Week 2

Shengtong Han

Posterior distribution

Suppose we are interested in k unknown quantities

$\theta = (\theta_1, \cdots, \theta_k)$,

and the prior belief of θ is $p(\theta)$. Assume we have n observations $\mathbf{X} = (X_1, \cdots, X_n)$ that have a density which depends on these k unknown parameters. But the PDF of $\mathbf{X}$ depends on θ in a known way via

$$p(\mathbf{X}|\theta)$$

Basic idea: **take into account both prior belief** $p(\theta)$ **and sample $\mathbf{X}$ to infer** $p(\theta|\mathbf{X})$. By Bayes' theorem

$$p(\theta|\mathbf{X}) \propto p(\theta)p(\mathbf{X}|\theta)$$

where $p(\mathbf{X}|\theta)$ is the **likelihood** function when viewing it as a function of θ, also written as $l(\theta|\mathbf{X}) = p(\mathbf{X}|\theta)$ It is more often to use **log-likelihood** function $L(\theta|\mathbf{X}) = logl(\theta|\mathbf{X})$ $Posterior \propto Prior \times Likelihood$

Standardized likelihood

Note that there is a proportionality in the posterior, and it doesn't alter the result if multiplying $l(\theta|\mathbf{X})$ by any constant free of θ. In fact

$\int l(\theta|\mathbf{X})d\theta$ could be multiple integral

is often finite. Then

$\frac{l(\theta|\mathbf{X})}{\int l(\theta|\mathbf{X})d\theta}$ is called **standardized** likelihood function.

Sequential use of Bayes' theorem

For one sample of observation $\mathbf{X}$,

$p(\theta|\mathbf{X}) \propto p(\theta)l(\theta|\mathbf{X})$.

If we have two independent samples $\mathbf{X}, \mathbf{Y}$,

$$\begin{aligned} p(\theta|\mathbf{X},\mathbf{Y}) &\propto p(\theta)l(\theta|\mathbf{X},\mathbf{Y}) \\ &= p(\theta)p(\mathbf{X},\mathbf{Y}|\theta) \\ &= p(\theta)p(\mathbf{X}|\theta)p(\mathbf{Y}|\theta) \\ &= p(\theta|\mathbf{X})l(\theta|\mathbf{Y}) \end{aligned}$$

where $p(\theta|\mathbf{X})$ could be treated as a prior for sample $\mathbf{Y}$.

Predictive distribution

Normal prior and likelihood

Suppose x is a normal random variable with mean θ and variance ϕ, i.e., $x \sim N(\theta, \phi)$, with PDF

$$p(x) = \frac{1}{\sqrt{2\pi\phi}} exp\{-\frac{(x-\theta)^2}{2\phi}\}$$

with ϕ known. Suppose the prior belief of unknown parameter θ is also normal, $\theta \sim N(\theta_0, \phi_0)$, where θ_0, ϕ_0 are known. That is

$$p(\theta) = \frac{1}{\sqrt{2\pi\phi_0}} exp\{-\frac{(\theta-\theta_0)^2}{2\phi_0}\}$$
$$p(x|\theta) = \frac{1}{\sqrt{2\pi\phi}} exp\{-\frac{(x-\theta)^2}{2\phi}\}$$

The posterior of θ, by Bayes' theorem

$$\begin{aligned} p(\theta|x) &\propto p(\theta)p(x|\theta) \\ &= \frac{1}{\sqrt{2\pi\phi_0}} exp\{-\frac{(\theta-\theta_0)^2}{2\phi_0}\} \times \frac{1}{\sqrt{2\pi\phi}} exp\{-\frac{(x-\theta)^2}{2\phi}\} \\ &\propto exp\Big\{-\frac{1}{2}\theta^2(\frac{1}{\phi_0}+\frac{1}{\phi}) + \theta(\frac{\theta_0}{\phi_0}+\frac{x}{\phi})\Big\} \end{aligned}$$

write $\phi_1 = \frac{1}{\frac{1}{\phi_0}+\frac{1}{\phi}}, \theta_1 = \phi_1(\frac{\theta_0}{\phi_0}+\frac{x}{\phi})$, so $\frac{1}{\phi_0}+\frac{1}{\phi} = \frac{1}{\phi_1}; \frac{\theta_0}{\phi_0}+\frac{x}{\phi} = \frac{\theta_1}{\phi_1}$

thus

$$\begin{aligned} p(\theta|x) &\propto exp\Big\{-\frac{\theta^2}{2\phi_1} + \theta\frac{\theta_1}{\phi_1}\Big\} \\ &\propto exp\Big\{-\frac{(\theta-\theta_1)^2}{2\phi_1}\Big\} \end{aligned}$$

Hence $\theta|x \sim N(\theta_1, \phi_1)$

- precision: $\frac{1}{\phi_1} = \frac{1}{\phi_0} + \frac{1}{\phi}$. Posterior precision=Prior precision+Data precision
- mean: $\theta_1 = \theta_0\frac{\phi_0^{-1}}{\phi_0^{-1}+\phi^{-1}} + x\frac{\phi^{-1}}{\phi_0^{-1}+\phi^{-1}}$

Posterior mean=weighted mean of prior mean and data value the weights is proportional to their respective precisions

Examples

The age of Ennerdale granophyre was measured as 370 ± 20 million by K/Ar method and 420 ± 8 million years by Rb/Sr method in 1960s and 1970s respectively. It is reasonable to assume that errors are normally distributed and the errors marked are meant to be standard deviations. Suppose a scientist has a measure in early 1970s, and his prior was represented as $\theta \sim N(370, 20^2)$, and suppose that Rb/Sr method results in $x \sim N(\theta, 8^2)$

The posterior of θ is

$$\theta|x \sim N(\theta_1, \phi_1)$$

where $\phi_1 = \frac{1}{\frac{1}{\phi_0}+\frac{1}{\phi}} = \frac{1}{20^{-2}+8^{-2}} = 7.4^2$, $\theta_1 = \theta_0 \frac{\phi_0^{-1}}{\phi_0^{-1}+\phi^{-1}} + x\frac{\phi^{-1}}{\phi_0^{-1}+\phi^{-1}} = 370\frac{20^{-2}}{20^{-2}+8^{-2}} + 421\frac{8^{-2}}{20^{-2}+8^{-2}} = 413$ hence

$\theta|x \sim N(431, 7.4^2)$

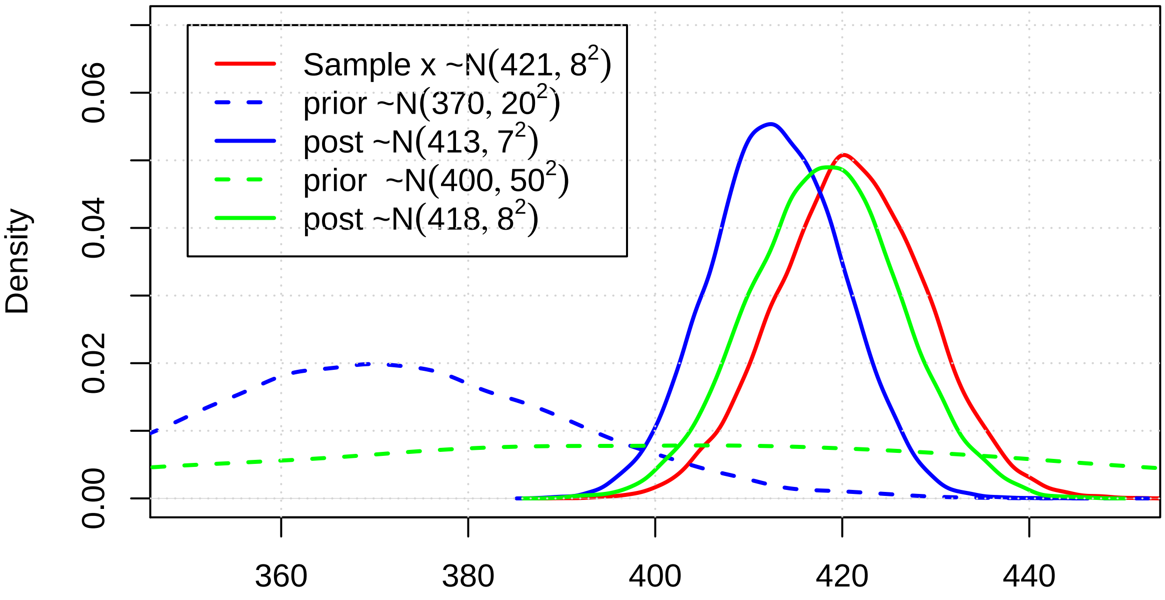
If another scientist does not have K/Ar measurement as prior, but had a vague idea that it is likely to be 400 ± 50 million years, i.e., $\theta \sim N(400, 50^2)$. Then the posterior mean and variance are

$$\phi_1 = (50^{-2} + 8^{-2})^{-1} = 8^2$$
$$\theta_1 = 62(400 \times 50^{-1} + 421 \times 8^{-2}) = 418$$

$\theta|x \sim N(418, 8^2)$.

In both cases, the posterior is almost determined by the data

Prior	Data	Posterior
$N(370, 20^2)$	$N(421, 8^2)$	$N(413, 7^2)$
$N(400, 50^2)$		$N(418, 8^2)$



Assumptions

- The distribution of observation x is assumed to be **normal**, but with only one unknown parameter θ.
- The variance in the normal distribution, ϕ is also assumed to be known, which is hard to justify.

Several normal observations and a normal prior

Posterior distributions

Suppose we have a prior $\theta \sim N(\theta_0, \phi_0)$ and n independent observations $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ such that $x_i \sim N(\theta, \phi)$. By Bayes' theorem,

$$\begin{aligned}
p(\theta|\mathbf{x}) &\propto p(\theta)p(\mathbf{x}|\theta) \\
&= p(\theta)p(x_1|\theta)p(x_2|\theta)\cdots p(x_n|\theta) \ (\textit{by independence}) \\
&= (2\pi\phi_0)^{-\frac{1}{2}} exp\Big\{-\frac{(\theta-\theta_0)^2}{2\phi_0}\Big\} \\
&\times (2\pi\phi)^{-\frac{1}{2}} exp\Big\{-\frac{(x_1-\theta)^2}{2\phi}\Big\} \times \cdots \times (2\pi\phi)^{-\frac{1}{2}} exp\Big\{-\frac{(x_n-\theta)^2}{2\phi}\Big\} \\
&\propto exp\Big\{-\frac{\theta^2}{2}(\frac{1}{\theta_0}+\frac{n}{\phi}) + \theta(\frac{\theta_0}{\phi_0}+\frac{\sum x_i}{\phi})\Big\}
\end{aligned}$$

write

$$\phi_1 = \frac{1}{\frac{1}{\phi_0}+\frac{n}{\phi}}$$

$$\theta_1 = \phi_1(\frac{\theta_0}{\phi_0}+\frac{\sum x_i}{\phi})$$

then $\theta|\mathbf{x} \sim N(\theta_1, \phi_1)$.

Alternatively, write mean and variance as

$$\phi_1 = \frac{1}{\phi_0^{-1}+(\phi/n)^{-1}}$$

$$\theta_1 = \phi_1\Big\{\theta_0/\phi_0 + \bar{x}/(\phi/n)\Big\}$$

same as the posterior obtained from the single observation of the mean $\bar{x}$ as $\bar{x} \sim N(\theta, \frac{\phi}{n})$

Example

Consider the chest measurement of 10000 men and a prior $N(38, 9)$. Whitaker and Robbinson's data show that the mean turned out to be 39.8 with a standard deviation of 2.0. By combining prior and sample data

$$\phi_1 = \frac{1}{9^{-1}+(2^2/10000)^{-1}} = \frac{1}{2500}$$

$$\theta_1 = \frac{1}{2500}(38/9 + \frac{39.8}{2^2/10000}) = 39.8$$

Predictive distribution

Consider the predictive distribution of one observation x_{n+1}, since $x_{n+1} = (x_{n+1}-\theta)+\theta$ because of the independence of one another

$$(x_{n+1}-\theta) \sim N(0, \theta)$$

$$\theta \sim N(\theta_1, \phi_1)$$

so $x_{n+1} \sim N(\theta_1, \phi + \phi_1)$

Dominant likelihoods

Improper priors

So far we know if we have several normal observations and a normal prior and known variance, the posterior for the mean is $N(\theta_1, \phi_1)$, where θ_1, ϕ_1 are given by appropriate formulae.

- Consider the prior $N(\theta_0, \infty)$ which has to be the uniform over the whole real line, couldn't be used for any proper density function. Similarly $p(\theta) = \kappa, (-\infty < \theta < \infty)$ can not represent a probability density whatever κ is.
- **improper** density $p(\theta)$, if $\int_{-\infty}^{\infty} p(\theta) d\theta = \infty$. Another example is $p(\theta) = \frac{\kappa}{\theta}, 0 < \theta < \infty$.

Sometimes improper prior could be combined with an ordinary likelihood to give a proper posterior. For instance, if we use prior $p(\theta) = \kappa, \kappa \neq 0$ on the whole real line, combining it with normal likelihood gives standard likelihood as posterior. **Dominant feature of posterior is the likelihood**.

Locally uniform priors

A prior which does not change very much over the region in which the likelihood is appreciable and does not take very large values outside the region is said to be locally uniform. For such a prior $p(\theta|x) \propto p(x|\theta) = l(\theta|x)$ because prior is uniform

Bayes postulate

The situation that we know “nothing" about θ should be represented by a uniform prior.

For example

$$p(\theta) = 1, \quad (0 < \theta < 1)$$

Let $\phi = \frac{1}{\theta}$. According to the change of variable rule $p(\phi)|d\phi| = p(\theta)|d\theta|$,

$$\begin{aligned} p(\phi) &= p(\theta|d\theta/d\phi|) \\ &= \frac{1}{\phi^2} \quad (1 < \phi < \infty) \end{aligned}$$

Not a uniform prior.

Data translated likelihoods

The likelihood is **data translated** if it is in a form of

$$l(\theta|x) = g(\theta - t(x))$$

for some function t.

For example, the likelihood of n sample normal distribution with unknown mean and known variance is

$$l(\theta|x) = exp\Big\{ - \frac{(\theta - \bar{x})^2}{2\phi/n} \Big\}$$

is of this from.

If k has a binomial distribution with parameter π,

$$l(\pi|k) \propto \pi^k(1-\pi)^{n-k}$$

can not be expressed in the form of $g(\pi - t(k))$.

If the likelihood is in data translated form, different values of the data will give rise to the same functional form for the likelihood except for a shift in location.

For the normal mean, suppose we have two experiments, one of which has $\bar{x}$ 5 larger than another, then both experiment have the same likelihood function except the corresponding value of θ differs by 5.

Transformation of unknown parameters

When the likelihood is not in data translated form, there may be a function $\psi = \psi(\theta)$ such that

$$l(\theta|x) = g(\psi(\theta) - t(x))$$

In such a case, the prior information may be put in ψ, rather than in θ.

Suppose $x \sim E(\theta)$, then

$$p(x|\theta) = \frac{exp(-x/\theta)}{\theta}$$

By multiplying x

$$\begin{aligned} l(\theta|x) &= \frac{x}{\theta}exp(-x/\theta) \\ &= exp\Big\{(logx - log\theta) - exp(logx - log\theta)\Big\} \end{aligned}$$

Let

$$g(y) = exp\{y - exp(-y)\}$$

$$\begin{aligned} t(x) &= logx \\ \psi(\theta) &= log\theta \end{aligned}$$

It is often difficult to express a likelihood function in this from when it is possible and it is not always possible.

Highest density regions (HDR)

Summary of posterior information

- One way to summarize the posterior information is simply presenting the distribution, say $\theta \sim N(413, 7^2)$.
- Sometimes the probability that the parameter lies in a $\{$*particular interval} may be of interest, say* $p\{\theta < 400\}$
- *Highest density regions: an interval in which "most of the distribution" lies; the density at any point inside it is greater than the density at any point outside it. It is also called {Bayesian confidence interval, credible interval.*

add a pdf with 95% area shading

Due to the fact that 95% of area of a normal distribution is within ±1.96 standard deviations of the mean, that is $413 \pm 1.96 \times 7$ which is $(399, 427)$.

Relation to classical statistics

- In classical approach, x is regarded as random and gives rise to random interval which has a probability of 95% containing the fixed but unknown parameter θ.

$$|\frac{\theta - \tilde{x}}{\sqrt{\phi}}| < 1.96$$

- In Bayesian approach, θ is regraded as random and interval is fixed once the data is available.

$$|\frac{\tilde{\theta} - x}{\sqrt{\phi}}| < 1.96$$

Normal variance

Prior for normal variance

Suppose we have sample $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ from $N(\mu, \phi)$ where μ is known and ϕ is unknown. So

$$\begin{aligned} p(\mathbf{x}|\phi) &= \frac{1}{\sqrt{2\pi\phi}} exp\{-\frac{(x_1-\mu)^2}{2\phi}\} \\ &\times \cdots \times \frac{1}{\sqrt{2\pi\phi}} exp\{-\frac{(x_n-\mu)^2}{2\phi}\} \\ &\propto \phi^{-\frac{n}{2}} exp\Big\{-\frac{\sum(x_i-\mu)^2}{2\phi}\Big\} \\ &= \phi^{-\frac{n}{2}} exp\Big\{-\frac{S}{2\phi}\Big\} \end{aligned}$$

where $S = \sum(x_i - \mu)^2$. In principle, the prior of ϕ can be of any form, but to make the posterior distribution easy to deal with, the prior may be chosen of a similar form to the likelihood, $p(\phi) \propto \phi^{-\frac{\kappa}{2}} exp(-\frac{S_0}{2\phi})$, where κ, S_0 are suitable constants.

The posterior is

$$\begin{aligned} p(\phi|\mathbf{x}) &\propto p(\phi)p(\mathbf{x}|\phi) \\ &\propto \phi^{-\frac{v+n}{2}-1} exp\Big\{-\frac{S_0+S}{2\phi}\Big\} \end{aligned}$$

where $\kappa = v + 2$. Note the true prior for ϕ may not be exactly represented by such a density, but in most cases they can be reasonably approximated by such kind of form. This posterior is close to density of χ^2 distribution.

Let $\lambda = \frac{1}{\phi}$, by the change of variable rule

$$\begin{aligned} p(\lambda|\mathbf{x}) &\propto p(\phi|\mathbf{x})|\frac{d\phi}{d\lambda}| \\ &\propto \lambda^{\frac{v+n}{2}+1} exp\Big\{-\frac{(S_0+S)\lambda}{2}\Big\} \times \lambda^{-2} \\ &\propto \lambda^{\frac{v+n}{2}-1} exp\Big\{-\frac{(S_0+S)\lambda}{2}\Big\} \end{aligned}$$

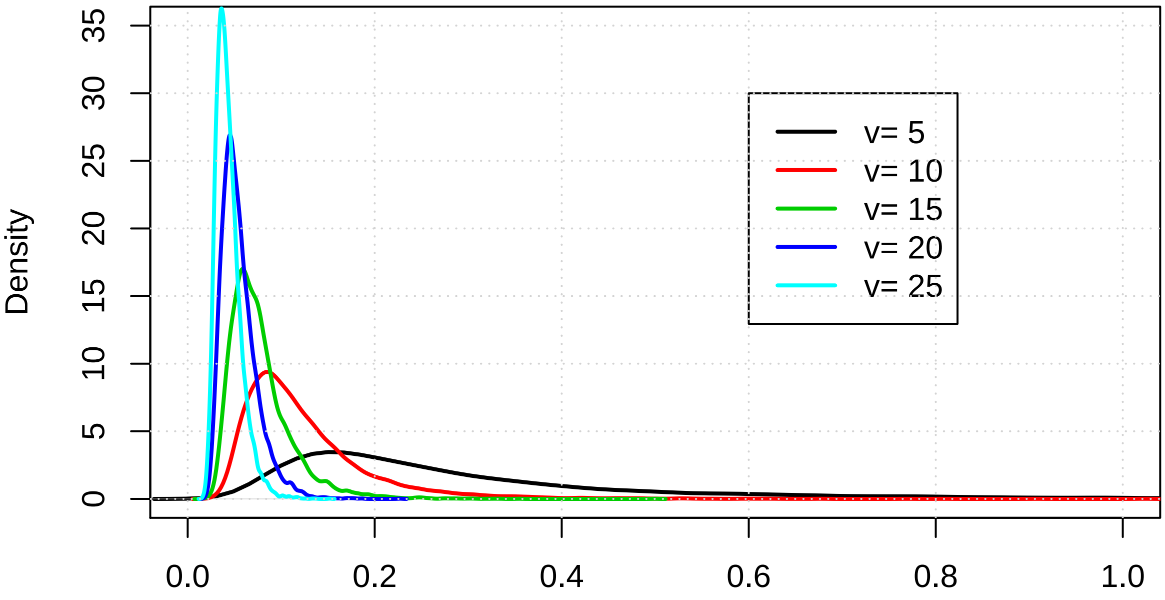
It follows that $(S_0 + S)\lambda \sim \chi^2_{v+n}$, furthermore $\phi \sim (S_0 + S)\chi^{-2}_{(v+n)}$, which is **inverse chi-squared distribution**.

So our prior can be chosen

$$\phi \sim S_0 \chi_v^{-2}$$

- it may not be straightforward to choose v, S_0
- the prior does not need to be too close since the likelihood will dominate
- consider the mean and variance, $E(\phi) = \frac{S_0}{v-2}, var(\phi) = \frac{2S_0^2}{(v-2)^2(v-4)}$ when choosing prior

```
## ----------------------------------------------------------------
##  Analysis of Geostatistical Data
##  For an Introduction to geoR go to http://www.leg.ufpr.br/geoR
##  geoR version 1.7-5.2.1 (built on 2016-05-02) is now loaded
## ----------------------------------------------------------------
```



Reference prior for normal variance

HDR for the normal variance

Which distribution to use

To find HDR of normal variance, we can use

- distribution table: because the distribution of variance is a multiple of the inverse chi-squared distribution
- reference prior: which was uniform in $log(\phi)$, use $log(\phi)$ in the posterior distribution and look for an interval insider which the posterior density of $log(\phi)$ is higher than anywhere outside.

Example

Uterine weight (in mg) of 20 rats drawn randomly from a large stock

Table 2: Uterine weights of 20 rats

9	18	21	26
14	18	22	27
15	19	22	29
15	19	24	30
16	20	24	32

Obviously $n = 20, \sum x_i = 420, \sum x_i^2 = 9484$ and $\bar{x} = 21$,

$$S = \sum (x_i - \bar{x})^2 = 664$$

Suppose we know the mean and can assume

$$\phi \propto 664 \chi_{20}^{-2}$$

- 95% HDR for log chi-squared are 9.958 and 35.227 and the interval for ϕ is $(\frac{664}{35.227}, \frac{664}{9.958})$, that is $(19, 67)$.
- 95 % HDR for inverse chi-squared is $(0.025, 0.094)$, so the interval for ϕ is $(664 \times 0.025, 664 \times 0.094)$, that is $(17, 62)$.