..., y_n, \sigma^2) = N(\mu_n, \tau_n^2)$ has been derived under the prior $N(\mu_0, \tau_0^2)$

- Now we are using $p(\mu) \propto 1$ as prior on μ. This can be considered as the limit of a $N(\mu_0, \tau_0^2)$ for $\tau_0^2 \to \infty$.

- This implies that the (conditional) posterior distribution $p(\mu|y_1, ..., y_n, \sigma^2)$ is $N(\mu_n, \tau_n^2)$ where now:

$$\begin{cases} \tau_n^2 = \frac{\sigma^2}{n} \\ \mu_n = \bar{y} \end{cases}$$

- So under an improper noninformative prior on μ: $p(\mu|y_1, ..., y_n, \sigma^2) = N(\bar{y}, \frac{\sigma^2}{n})$.

unknown (μ, σ^2): noninformative prior

- To derive the marginal posterior distribution $p(\sigma^2|y_1, ..., y_n)$, we need to integrate μ out of the joint posterior distribution $p(\mu, \sigma^2|y_1, ..., y_n)$:

$$\begin{aligned} p(\sigma^2|y_1, ..., y_n) &= \int_{-\infty}^{\infty} p(\mu, \sigma^2|y_1, ..., y_n) d\mu \\ &\propto \int_{-\infty}^{\infty} \frac{1}{\sigma^{n+2}} \exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]) d\mu \\ &= \frac{1}{\sigma^{n+2}} \exp(-\frac{1}{2\sigma^2}(n-1)s^2) \sqrt{\frac{2\pi\sigma^2}{n}} \\ &\propto \frac{1}{(\sigma^2)^{\frac{n+1}{2}}} \exp(-\frac{(n-1)s^2}{2\sigma^2}) \end{aligned}$$

This means that $p(\sigma^2|y_1, ..., y_n) = \text{InverseGamma}(\frac{n-1}{2}, \frac{(n-1)s^2}{2})$.

unknown (μ, σ^2): noninformative prior

- A positive random variable X is said to have a Scaled Inverse $\chi^2(\nu, \kappa)$ with ν degrees of freedom distribution if its density is given by:

  $$p(x; \nu, \kappa) = \frac{\frac{\kappa\nu}{2}^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \cdot \frac{1}{x^{1+\frac{\nu}{2}}} \cdot \exp(-\frac{\nu\kappa}{2x})$$

  where $\nu, \kappa > 0$. κ is called the scale parameter.

  For such a random variable: Mode$(X) = \frac{\nu}{\nu+2}\kappa$

  $$\mathsf{E}(X) = \frac{\nu}{\nu-2}\kappa \text{ for } \nu > 2; \quad \mathsf{Var}(X) = \frac{2\nu^2}{(\nu-2)^2(\nu-4)}\kappa^2 \text{ for } \nu > 4.$$

- Since

  $$p(\sigma^2|y_1, ..., y_n) \propto \frac{1}{(\sigma^2)^{\frac{n+1}{2}}} \exp(-\frac{(n-1)s^2}{2\sigma^2}),$$

  we have $p(\sigma^2|y_1, ..., y_n) =$ Scaled Inverse $\chi^2(n-1, s^2)$.

unknown (μ, σ^2): noninformative prior

- Note that the similarity between this result and the result in classical (frequentist) statistics: conditional on σ^2 (and μ)

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

- The decomposition of the joint posterior distribution $p(\mu, \sigma^2|y_1, ..., y_n)$ into the product of the conditional posterior distribution $p(\mu|\sigma^2, y_1, ..., y_n) = N(\bar{y}, \frac{\sigma^2}{n})$ and of the marginal posterior distribution $p(\sigma^2|y_1, ..., y_n) =$ Scaled Inverse $\chi^2(n-1, s^2)$ can be used to obtain samples from the joint posterior distribution:

  Repeat the following B times:

  1. sample ${\sigma^2}^{(i)}$ from $p(\sigma^2|y_1, ..., y_n) =$ Scaled Inverse $\chi^2(n-1, s^2)$;
  2. sample $\mu^{(i)}$ from $p(\mu|{\sigma^2}^{(i)}, y_1, ..., y_n) = N(\bar{y}, \frac{{\sigma^2}^{(i)}}{n})$.

unknown (μ, σ^2): noninformative prior

- Then: $(\mu^{(1)}, \sigma^{2^{(1)}}), ..., (\mu^{(B)}, \sigma^{2^{(B)}})$ are B independent samples from the joint posterior distribution $p(\mu, \sigma^2 | y_1, ..., y_n)$

- If we look only at samples $(\mu^{(1)}, ..., \mu^{(B)})$: these are B independent samples from the marginal posterior distribution $p(\mu | y_1, ..., y_n)$

unknown (μ, σ^2): noninformative prior: an example

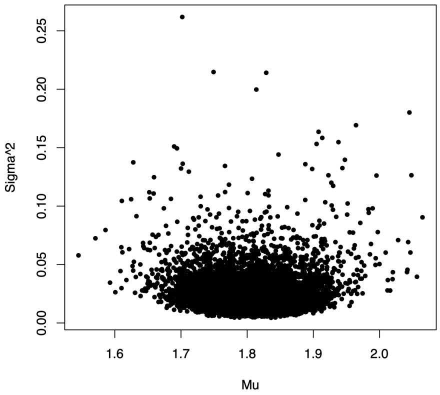
- Data $y_1, ..., y_n$ on the wing length in millimeters of nine members of a species of midge (small, two-winged flies) can be modeled, conditional on the mean μ and variance σ^2, as realizations of n random variables that are conditionally i.i.d. $N(\mu, \sigma^2)$. The data collected (ordered in increasing magnitude) are:
  $\boldsymbol{y} = (1.64, 1.70, 1.72, 1.74, 1.82, 1.82, 1.82, 1.90, 2.08)$, which gives:

$$\bar{y} = 1.804 \qquad \text{and} \qquad s^2 = 0.017$$

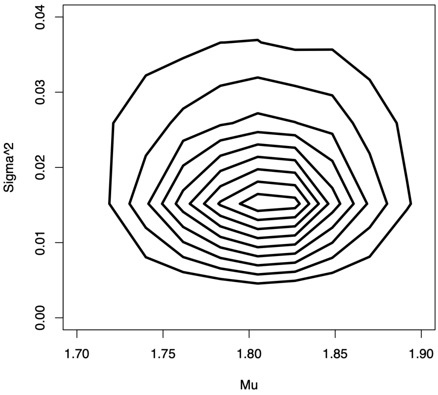
unknown (μ, σ^2): noninformative prior: an example

- We sampled $B = 10,000$ from the joint posterior distribution $p(\mu, \sigma^2|y_1, ..., y_n)$.
- Monte Carlo estimate of the joint posterior distribution $p(\mu, \sigma^2|y_1, ..., y_n)$.

Samples from the joint posterior distribution

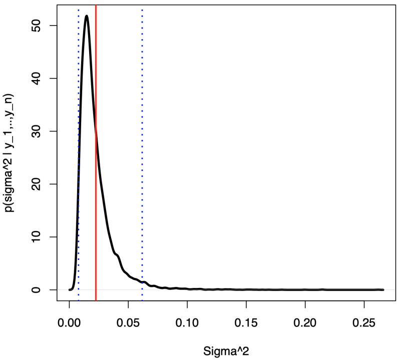


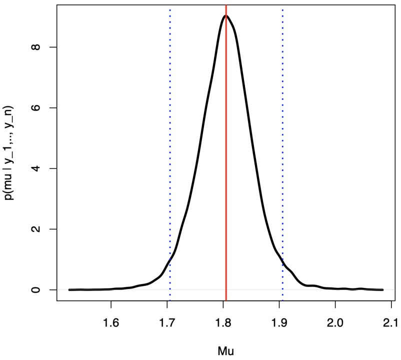
Monte Carlo estimate of the joint posterior density



unknown (μ, σ^2): noninformative prior: an example

- Marginal posterior distributions $p(\sigma^2|y_1, ..., y_n)$ and $p(\mu|y_1, ..., y_n)$.
- Displayed are also the Monte Carlo estimate of the posterior mean and the 95% Credible interval for σ^2 and μ, respectively.
- Theoretical posterior mean and variance for σ^2: 0.023 and 0.0002, respectively.
- Monte Carlo estimate of posterior mean and variance for σ^2: 0.022 and 0.0002, respectively.





unknown (μ, σ^2): noninformative prior

- Often, when we work with normal data, the parameter of interest is the mean, μ: we might be interested in deriving the marginal posterior distribution $p(\mu|y_1, ..., y_n)$ of μ.
- We can approximate this via the Monte Carlo method as shown in the example before or we can compute it in closed form from the joint posterior distribution $p(\mu, \sigma^2|y_1, ..., y_n)$:

$$\begin{aligned} p(\mu|y_1, ..., y_n) &= \int_0^\infty p(\mu, \sigma^2|y_1, ..., y_n) d\sigma^2 \\ &\propto \int_0^\infty \tfrac{1}{\sigma^{n+2}} \exp(-\tfrac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]) d\sigma^2 \\ &\propto \int_0^\infty A^{-\frac{n}{2}} z^{\frac{n-2}{2}} \exp(-z) dz \\ &\text{(where } A = (n-1)s^2 + n(\bar{y}-\mu)^2 \text{ and } z = \tfrac{A}{2\sigma^2}) \\ &\propto [(n-1)s^2 + n(\bar{y}-\mu)^2]^{-\frac{n}{2}} \\ &\propto [1 + \tfrac{n(\mu-\bar{y})^2}{(n-1)s^2}]^{-\frac{n}{2}} \end{aligned}$$

unknown (μ, σ^2): noninformative prior

- A random variable X is said to have a scaled non-central $t_\nu(\eta, \kappa)$ distribution with ν degrees of freedom if its density is given by:

  $$p(x; \nu, \eta, \kappa) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}\kappa}\Big(1 + \frac{1}{\nu}(\frac{x-\eta}{\kappa})^2\Big)^{-\frac{\nu+1}{2}}$$

  where $\nu, \kappa > 0$. The parameter η is called the location parameter and κ is called the scale parameter.For such a random variable:

  $$\text{E}(X) = \eta, \nu > 1; \quad \text{Var}(X) = \frac{\nu}{\nu-2}\kappa^2, \nu > 2; \quad \text{Mode}(X) = \eta$$

- Since

  $$p(\mu|y_1, ..., y_n) \propto [1 + \tfrac{n(\mu-\bar{y})^2}{(n-1)s^2}]^{-\frac{n}{2}},$$

  we have

  $$p(\mu|y_1, ..., y_n) = t_{n-1}(\bar{y}, \frac{s}{\sqrt{n}}).$$

unknown (μ, σ^2): noninformative prior

- Note that there is the following relationship between the scaled non-central $t_\nu(\eta, \kappa)$ distribution and the t_ν distribution:

$$\text{if } X \sim t_\nu(\eta, \kappa), \quad \text{then } \frac{X - \eta}{\kappa} \sim t_\nu$$

- Since we have just shown that $\mu | y_1, ..., y_n \sim t_{n-1}(\bar{y}, \frac{s}{\sqrt{n}})$ under $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$, then

$$\frac{\mu - \bar{y}}{\frac{s}{\sqrt{n}}} | y_1, ..., y_n \sim t_{n-1}$$

  Again, note the similarity with the frequentist result: if $y_1, ..., y_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables, then:

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

unknown (μ, σ^2): noninformative prior

- The difference between the two statements is:
  In classical, frequentist statistics

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

before we observe the data, our uncertainty about the scaled deviation of $\bar{y}$ from the mean μ can be represented by a t_{n-1} distribution (here both $\bar{y}$ and μ are unknown).

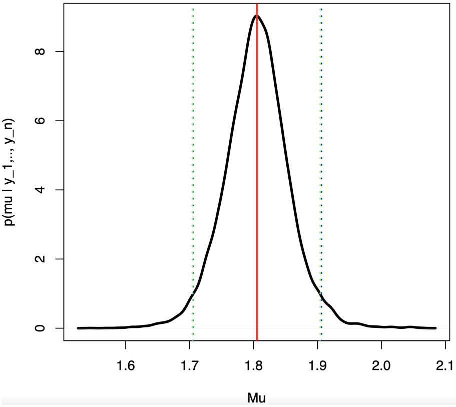
In Bayesian statistics

$$\frac{\mu - \bar{y}}{\frac{s}{\sqrt{n}}} | y_1, ..., y_n \sim t_{n-1}$$

after we observe the data, our uncertainty about the scaled deviation of $\bar{y}$ from the mean μ can be represented by a t_{n-1} distribution (here only μ is unknown).

unknown (μ, σ^2): noninformative prior: an example

- Using the same B independent samples of μ values that we have considered before, below is a plot of the marginal posterior distribution of μ with the Monte Carlo estimate (in blue) and the theoretical (in green) 95% credible interval.



unknown (μ, σ^2): noninformative prior: posterior prediction

- The posterior predictive distribution for a new observation $\tilde{y}$ can be obtained using the usual representation:

$$
\begin{aligned}
&p(\tilde{y}|y_1, ..., y_n) \\
&= \int_{-\infty}^{\infty} \int_0^{\infty} p(\tilde{y}|\mu, \sigma^2) p(\mu, \sigma^2 | y_1, ..., y_n) d\mu d\sigma^2 \\
&= \int_{-\infty}^{\infty} \int_0^{\infty} p(\tilde{y}|\mu, \sigma^2) p(\mu|\sigma^2, y_1, ..., y_n) p(\sigma^2 | y_1, ..., y_n) d\mu d\sigma^2 \\
&= \int_{-\infty}^{\infty} \int_0^{\infty} N(\tilde{y}; \mu, \sigma^2) \cdot N(\mu; \bar{y}, \tfrac{\sigma^2}{n}) \\
&\qquad\qquad \cdot \text{Scaled Inverse } \chi^2(n-1, s^2) d\mu d\sigma^2
\end{aligned}
$$

- We can sample from the posterior predictive distribution using Monte Carlo methods. How?

- It can be shown (using the same techniques used to derive $p(\mu|y_1, ..., y_n)$) that $p(\tilde{y}|y_1, ..., y_n)$ is a scaled t distribution with $n-1$ degrees of freedom, location $\bar{y}$ and scale $\sqrt{1 + \frac{1}{n}} s$

With unknown μ and σ^2: conjugate prior

- We have looked at the following case:

$$\begin{cases} Y_1, ..., Y_n | \mu, \sigma^2 & \overset{i.i.d.}{\sim} & p(y|\mu, \sigma^2) = N(\mu, \sigma^2) \\ \mu, \sigma^2 & \sim & p(\mu, \sigma^2) \propto \frac{1}{\sigma^2} \end{cases}$$

$$\Longrightarrow \begin{cases} p(\mu | y_1, ..., y_n, \sigma^2) & = & N(\bar{y}, \frac{\sigma^2}{n}) \\ p(\sigma^2 | y_1, ..., y_n) & = & \text{Inverse Gamma}(\frac{n-1}{2}, \frac{(n-1)s^2}{2}) \\ & = & \text{(Scaled) Inverse } \chi^2(n-1, s^2) \end{cases}$$

- Now, we consider a more general model where we place a conjugate prior $p(\mu, \sigma^2)$ on μ, σ^2.

With unknown μ and σ^2: conjugate prior

- To determine a conjugate prior $p(\mu, \sigma^2)$ on μ, σ^2, we look at the likelihood $p(y_1, ..., y_n|\mu, \sigma^2)$:

$$\begin{aligned} p(y_1, ..., y_n|\mu, \sigma^2) &= \textstyle\prod_{i=1}^n \left\{\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2})\right\} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}}\frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}[\textstyle\sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2]\right) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}}\frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2]\right) \end{aligned}$$

- Remembering the form of the Inverse Gamma density and of the Normal density, it looks that a conjugate prior distribution $p(\mu, \sigma^2)$ on μ, σ^2 could be of the form:

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

  with $p(\sigma^2)$ Inverse Gamma (or Scaled Inverse χ^2) density and $p(\mu|\sigma^2)$ Normal density.

With unknown μ and σ^2: conjugate prior

- We have chosen a prior $p(\mu, \sigma^2) = p(\sigma^2) \cdot p(\mu|\sigma^2)$ where

$$p(\mu|\sigma^2) = N(\mu_0, \tfrac{\sigma^2}{\kappa_0}) \qquad \text{(page 8)}$$

$$p(\sigma^2) = \text{InverseGamma}(\tfrac{\nu_0}{2}, \tfrac{\nu_0\sigma_0^2}{2}) = \text{Inverse } \chi^2(\nu_0, \sigma_0^2) \qquad \text{(page 22)}$$

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

- Similarly, the parameters σ_0^2 and ν_0 can be interpreted as:
  - σ_0^2: sample variance of prior observations.
  - ν_0: sample size of prior observations (or prior degrees of freedom)
  - $\nu_0\sigma_0^2$: sum of squares of prior observations, also called prior sum of squares.

With unknown μ and σ^2: conjugate prior

- With these choices, the joint prior density $p(\mu, \sigma^2)$ is given by:

$$\begin{aligned} p(\mu, \sigma^2) &= p(\sigma^2) p(\mu|\sigma^2) \\ &\propto \frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}} \exp(-\frac{\nu_0 \sigma_0^2}{2\sigma^2}) \cdot \frac{1}{\sqrt{2\pi\frac{\sigma^2}{\kappa_0}}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\frac{\sigma^2}{\kappa_0}}\right) \\ &\propto \frac{1}{\sigma} \cdot \frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}} \exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu-\mu_0)^2]\right) \end{aligned}$$

- This joint prior density is sometimes denoted as the Normal-Inverse $\chi^2(\mu_0; \frac{\sigma_0^2}{\kappa_0}; \nu_0; \sigma_0^2)$ density.

- Note that μ and σ^2 are dependent in their joint conjugate prior distribution: if σ^2 is large, then the prior distribution on μ is a high-variance distribution.

With unknown μ and σ^2: conjugate prior

- Multiplying the joint prior distribution times the likelihood, we obtain the joint posterior distribution:

$$\begin{aligned}
p(\mu,\sigma^2|y_1,...,y_n) &\propto p(y_1,...,y_n|\mu,\sigma^2)\cdot p(\mu,\sigma^2) \\
&\propto \frac{1}{(2\pi)^{\frac{n}{2}}}\frac{1}{(\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2+n(\bar{y}-\mu)^2]\right) \\
&\quad\cdot\frac{1}{\sigma}\cdot\frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}}\exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2+\kappa_0(\mu-\mu_0)^2]\right) \\
&\propto \frac{1}{\sigma}\cdot\frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+\frac{n}{2}+1}}\exp(-\frac{1}{2\sigma^2}[(n-1)s^2+\nu_0\sigma_0^2 \\
&\quad+n(\mu-\bar{y})^2+\kappa_0(\mu-\mu_0)^2])
\end{aligned}$$

With unknown μ and σ^2: conjugate prior

- After some algebra, we can see that:

$$p(\mu, \sigma^2 | y_1, ..., y_n) = \text{Normal-Inverse } \chi^2(\mu_n; \frac{\sigma_n^2}{\kappa_n}; \nu_n; \sigma_n^2)$$

where

$$\begin{aligned}
\mu_n &= \tfrac{\kappa_0}{\kappa_0+n} \cdot \mu_0 + \tfrac{n}{\kappa_0+n} \cdot \bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\nu_n \sigma_n^2 &= \nu_0 \sigma_0^2 + (n-1)s^2 + \tfrac{\kappa_0 n}{\kappa_0+n}(\bar{y} - \mu_0)^2
\end{aligned}$$

This means that: $p(\mu | y_1, ..., y_n, \sigma^2) = N(\mu_n, \frac{\sigma^2}{\kappa_n})$ and $p(\sigma^2 | y_1, ..., y_n) =$ Inverse $\chi^2(\nu_n, \sigma_n^2) =$ InverseGamma$(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2})$

With unknown μ and σ^2: conjugate prior

- $p(\mu|y_1, ..., y_n, \sigma^2) = N(\mu_n, \frac{\sigma^2}{\kappa_n})$ where

$$\mu_n = \tfrac{\kappa_0}{\kappa_0+n} \cdot \mu_0 + \tfrac{n}{\kappa_0+n} \cdot \bar{y}, \qquad \kappa_n = \kappa_0 + n$$

The posterior mean μ_n is a weighted average of the prior mean and the sample mean with weights that depend on the relative precision of the two pieces of information (cf. page 14).

The posterior precision (i.e., the inverse of the posterior variance) is:

$$\frac{\kappa_0 + n}{\sigma^2} = \frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2} = \kappa\tilde{\sigma}^2 + n\tilde{\sigma}^2$$

which is the form of the posterior precision shown on page 12.

With unknown μ and σ^2: conjugate prior

- $p(\sigma^2|y_1, ..., y_n) =$ Inverse $\chi^2(\nu_n, \sigma_n^2) =$ InverseGamma$(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2})$ where

  $$\nu_n = \nu_0 + n \qquad \nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + (n-1)s^2 + \tfrac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2$$

  The posterior degrees of freedom, ν_n, is equal to the prior degrees of freedom plus the sample size.

  The posterior sum of squares, $\nu_n \sigma_n^2$, is a combination of the prior sum of squares, the sample sum of squares and the squared difference between the sample mean and the prior mean.

With unknown μ and σ^2: conjugate prior

- The marginal posterior distribution of σ^2 is:

$$p(\sigma^2|y_1,...,y_n) = \text{InverseGamma}(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2}) = \text{Inverse } \chi^2(\nu_n, \sigma_n^2)$$

- If we want to obtain the marginal posterior distribution of μ, we need to integrate out σ^2 from the joint posterior distribution:

$$\begin{aligned} p(\mu|y_1,...,y_n) &= \int_0^\infty p(\mu,\sigma^2|y_1,...,y_n)d\sigma^2 \\ &= \int_0^\infty \text{Normal-Inverse } \chi^2(\mu_n; \tfrac{\sigma_n^2}{\kappa_n}; \nu_n; \sigma_n^2)d\sigma^2 \end{aligned}$$

Using the same technique used on pages 44-45, we can show that

$$p(\mu|y_1,...,y_n) \propto \left[1 + \tfrac{\kappa_n(\mu-\mu_n)^2}{\nu_n\sigma_n^2}\right]^{-\frac{\nu_n+1}{2}} = t_{\nu_n}(\mu_n, \tfrac{\sigma_n}{\sqrt{\kappa_n}})$$

that is, the marginal posterior distribution of μ is a scaled non-central t distribution with ν_n degrees of freedom, location parameter μ_n and scale parameter $\frac{\sigma_n}{\sqrt{\kappa_n}}$.

With unknown μ and σ^2: conjugate prior

- In practice, we will derive the marginal posterior distribution of μ using a Monte Carlo approach.
- We will generate B samples from the joint posterior distribution $p(\mu, \sigma^2 | y_1, ..., y_n)$ by doing the following:

  For $i = 1, ..., B$ repeat the following:

  1. Sample $\sigma^{2^{(i)}}$ from $p(\sigma^2 | y_1, ..., y_n) =$ InverseGamma$(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2})$.
  2. Sample $\mu^{(i)}$ from $p(\mu | y_1, ..., y_n, \sigma^{2^{(i)}}) = N(\mu_n, \frac{\sigma^{2^{(i)}}}{\kappa_n})$.
- The samples $\{(\mu^{(1)}, \sigma^{2^{(1)}}), ..., (\mu^{(B)}, \sigma^{2^{(B)}})\}$ are B independent samples from the joint posterior distribution $p(\mu, \sigma^2 | y_1, ..., y_n)$.
- $(\sigma^{2^{(1)}}, ..., \sigma^{2^{(B)}})$ are B independent samples from the marginal posterior distribution $p(\sigma^2 | y_1, ..., y_n)$.
- $(\mu^{(1)}, ..., \mu^{(B)})$ are B independent samples from the marginal posterior distribution $p(\mu | y_1, ..., y_n)$.

Example: Problem 5.1

- Problem 5.1 (p. 235): The files `school1.dat`, `school2.dat` and `school3.dat` contain data on the amount of time students from three high schools spent on studying or homework during an exam period.

  Analyze data from each of these schools separately, using the normal model with a conjugate prior distribution in which $\{\mu_0 = 5, \sigma_0^2 = 4, \kappa_0 = 1, \nu_0 = 2\}$ and compute or approximate:

  1. the posterior means and 95% credible intervals for the mean θ and standard deviation σ from each school;
  2. the posterior probability that $\theta_i < \theta_j < \theta_k$ for all six permutations $\{i, j, k\}$ of $\{1, 2, 3\}$;
  3. the posterior probability that $\tilde{Y}_i < \tilde{Y}_j < \tilde{Y}_k$ for all six permutations $\{i, j, k\}$ of $\{1, 2, 3\}$, where $\tilde{Y}_i$ is a sample from the posterior predictive distribution of school i.
  4. the posterior probability that θ_1 is bigger than both θ_2 and θ_3, and the posterior probability that $\tilde{Y}_1$ is bigger than both $\tilde{Y}_2$ and $\tilde{Y}_3$.

Example: Problem 5.1

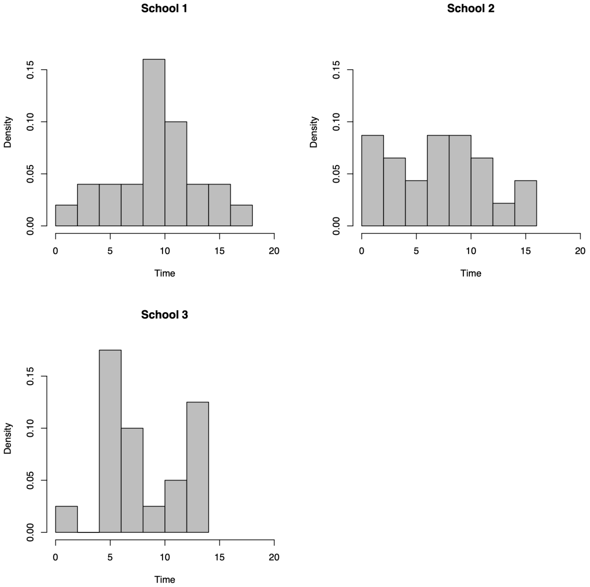
- For each school, we use the same prior $p(\theta_i, \sigma_i^2)$. Thus, our model is: for $i = 1, 2, 3$

$$Y_{1,i}, \ldots, Y_{n_i,i} | \theta_i, \sigma_i^2 \quad \overset{iid}{\sim} \quad p(y|\theta_i, \sigma_i^2) = N(\theta_i, \sigma_i^2)$$

$$\begin{aligned} \theta_i, \sigma_i^2 \quad \sim \quad & p(\theta_i, \sigma_i^2) = p(\theta_i|\sigma_i^2) \cdot p(\sigma_i^2) \\ & = N(\theta_i; \mu_0 = 5, \tfrac{\sigma_i^2}{\kappa_0} = \sigma_i^2) \\ & \quad \cdot \text{InverseGamma}(\tfrac{\nu_0}{2} = \tfrac{2}{2} = 1, \tfrac{\nu_0 \sigma_0^2}{2} = \tfrac{2 \times 4}{2} = 4) \end{aligned}$$

Example: Problem 5.1

- Histogram of the data at the three schools



Example: Problem 5.1

- Sample mean and sample variance for the three datasets:

School	n	$\bar{y}$	s^2
1	25	9.46	15.10
2	23	7.03	20.11
3	20	7.95	14.30

- Parameters for the posterior distribution:

School	μ_n	κ_n	ν_n	σ_n^2
1	9.29	26	27	14.42
2	6.95	24	25	18.18
3	7.81	21	22	13.09

Example: Problem 5.1

- We can compute the posterior mean and a 95% credible interval for θ_i for $i = 1, 2, 3$ from the marginal posterior distribution of θ_i,which is $t_{\nu_n}(\mu_n, \frac{\sigma_n}{\sqrt{\kappa_n}})$:

School	Posterior mean	95% credible interval
1	9.29	(7.76,10.82)
2	6.95	(5.16,8.74)
3	7.81	(6.17,9.45)

Example: Problem 5.1

- To get the posterior mean and a 95% credible interval for σ_i, we will use a Monte Carlo approximation.
- If we want to use the exact distribution to determine the posterior mean and a 95% credible interval (l_i, u_i), we need to do the following:
  1. Derive the marginal posterior distribution of σ_i knowing that the marginal posterior distribution $p(\sigma_i^2|y_{1,1}, ..., y_{n_i,i})$ is an InverseGamma$(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2})$.
  2. Compute the posterior mean:

$$E(\sigma_i|y_{1,i}, ..., y_{n_i,i}) = \int_0^\infty \sigma_i \cdot p(\sigma_i|y_{1,i}, ..., y_{n_i,i}) d\sigma_i$$

  3. Solve the following equations for (l_i, u_i):

$$\int_0^{l_i} p(\sigma_i|y_{1,i}, ..., y_{n_i,i}) d\sigma_i = 0.025$$

$$\int_0^{u_i} p(\sigma_i|y_{1,i}, ..., y_{n_i,i}) d\sigma_i = 0.975$$

Example: Problem 5.1

- We sampled $B = 10{,}000$ values from the joint posterior distribution $p(\theta_i, \sigma_i^2 | y_{1,i}, ..., y_{n_i,i})$.
- For θ_i, $i = 1, 2, 3$

School	MC estimate of Posterior mean	MC estimate of 95% credible interval
1	9.29	(7.77,10.79)
2	6.94	(5.14,8.73)
3	7.82	(6.18,9.45)

- For σ_i, $i = 1, 2, 3$

School	MC estimate of posterior SD	MC estimate of 95% credible interval
1	3.91	(3.02,5.20)
2	4.40	(3.33,5.92)
3	3.75	(2.79,5.12)

Example: Problem 5.1

- Using the $B = 10{,}000$ independent samples from the joint posterior distribution $p(\theta_i, \sigma_i^2 | y_{1,i}, ..., y_{n_i,i})$ for $i = 1, 2, 3$ we can approximate the posterior probability $P(\theta_i < \theta_j < \theta_k|\text{data})$ for different permutations of (i, j, k).

(i,j,k)	MC estimate of $P(\theta_i < \theta_j < \theta_k\|\text{data})$
(1,2,3)	0.0065
(1,3,2)	0.0049
(2,1,3)	0.0877
(2,3,1)	0.6669
(3,1,2)	0.0137
(3,2,1)	0.2203

Example: Problem 5.1

- We compute the posterior probability that $\tilde{Y}_i < \tilde{Y}_j < \tilde{Y}_k$ for all six permutations $\{i, j, k\}$ of $\{1, 2, 3\}$ by sampling from the posterior predictive distribution.

- We achieve this by doing the following:
  Repeat the following B times: for $j = 1, ..., B$
  1. Sample $(\theta_i^{(j)}, \sigma_i^{2\,(j)})$ from the joint posterior distribution $p(\theta_i, \sigma_i^2 | y_{1,i}, ..., y_{n_i,i})$
  2. Sample $\tilde{y}_i^{(j)}$ from the sampling model $p(y_i | \theta_i^{(j)}, \sigma_i^{2\,(j)})$

- The sample from the joint posterior distribution $p(\theta_i, \sigma_i^2 | y_{1,i}, ..., y_{n_i,i})$ is obtained as before: first sampling σ_i^2 from the marginal posterior distribution and then θ_i from the conditional posterior distribution given the current value of σ_i^2.

Example: Problem 5.1

- Using the $B = 10{,}000$ independent samples from the posterior predictive distribution $p(\tilde{y}_i | y_{1,i}, ..., y_{n_i,i})$ for $i = 1, 2, 3$, we can approximate the posterior probabilities $P(\tilde{Y}_i < \tilde{Y}_j < \tilde{Y}_k | \text{data})$ for different permutations of (i, j, k).

	MC estimate of
(i, j, k)	$P(\tilde{Y}_i < \tilde{Y}_j < \tilde{Y}_k \| \text{data})$
(1,2,3)	0.105
(1,3,2)	0.1037
(2,1,3)	0.1886
(2,3,1)	0.2667
(3,1,2)	0.1397
(3,2,1)	0.1963

Example: Problem 5.1

- Using again the $B = 10{,}000$ independent samples $\{\theta_i^{(j)}, \tilde{y}_i^{(j)}\}_{j=1}^{B}$, we can approximate the posterior probabilities $P(\theta_1 > \theta_2 \text{ and } \theta_1 > \theta_3|\text{data})$ and $P(\tilde{Y}_1 > \tilde{Y}_2 \text{ and } \tilde{Y}_1 > \tilde{Y}_3|\text{data})$.

MC estimate of $P(\theta_1 > \theta_2 \text{ and } \theta_1 > \theta_3|\text{data})$	MC estimate of $P(\tilde{Y}_1 > \tilde{Y}_2 \text{ and } \tilde{Y}_1 > \tilde{Y}_3|\text{data})$
0.8854	0.463