Introduction to Bayesian Inference

Notes by Best, Mason, and Li
(edited by Heejung Shim)

Learning goals

- Review Bayesian Statistics (taught in MAST20005).
- Understand the posterior predictive distribution.

Probability

Conventional frequentist statistics
The relative frequency of occurrence of an event, observed in repetitions of an experiment.

Bayesian statistics
The degree of plausibility, or strength of belief, of an event.

Thus, direct expression of uncertainty about unknown parameters.

credible interval

- No (difficult to interpret) confidence intervals: just report, say, central area that contains 95% of distribution for the parameter of interest.
- There is a procedure for adapting the distribution in the light of additional evidence: i.e. Bayes theorem allows us to learn from experience.

Example

A clinical trial is carried out to collect evidence about an unknown 'treatment effect'.

Conventional frequentist analysis

- p-value for H0: treatment effect is zero.
- Point estimate and CI as summaries of size of treatment effect.
- Aim is to learn what this trial tells us about the treatment effect.

Bayesian analysis

- Inference is based on probability statements summarising the posterior distribution of the treatment effect.
- Asks: 'how should this trial change our opinion about the treatment effect?'.

Components of a Bayesian analysis

The Bayesian analyst needs to explicitly state

- a reasonable opinion concerning the plausibility of different values of the treatment effect excluding the evidence from the trial (the **prior distribution**)
- the support for different values of the treatment effect based solely on data from the trial (the **likelihood**),

and to combine these two sources to produce

- a final opinion about the treatment effect (the **posterior distribution**)

The final combination is done using Bayes theorem (and only simple rules of probability), which essentially weights the likelihood from the trial with the relative plausibilities defined by the prior distribution.

One can view the Bayesian approach as a formalisation of the process of learning from experience.

Bayesian inference - the posterior distribution

Posterior distribution forms basis for all inference - can be summarised to provide

- point and interval estimates for the parameters of interest (e.g. treatment effect)
- point and interval estimates of any function of the parameters
- probability that the parameter of interest exceeds a critical threshold
- prior information for future experiments, trials, surveys, etc.
- ...

Bayesian inference

Makes fundamental distinction between

- Observable quantities x, i.e. the data
- Unknown quantities θ

θ can be statistical parameters.

- parameters are treated as random variables.
- in the Bayesian framework, we make probability statements about model parameters.

prior : probable distribution about parameters before observing the data

In the frequentist framework, parameters are fixed non-random quantities and the probability statements concern the data.

Bayesian inference

As with any analysis, we start by positing a model, $p(x|\theta)$, which is the **likelihood**.

From a Bayesian point of view

- θ is unknown so should have a **probability distribution** reflecting our uncertainty about it before seeing the data
  - Need to specify a **prior distribution** $p(\theta)$.
- x is known so we should condition on it.
  - Use Bayes theorem to obtain conditional probability distribution for unobserved parameters of interest given the data:

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)} = \frac{p(\theta)p(x|\theta)}{\int p(\theta)p(x|\theta)d\theta} \propto p(\theta)p(x|\theta)$$

  - This is the **posterior distribution**.

Bayesian inference

- The prior distribution $p(\theta)$, expresses our uncertainty about θ **before** seeing the data.
- The posterior distribution $p(\theta|x)$, expresses our uncertainty about θ **after** seeing the data.

Bayesian inference - remarks

1. $p(\theta)$: the prior distribution on θ
2. $f(x|\theta)$: the likelihood of θ
3. $p(\theta|x) = \frac{f(x|\theta)p(\theta)}{f(x)}$: the posterior distribution of θ
4. $f(x) = \int_\Theta f(x|\theta)p(\theta)d\theta$: the marginal distribution of X

$p(x,\theta)$

Bayesian inference for the Normal distribution

Known variance, unknown mean

- Suppose we have a sample of Normal data $x_i \sim \mathrm{N}(\theta, \sigma^2)$ $i = 1, \ldots, n$.
- For now assume σ^2 is known and θ has a Normal prior $\theta \sim \mathrm{N}(\mu, \sigma^2/n_0)$.
  - Same standard deviation σ is used in the likelihood and the prior.
  - Prior variance is based on an 'implicit' sample size n_0.
- Then straightforward to show that the posterior distribution is

$$\theta|\mathrm{x} \sim \mathrm{N}\left(\frac{n_0\mu + n\bar{x}}{n_0 + n}, \frac{\sigma^2}{n_0 + n}\right)$$

Bayesian inference for the Normal distribution - comments

Model : $x_i \sim \mathrm{N}(\theta, \sigma^2)$ $i = 1, \ldots, n$.
Prior : $\theta \sim \mathrm{N}(\mu, \sigma^2/n_0)$
Posterior : $\theta|\mathrm{x} \sim \mathrm{N}(\frac{n_0\mu+n\bar{x}}{n_0+n}, \frac{\sigma^2}{n_0+n})$

- Compare with frequentist setting: the MLE is $\hat{\theta} = \bar{x}$ with $\mathrm{SE}(\hat{\theta}) = \sigma/\sqrt{n}$, and sampling distribution is

$$p(\hat{\theta}|\theta) = p(\bar{x}|\theta) = \mathrm{N}(\theta, \sigma^2/n).$$

- As n_0 tends to 0, the prior variance becomes larger and the distribution becomes 'flatter', and in the limit the prior distribution becomes essentially uniform over (-∞, ∞).

Bayesian inference for the Normal distribution - comments

Model : $x_i \sim \mathrm{N}(\theta, \sigma^2)$ $i = 1, \ldots, n$.
Prior : $\theta \sim \mathrm{N}(\mu, \sigma^2/n_0)$
Posterior : $\theta|\mathrm{x} \sim \mathrm{N}(\frac{n_0\mu + n\bar{x}}{n_0+n}, \frac{\sigma^2}{n_0+n})$

- Posterior mean is
  - a weighted average of the prior mean μ and the MLE $\bar{x}$, weighted by their precisions (relative 'sample sizes').
  - always a compromise between the prior mean and the MLE.
- Posterior variance is
  - based on an implicit sample size equivalent to the sum of the prior 'sample size' n_0 and the sample size of the data n.
  - less than each of the prior variance and the variance of MLE.
- As $n \to \infty$, $p(\theta|\mathrm{x}) \to \mathrm{N}(\bar{x}, \sigma^2/n)$
  - posterior mean $\to$ MLE; posterior variance $\to$ variance of MLE.
  - posterior does not depend on the prior.
  - posterior distribution gets more concentrated on MLE.

Bayesian inference for the Normal distribution - comments

Model : $x_i \sim \mathrm{N}(\theta, \sigma^2)$ $i = 1, \ldots, n$.
Prior : $\theta \sim \mathrm{N}(\mu, \sigma^2/n_0)$
Posterior : $\theta|\mathrm{x} \sim \mathrm{N}(\frac{n_0\mu + n\bar{x}}{n_0+n}, \frac{\sigma^2}{n_0+n})$

- Posterior mean is
  - a weighted average of the prior mean μ and the MLE $\bar{x}$, weighted by their precisions (relative 'sample sizes').
  - always a compromise between the prior mean and the MLE.
- Posterior variance is
  - based on an implicit sample size equivalent to the sum of the prior 'sample size' n_0 and the sample size of the data n.
  - less than each of the prior variance and the variance of MLE.
- As $n \to \infty$, $p(\theta|\mathrm{x}) \to \mathrm{N}(\bar{x}, \sigma^2/n)$
  - posterior mean $\to$ MLE; posterior variance $\to$ variance of MLE.
  - posterior does not depend on the prior.
  - posterior distribution gets more concentrated on MLE.

These observations in blue are generally true, when MLE exists and is unique.

Bayesian inference for the Normal distribution - example : THM concentrations

$\sigma = 5$

$x_z^1, x_z^2 \sim N(\theta_z, 5^2)$

$\bar{x}_z = 130 \sim N(\theta_z, \frac{5^2}{2})$

Regional water companies in the UK are required to take routine measurements of trihalomethane (THM) concentrations in tap water samples for regulatory purposes. Samples tested throughout year in each water supply zone.

Suppose we want to estimate the THM concentration in a particular water zone, z. Two independent measurements, x_z^1 and x_z^2 are taken and their mean, $\bar{x}_z$ is 130 μg/l. Suppose we know that the assay measurement error has a standard deviation $\sigma = 5$ μg/l.

Let the THM concentration in this water zone be denoted θ_z. Perform statistical inference on θ_z?

Bayesian inference for the Normal distribution - example : THM concentrations

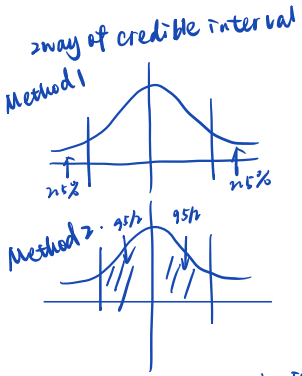
Conventional frequentist analysis

Conventional frequentist analysis would use sample mean $\bar{x}_z = 130\mu$g/l as an estimate of θ_z, with standard error $\sigma/\sqrt{n} = 5/\sqrt{2} = 3.5\mu$g/l.

95% CI for θ_z: $\bar{x}_z \pm 1.96 \times \sigma/\sqrt{n}$, i.e., 123.1 to 136.9 μg/l.

$\bar{x}_z \sim N(\theta_z, \frac{\sigma^2}{n})$

Bayesian inference for the Normal distribution - example : THM concentrations



Bayesian analysis

Suppose historical data on THM concentrations in other zones supplied from the same source showed that the THM concentration was 120 μg/l with standard deviation 10 μg/l. It suggests Normal(120, 10^2) prior for θ_z.

prior
$\theta_z \sim N(120, \frac{\sigma^2}{n_0})$
$\sigma = 5 \Rightarrow n_0 = 0.25$

If we express the prior standard deviation as $\sigma/\sqrt{n_0}$, $n_0 = (\sigma/10)^2 = 0.25$. So the prior can be written as $\theta_z \sim \text{Normal}(120, \sigma^2/0.25)$, where $\sigma = 5$.

Then, the posterior for θ_z is then

$$\begin{aligned} p(\theta_z|x) &= \text{Normal}\left(\frac{0.25 \times 120 + 2 \times 130}{0.25 + 2}, \frac{5^2}{0.25 + 2}\right) \\ &= \text{Normal}(128.9, 3.33^2) \end{aligned}$$

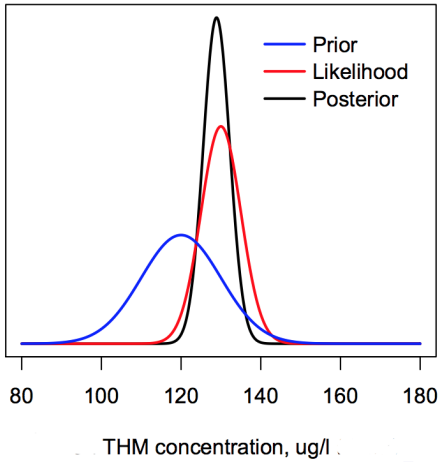
posterior
symmetric
2.5%
2.5%

when you report credible interval, should report how you compute.

giving 95% credible interval for θ_z of 122.4 to 135.4 μg/l.

Bayesian inference for the Normal distribution - example : THM concentrations

Visualisation of the prior, likelihood and posterior for the THM concentration θ_z



Posterior predictive distribution - perhaps new concept?

The **posterior predictive distribution** for a new observation $\tilde{x}$ is

$$p(\tilde{x}|\mathrm{x}) = \int p(\tilde{x}|\mathrm{x}, \theta)p(\theta|\mathrm{x})d\theta$$

which generally simplifies to

$$p(\tilde{x}|\mathrm{x}) = \int p(\tilde{x}|\theta)p(\theta|\mathrm{x})d\theta.$$

$X=(x_1,\cdots x_n)$ number

$\tilde{X}$ r.v

$p(\tilde{x}|x)=\int p(\tilde{x},\theta|x)\,d\theta$

$=\int p(\tilde{x}|x,\theta)\cdot p(\theta|x)\,d\theta$

$x_1,\cdots x_n,\ \tilde{x}\sim f(x|\theta)$

$\tilde{x}$, condition on θ, doesn't depend on $x_1,\cdots x_n$

Posterior predictive distribution for the Normal distribution

Model : $x_i \sim \mathrm{N}(\theta, \sigma^2)$ $i = 1, \ldots, n$.
Prior : $\theta \sim \mathrm{N}(\mu, \sigma^2/n_0)$
Posterior : $\theta|\mathrm{x} \sim \mathrm{N}(\frac{n_0\mu+n\bar{x}}{n_0+n}, \frac{\sigma^2}{n_0+n})$

$\theta|\mathrm{x} \sim \mathrm{N}(\frac{n_0\mu + n\bar{x}}{n_0 + n}, \frac{\sigma^2}{n_0 + n})$

$p(\tilde{x}|x) = \int p(\tilde{x}|\theta)\, p(\theta|x)\, d\theta$

$\propto \int \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\tilde{x}-\theta)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n_0+n}}} e^{-\frac{(n_0+n)(\theta - \frac{n_0\mu+n\bar{x}}{n_0+n})^2}{2\sigma^2}} d\theta$

Exercise: you can show that the posterior predictive distribution for a new observation $\tilde{x}$:

$$p(\tilde{x}|\mathrm{x}) \sim \mathrm{N}(\frac{n_0\mu + n\bar{x}}{n_0 + n}, \frac{\sigma^2}{n_0 + n} + \sigma^2)$$

$\frac{\sqrt{n_0+n}}{2\pi\sigma^2} \int e^{-\frac{1}{2\sigma^2}\left[\theta^2 - 2\tilde{x}\theta + \tilde{x}^2 + (n_0+n)\theta^2 - 2(n_0\mu+n\bar{x})\theta + \cdots\right]^2} d\theta$

$\Downarrow$

$(1+n_0+n)\theta^2$

Comment: the posterior predictive distribution is centred around the posterior mean with variance equal to sum of the posterior variance and the sample variance of $\tilde{x}$.

Posterior predictive distribution for the Normal distribution - example : THM concentrations

Suppose the water company will be fined if THM concentration in the water supply exceed 145 μg/l. Compute the posterior probability that THM concentration in a future sample taken from the water zone exceeds 145μg/l.

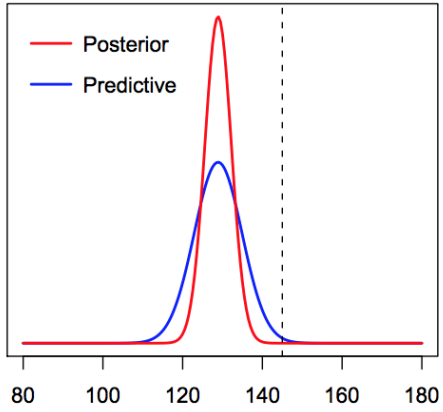
The posterior predictive distribution for THM concentration in a future sample taken from the water zone is

$$\mathrm{N}(128.9, 3.33^2 + 5^2) = \mathrm{N}(128.9, 36.1).$$

The posterior probability that THM concentration in a future sample exceeds 145μg/l is $1 - \Phi[(145 - 128.9)/\sqrt{36.1}] = 0.004$.

Posterior predictive distribution for the Normal distribution - example : THM concentrations

Visualisation of the posterior and posterior predictive distribution for the THM concentration θ_z



$$\theta|x \sim N(\frac{n_0\mu + n\bar{x}}{n_0 + n}, \frac{\sigma^2}{n_0 + n})$$

posterior
Is not closed form

Learning about the posterior on parameters using Monte Carlo methods

In the Normal distribution example, we have calculated the posterior distributions in closed form. So we can make use of known properties of the closed-form posterior distribution to make Bayesian inference. For example, posterior mean and tail-area probability are known analytically.

However, the posterior distribution does not always have a closed form.

Using Monte Carlo methods, we can make Bayesian inference without explicitly specifying posterior.

- Just specify the prior and likelihood.
- There are algorithms to summarise the posterior given (almost) arbitrary specification of the prior and likelihood.

Learning goals

- Review Bayesian Statistics (taught in MAST20005).
- Understand the posterior predictive distribution.