§2 Bayesian Inference

Outline

1. Bayesian inference for the normal distribution
2. Summarisation of posterior distributions
3. Statistical decision theory
4. Comparison of Bayesian and classical inferences
5. Advantages and disadvantages of Bayesian approaches

1. Bayesian inference for the normal distribution

A) Unknown mean, known precision

Suppose we have a sample $\mathbf{y}$ of n independent observations

$$Y_i \mid \theta \sim \text{Normal}(\theta, \sigma^2) , \quad i = 1, \ldots, n,$$

where σ^2 is known and θ is unknown; ie

$$\begin{aligned} p(y_i \mid \theta) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \theta)^2\right] \\ &\propto \exp\left[-\frac{1}{2}\left\{\frac{1}{\sigma^2}(y_i - \theta)^2\right\}\right] . \end{aligned}$$

Sometimes, it is more convenient to work with the *precision* parameter $\tau = \frac{1}{\sigma^2}$. So,

$$Y_i \mid \theta \sim \text{Normal}(\theta, \tau^{-1}) , \quad i = 1, \ldots, n,$$

and thus

$$p(y_i \mid \theta) \propto \exp\left[-\frac{1}{2}\left\{\tau(y_i - \theta)^2\right\}\right] .$$

Classical inference

Maximum likelihood estimation: $\hat{\theta} = \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $\text{Var}(\hat{\theta}) = \sigma^2/n = (n\tau)^{-1}$.

Bayesian inference

Suppose we have a normal prior distribution for the unknown parameter θ:

$$\theta \sim \text{Normal}(\mu_0, \phi_0^{-1}),$$

where μ_0 (prior mean) and ϕ_0 (prior precision) are fixed. Hence,

$$p(\theta) \propto \exp\left[-\frac{1}{2}\left\{\phi_0(\theta - \mu_0)^2\right\}\right] .$$

Then the posterior distribution for θ is

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto p(\theta)\, p(\mathbf{y} \mid \theta) = p(\theta) \prod_{i=1}^n p(y_i \mid \theta) \\ &\propto \exp\left[-\frac{1}{2}\left\{\phi_0(\theta - \mu_0)^2 + \tau\sum_{i=1}^n (y_i - \theta)^2\right\}\right] \end{aligned}$$

$$p(\theta|\mathbf{y}) \propto \exp\left[-\frac{1}{2}\left\{\phi_0(\theta - \mu_0)^2 + \tau\sum_{i=1}^n (y_i - \theta)^2\right\}\right]$$

The part inside $\{\ldots\}$

$$\begin{aligned} &= \theta^2(\phi_0 + n\tau) - 2\theta(\mu_0\phi_0 + n\bar{y}\tau) + \text{const}_1 \\ &= (\phi_0 + n\tau)\left(\theta - \frac{\mu_0\phi_0 + n\bar{y}\tau}{\phi_0 + n\tau}\right)^2 + \text{const}_2 \end{aligned}$$

Therefore,

$$p(\theta|\mathbf{y}) \propto \exp\left[-\frac{1}{2}\left\{(\phi_0 + n\tau)\left(\theta - \frac{\mu_0\phi_0 + n\bar{y}\tau}{\phi_0 + n\tau}\right)^2\right\}\right] .$$

Hence $\theta \mid \mathbf{y} \sim \text{Normal}(\mu_1, \phi_1^{-1})$,

where
$$\begin{aligned} \phi_1 &= \phi_0 + n\tau , \\ \mu_1 &= \frac{\phi_0\mu_0 + n\tau\bar{y}}{\phi_0 + n\tau} = w\mu_0 + (1 - w)\bar{y} , \\ w &= \frac{\phi_0}{\phi_0 + n\tau}, \quad 1 - w = \frac{n\tau}{\phi_0 + n\tau} . \end{aligned}$$

Three interesting notes:

1. The posterior precision ϕ_1 is the sum of the prior precision ϕ_0 and the sample precision $n\tau$.
2. The posterior mean μ_1 is a weighted average of the prior mean μ_0 and the sample mean $\bar{y}$, weighted by their relative precisions.
3. Both θ and $\theta|\mathbf{y}$ follow normal distributions.

$$\phi_1 = \phi_0 + n\tau\ ,\ \mu_1 = w\mu_0 + (1-w)\bar{y}\ ,$$
$$w = \frac{\phi_0}{\phi_0 + n\tau}\ .$$

Three important comments

1. If $n \to \infty$ with ϕ_0 fixed, then $w \to 0$ and $1 - w \to 1$. So, for large enough n,
$$\theta \mid \mathbf{y} \sim \text{Normal}(\bar{y},\, (n\tau)^{-1})$$
approximately, ie posterior does not depend on the prior.
2. If $\phi_0 \to 0$ with n fixed, then $w \to 0$ and $1 - w \to 1$. So, for very diffused prior beliefs (ie non-informative priors),
$$\theta \mid \mathbf{y} \sim \text{Normal}(\bar{y},\, (n\tau)^{-1})$$
approximately, ie posterior does not depend on the prior.
3. If we write $\phi_0 = \kappa_0\tau$, so that prior is Normal $\left(\mu_0, (\kappa_0\tau)^{-1}\right)$, then
$$\theta \mid \mathbf{y} \sim \text{Normal}\left(\frac{n}{n+\kappa_0}\bar{y} + \frac{\kappa_0}{n+\kappa_0}\mu_0,\ \ [(n+\kappa_0)\tau]^{-1}\right).$$
Hence κ_0 may be viewed as a 'prior sample size'.

Example 2.1

The first reliable geochemical datings for the age of the Ennerdale granophyre rock strata were obtained in the 1960s using the K/Ar method (based on the relative proportions of potassium 40 and argon 40 in the rock). **The resulting estimate was 370 (SE 20) million years**. In the 1970s a more accurate method based on the relative proportions of rubidium 87 and strontium 87 became available. **We shall assume that for the latter method the standard deviation of a measurement is known to be 8 million years.**

Rb/Sr measurements were made on **a sample of 5 observations** taken from the rocks and the estimated age calculated from each measurement. **The mean of these estimated ages was 421 million years.**

Using the information from the K/Ar method as a prior, and assuming normally distributed measurement errors, obtain a posterior distribution for θ, the true age of the Ennerdale granophyre.

Prior: $\theta \sim \text{Normal}(370, 20^2)$

Likelihood: $Y_i \mid \theta \sim \text{Normal}(\theta, 8^2)$, $i = 1, \ldots, 5$, with $\bar{y} = 421$

$\Rightarrow$ Posterior: $\theta \mid \mathbf{y} \sim \text{Normal}(\mu_1, \phi_1^{-1})$, where

$$\phi_1 = \frac{1}{20^2} + \frac{5}{8^2} = 0.081\ ,$$
$$w = \frac{1/20^2}{1/20^2 + 5/8^2} = 0.031\ ,$$
$$\mu_1 = 0.031 \times 370 + (1 - 0.031) \times 421 = 419.42\ .$$

B) Known mean, unknown precision

Suppose we have an independent sample $\mathbf{y}$:

$$Y_i \mid \tau \ \sim\ \text{Normal}(\theta, \tau^{-1})\ , \quad i = 1, \ldots, n,$$

where this time τ is unknown and θ is known; ie

$$p(y_i \mid \tau)\ \propto\ \tau^{1/2}\exp\left[-\frac{1}{2}\left\{\tau(y_i - \theta)^2\right\}\right]\ .$$

For reasons to be discussed later ($\tau > 0$; a conjugate prior for this likelihood) it is convenient to choose a gamma distribution as a prior for τ, i.e. $\tau \sim \text{Gamma}(\alpha, \beta)$:

$$p(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)}\tau^{\alpha-1}e^{-\beta\tau} \quad (\tau > 0)$$

where α (the shape parameter, > 0) and β (the inverse scale parameter, > 0) are fixed.

Note that if $\tau \sim \text{Gamma}(\alpha, \beta)$ (precision) then $\tau^{-1} \sim \text{InvGamma}(\alpha, \beta)$ (variance).

The posterior distribution for τ is

$$\begin{aligned}
p(\tau \mid \mathbf{y}) &\propto p(\tau)\, p(\mathbf{y} \mid \tau) = p(\tau) \prod_{i=1}^{n} p(y_i \mid \tau) \\
&\propto p(\tau) \prod_{i=1}^{n} \left\{\tau^{1/2}\exp\left[-\frac{\tau}{2}(y_i - \theta)^2\right]\right\} \\
&\propto \tau^{\alpha-1}\exp\left[-\beta\tau\right] \times \tau^{n/2}\exp\left[-\frac{\tau}{2}ns^2_{(n)}\right] \\
&= \tau^{\frac{n}{2}+\alpha-1}\exp\left[-\left(\frac{ns^2_{(n)}}{2} + \beta\right)\tau\right] \\
\tau \mid \mathbf{y} &\sim \text{Gamma}\left(\frac{n}{2} + \alpha\ ,\ \frac{ns^2_{(n)}}{2} + \beta\right)
\end{aligned}$$

where $s^2_{(n)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$

- Both τ and $\tau|\mathbf{y}$ follow Gamma distributions.

$$\tau \mid \mathbf{y} \sim \text{Gamma}\left(\frac{n}{2}+\alpha\ ,\ \ \frac{ns^2_{(n)}}{2}+\beta\right)$$

Comments

1. shape and inverse scale parameters of posterior distribution are sums of the prior parameters and statistics of the data.
2. posterior shape is $\frac{n}{2}+\alpha = \frac{n+2\alpha}{2}$;
   posterior inverse scale is $\frac{ns^2_{(n)}}{2}+\beta = \frac{ns^2_{(n)}+2\beta}{2}$.

   So, prior Gamma(α, β) can be thought of as providing information equivalent to 2α observations with total sum of squares (ie $\sum_{i=1}^{2\alpha}(y_i-\theta)^2$) equal to 2β.
3. Mean of the posterior distribution
   $$E[\tau \mid \mathbf{y}] = \frac{n/2+\alpha}{ns^2_{(n)}/2+\beta}$$
   - If $s^2_{(n)}$ is large (for fixed n), $E[\tau \mid \mathbf{y}]$ is small; if $s^2_{(n)}$ is small, $E[\tau \mid \mathbf{y}]$ is large. Make sense?
   - When $\alpha \to 0$ and $\beta \to 0$, $E[\tau \mid \mathbf{y}] \to 1/s^2_{(n)}$. Make sense?

C) Unknown mean, unknown precision

Suppose we have an independent sample $\mathbf{y}$:

$$Y_i \mid \theta, \tau \ \sim\ \text{Normal}(\theta, \tau^{-1})\ ,\quad i = 1, \ldots, n,$$

where both τ and θ are unknown.

We now need to specify a *joint* prior distribution $p(\theta, \tau)$. One way to do this is as follows:

$$\begin{aligned} \tau &\sim \text{Gamma}(\alpha, \beta) \\ \theta \mid \tau &\sim \text{Normal}(\mu_0, (\kappa_0\tau)^{-1}) \\ p(\theta, \tau) &= p(\theta \mid \tau)p(\tau) \end{aligned}$$

However, for now we shall simplify the maths by instead assuming a *non-informative* prior for θ and τ:

$$p(\theta, \tau) \propto \tau^{-1}$$

(This is the product of a uniform prior for θ and the Jeffreys' prior for τ. See Lee p84. More on non-informative priors later.)

Joint posterior: $p(\theta, \tau \mid \mathbf{y})$

Multiplying this prior by the normal likelihood $\prod_{i=1}^n p(y_i \mid \theta, \tau)$ gives the joint posterior

$$\begin{aligned} p(\theta, \tau \mid \mathbf{y}) &\propto \tau^{-1}\tau^{n/2} \exp\left[-\frac{\tau}{2}\sum_{i=1}^n (y_i-\theta)^2\right] \\ &= \tau^{n/2-1} \exp\left[-\frac{\tau}{2}\sum_{i=1}^n (y_i-\bar{y}+\bar{y}-\theta)^2\right] \\ &= \tau^{n/2-1} \times \\ &\quad \exp\left[-\frac{\tau}{2}\left\{\sum_{i=1}^n (y_i-\bar{y})^2 + n(\bar{y}-\theta)^2\right\}\right] \\ &= \tau^{n/2-1} \times \\ &\quad \exp\left[-\frac{\tau}{2}\left\{(n-1)s^2 + n(\bar{y}-\theta)^2\right\}\right] \end{aligned}$$

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

Marginal posterior for τ: $p(\tau \mid \mathbf{y})$

$$\begin{aligned} p(\tau \mid \mathbf{y}) &= \int_{-\infty}^{\infty} p(\theta, \tau \mid \mathbf{y}) d\theta \\ &\propto \int_{-\infty}^{\infty} \tau^{\frac{n}{2}-1} \exp\left[-\frac{\tau}{2}\left((n-1)s^2 + n(\bar{y}-\theta)^2\right)\right] d\theta \\ &= \tau^{\frac{n}{2}-1} \exp\left[-\frac{\tau}{2}(n-1)s^2\right] \int_{-\infty}^{\infty} \exp\left[-\frac{\tau n}{2}(\bar{y}-\theta)^2\right] d\theta \end{aligned}$$

Noting that $\int_{-\infty}^{\infty} \exp\left[-\frac{\tau n}{2}(\theta-\bar{y})^2\right] d\theta$ is an un-normalised normal integral gives:

$$\begin{aligned} &= \tau^{\frac{n}{2}-1} \exp\left[-\frac{\tau}{2}(n-1)s^2\right] \sqrt{\frac{2\pi}{n\tau}} \\ &\propto \tau^{\frac{n-1}{2}-1} \exp\left[-\frac{(n-1)s^2}{2}\tau\right]\ . \end{aligned}$$

Therefore,

$$\tau \mid y \sim \text{Gamma}\left(\frac{n-1}{2},\ \ \frac{(n-1)s^2}{2}\right)\ .$$

Note: $E[\tau \mid \mathbf{y}] = \frac{n-1}{2}\Big/\frac{(n-1)s^2}{2} = \frac{1}{s^2}$, where s^2 is the usual unbiased classical estimator of the variance.

Marginal posterior for θ: $p(\theta \mid \mathbf{y})$

$$\begin{aligned} p(\theta \mid \mathbf{y}) &= \int_0^\infty p(\theta, \tau \mid \mathbf{y}) d\tau \\ &\propto \Gamma(\frac{n}{2}) \left[\frac{(n-1)s^2 + n(\bar{y}-\theta)^2}{2}\right]^{-n/2} \\ &\propto \left[1 + \frac{n(\theta-\bar{y})^2}{(n-1)s^2}\right]^{-n/2} . \end{aligned}$$

Using transformation of variable: $\phi = \frac{\theta-\bar{y}}{\sqrt{s^2/n}}$, we get

$$\frac{\theta-\bar{y}}{\sqrt{s^2/n}} \mid \mathbf{y} \sim t_{n-1} .$$

Conditional posterior for $\theta \mid \tau, \mathbf{y}$

$$\theta \mid \tau, \mathbf{y} \sim \text{Normal}\left(\bar{y}, (n\tau)^{-1}\right) .$$

D) Inference about functions of parameters

Let $\phi = g(\theta)$ be a one-to-one transformation. It follows that, given the posterior distribution $p(\theta \mid \mathbf{y})$, we have

$$p_{\phi|\mathbf{y}}(\phi \mid \mathbf{y}) = p_{\theta|\mathbf{y}}(g^{-1}(\phi) \mid \mathbf{y}) \left|\frac{d\theta}{d\phi}\right|$$

Example 2.2

Recall Example 1.2 (Drug efficacy).
Suppose we are interested in inference about the *odds* of being cured by the drug, rather than the probability of cure, where

$$\text{Odds: } \phi = \frac{\theta}{1-\theta}$$

Recall that $\theta \mid \mathbf{y} \sim \text{Beta}(24.2, 18.8)$. What is the posterior of ϕ?

$\theta \mid y \sim \text{Beta}(24.2, 18.8)$. So,

$$p_\Theta(\theta \mid y) = \frac{\Gamma(43)}{\Gamma(24.2)\Gamma(18.8)} \theta^{23.2}(1-\theta)^{17.8} .$$

$$\begin{aligned} \phi &= g(\theta) = \frac{\theta}{1-\theta} \\ \Rightarrow \theta &= g^{-1}(\phi) = \frac{\phi}{1+\phi} \\ \Rightarrow \frac{d\theta}{d\phi} &= \frac{1}{(1+\phi)^2} \end{aligned}$$

So

$$\begin{aligned} p_\Phi(\phi|y) &= \frac{1}{(1+\phi)^2} \frac{\Gamma(43)}{\Gamma(24.2)\Gamma(18.8)} \left(\frac{\phi}{1+\phi}\right)^{23.2} \\ &\quad \times \left(1 - \frac{\phi}{1+\phi}\right)^{17.8} \\ &= \frac{1}{B(24.2, 18.8)} \left(\frac{1}{1+\phi}\right)^2 \left(\frac{\phi}{1+\phi}\right)^{23.2} \left(\frac{1}{1+\phi}\right)^{17.8} \\ &= \frac{1}{B(24.2, 18.8)} \left(\frac{\phi}{1+\phi}\right)^{23.2} \left(\frac{1}{1+\phi}\right)^{19.8} \end{aligned}$$

Note: $B(a,b)$ denotes the Beta function: $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

2. Summarisation of posterior distributions

Bayesian inference is based on the posterior distribution $p(\theta \mid y)$. The posterior encapsulates everything that is known about θ following observation of the data y.

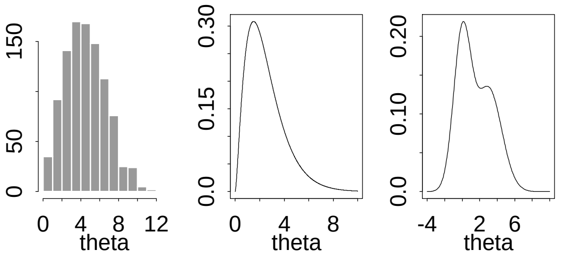
To help understand the mathematical formula for $p(\theta \mid y)$ and to identify its interesting features in a clear and concise way, we could *summarise* the posterior.

Three main ways to summarise $p(\theta \mid y)$:

1. Graphical summaries - e.g. plot the shape of the posterior
2. Quantitative summaries - e.g. measures of location and dispersion
3. Summaries relating to specific hypotheses - e.g. $P(\theta \in \Theta)$

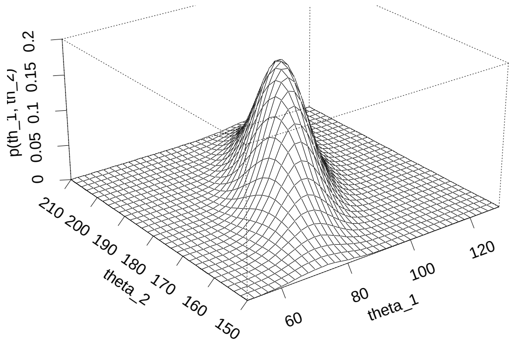
A) Graphical summaries

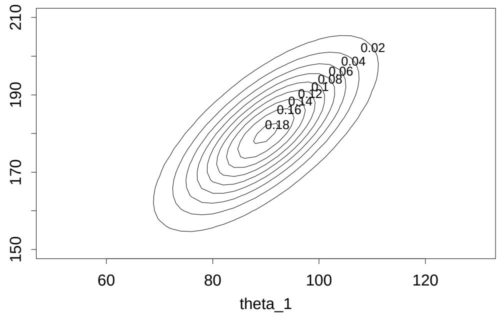
Univariate posteriors, $p(\theta \mid y)$



- Summarises shape of posterior
- Shows range of θ values that are with highest probabilities

Bivariate posteriors, $p(\theta_1, \theta_2 \mid y)$





Multivariate posteriors, $p(\boldsymbol{\theta} \mid y)$

- Most real-life problems involve many parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)$
- Difficult to visualise more than 2 dimensions graphically
- We can reduce the dimension by integrating to obtain one-dimensional marginal posteriors, $p(\theta_i \mid y) = \int p(\boldsymbol{\theta} \mid y) d\boldsymbol{\theta}_{-i}$, and conditional posteriors, $p(\theta_i \mid \boldsymbol{\theta}_{-i}, y)$, for different values of $\boldsymbol{\theta}_{-i}$ (see e.g. pp12-13). However, in many real-life problems these integrals are analytically intractable. $\Rightarrow$ need simulation methods (e.g. MCMC) to evaluate them.

B) Quantitative summaries

Quantitative summaries of Bayesian posteriors usually take the form of *point* and *interval estimates* for the parameters of interest.

- In principle, a Bayesian is free to choose *any* (sensible) numerical summary of the 'location' of θ (ie that provides an indication of the 'typical' value of θ) as a point estimate.

  Common choices include the *mean*, *median* and *mode* of the posterior.

- Likewise, any (sensible) numerical summary of the 'dispersion' of the posterior may be used.

  Common choices include the standard deviation of the posterior distribution, quantile ranges/credible intervals, and highest posterior density (HPD) regions.

Formal 'rules' for choosing optimal point or interval estimators can be obtained by appealing to *Statistical Decision Theory*. A bit more on this later.

Bayesian point estimates

- *Posterior mean*: $E(\theta) = \int \theta p(\theta \mid y) d\theta$
- *Posterior mode*: value θ_{mode} of θ for which $p(\theta \mid y)$ equals its maximum (there may be more than one modes)
- *Posterior median*: value θ_{med} of θ for which $P(\theta \geq \theta_{\text{med}} \mid y) = P(\theta \leq \theta_{\text{med}} \mid y) = 0.5$

Recall Example 2.1: $\theta \mid \mathbf{y} \sim \text{Normal}(419.42, 12.4)$. A point estimate for the age of the rocks is

$E(\theta|\mathbf{y}) = \theta_{\text{mode}} = \theta_{\text{med}} = 419.42$ million years

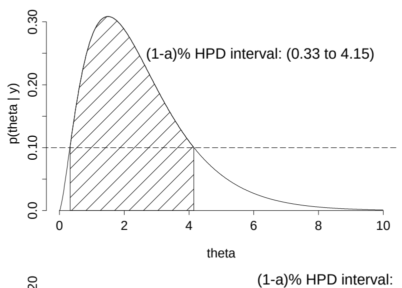
Bayesian interval estimates

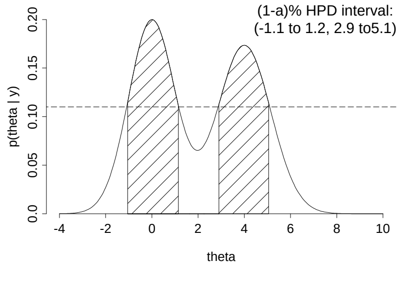
- *Credible interval*: An interval $[c_1, c_2]$ is a 100(1-α)% credible interval for θ if

$$P(\theta \in [c_1, c_2] \mid y) = 1 - \alpha \ .$$

  - E.g.: Let θ_q be the $(100 \times q)$-th percentile of $p(\theta \mid y)$, ie $P(\theta \leq \theta_q \mid y) = q$. Then $[\theta_{.025}, \theta_{.975}]$ is a 95% (central) credible interval for θ.
  - If $p(\theta \mid y)$ is (approximately) Normal(μ_θ, s^2), then the interval $[\mu_\theta - z_{\frac{\alpha}{2}} s, \mu_\theta + z_{\frac{\alpha}{2}} s]$ is a 100(1-α)% credible interval for θ, where $z_{\frac{\alpha}{2}}$ is the value such that $P(Z \leq -z_{\frac{\alpha}{2}}) = P(Z \geq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ for a standard Normal random variable Z.
- *Credible region*: For multivariate $\boldsymbol{\theta}$, a region $\mathbf{C}$ is a 100(1-α)% (simultaneous) credible region for $\boldsymbol{\theta}$ if $P(\boldsymbol{\theta} \in \mathbf{C} \mid y) = 1 - \alpha$.

- *Highest posterior density (HPD) interval*: An interval such that the posterior density at any point inside the interval is greater than that at any point outside.





  - A 100(1-α)% HPD interval is the 100(1-α)% credible interval with the shortest width.
  - Calculation of HPD intervals usually requires the aid of a computer.

Example 2.1 again

Posterior of age of rocks

$$\theta \mid \mathbf{y} \sim \text{Normal}(419.42, 12.4)$$

The 2.5th percentile of this posterior is the value $\theta_{0.025}$ such that

$$P(\theta < \theta_{0.025} \mid \mathbf{y}) = 0.025$$

Now

$$\begin{aligned} P(\theta < \theta_{0.025} \mid \mathbf{y}) &= P\left(\frac{\theta - 419}{\sqrt{12.4}} < \frac{\theta_{0.025} - 419}{\sqrt{12.4}} \mid \mathbf{y}\right) \\ &= \Phi\left(\frac{\theta_{0.025} - 419.42}{\sqrt{12.4}}\right) \end{aligned}$$

$\Phi(-1.96) = 0.025$, so

$$\frac{\theta_{0.025} - 419.42}{\sqrt{12.4}} = -1.96$$

$$\theta_{0.025} = 419.42 - 1.96 \times \sqrt{12.4} = 412.52$$

Likewise, the 97.5th percentile is

$$\theta_{0.975} = 419.42 + 1.96 \times \sqrt{12.4} = 426.32$$

So, a central 95% credible interval for the age of the rocks is [412.52, 426.32] million years.

- Note 1: Due to symmetric and unimodal shape of the normal distribution, this is also the 95% HPD interval.
- Note 2: The interval $(-\infty,\ \theta_{0.95})$ is also a 95% credible interval for the age of the rocks, but it is not necessarily a helpful summary.