C) Summaries for specific hypotheses

One of the most powerful features of Bayesian inference is that it can answer *any* question of interest by appropriate summarisation of the posterior.

E.g. recall Example 2.1. Suppose the hypotheses are $H_0 : \theta < \theta_0$ and $H_1 : \theta \geq \theta_0$ with equal losses for type I and type II errors. Then we reject H_0 if

$$P(\theta < \theta_0 \mid \mathbf{y}) < P(\theta > \theta_0 \mid \mathbf{y})$$

$$\begin{aligned} P(\theta < 420 \mid \mathbf{y}) &= \int_{-\infty}^{420} p(\theta \mid \mathbf{y}) d\theta \\ &= \Phi\left(\frac{420 - 419.42}{\sqrt{12.4}}\right) \\ &= 0.57 \end{aligned}$$

Suppose the losses for type I and II errors are c_α and c_β, respectively. Then we reject H_0 if

$$c_\alpha P(\theta < \theta_0 \mid \mathbf{y}) < c_\beta P(\theta > \theta_0 \mid \mathbf{y}) \ .$$

Comments

Consider a classical hypothesis test of

$$H_0 : \theta < \theta_0 \text{ vs } H_1 : \theta \geq \theta_0,$$

and suppose that we got a p-value of 0.02.

- Does this mean that $P(\theta < \theta_0) = 0.02$?
- This means that: if H_0 is true (ie $\theta < \theta_0$), then the probability of observing data at least as extreme as the data actually observed is at most 0.02.

Classical inference cannot provide probability statements about parameters!

Brief summary of Bayesian inference: Bayesian inference involves calculating posterior distributions and summarising them using appropriate graphical and numerical summaries.

3. Statistical decision theory

"Statistical decision theory is an approach to decision-making in the presence of uncertainty."

- Statistical decision theory is both an important application and justification of Bayesian inference.
- It places *all* statistical inference in the context of decision-making, by associating different actions with different values of the unknown parameter(s) through a loss function.
- The aim is to choose the action that minimises the expected loss.

Example 2.3

Suppose James Bond applies for life insurance. If Bond dies within ten years he wants the insurance company to pay out. He tells the company he can afford to pay a premium of Q pounds and wants a payout of S pounds in the event of his death. Should the company agree to insure him?

- Let $\theta = 1$ if Bond survives the next ten years, and $\theta = 0$ if he dies.
- The company calculates $p(\theta|y)$ with the data y available, e.g. his age, sex, medical history and profits made from the last Bond film.
- On the basis of $p(\theta|y)$, the company works out its expected loss if it insures or does not insure him.
- The company must take a decision: one 'action' is to agree to insure; the other 'action' is to refuse. How to?: if the expected loss for insuring him is less than that for no insuring him, the company agrees; otherwise it refuses.

Formally, a decision problem has three elements:

1. Parameter space Θ (i.e. set of possible values of θ)
2. Set of actions A (i.e. set of available actions a)
3. Loss function $L(\theta, a)$ defining the loss suffered from taking action a when the true parameter value is θ. $L(\theta, a) \geq 0$

A *decision function* $d(y)$ maps the observed data y onto an action $a = d(y)$. Hence, for $d(y)$, the posterior expected loss is

$$E_{\theta|y}[L(\theta, d(y))] = \int L(\theta, d(y))p(\theta \mid y)d\theta$$

The *optimal* decision, which is called the Bayes rule, is the $d(y)$ that minimises the posterior expected loss.

Optimal Bayes estimators

Decision theory can be used to define optimal point and interval estimators, which are the optimal decisions.

Point estimators

- Under squared-error (quadratic) loss,

  $$L(\theta, d(y)) = (\theta - d(y))^2 \ ,$$

  the optimal (Bayes) estimator of θ is $d(y) = E(\theta \mid y)$, the posterior mean of θ.

- Under absolute error loss,

  $$L(\theta, d(y)) = |\theta - d(y)| \ ,$$

  the optimal (Bayes) estimator is $d(y) = \theta_{\text{med}}$, the posterior median of θ.

- Under zero-one loss,

  $$L(\theta, d(y)) = \begin{cases} 0 & \text{if } |\theta - d(y)| \leq \epsilon \\ 1 & \text{if } |\theta - d(y)| > \epsilon \end{cases}$$

  with ϵ very small, the optimal (Bayes) estimator is $d(y) = \theta_{\text{mode}}$, the posterior mode of θ.

Proof for the case under zero-one loss

$$\begin{aligned} & \int_{-\infty}^{\infty} L(\theta, d(y))p(\theta \mid y)d\theta \\ = & \int_{-\infty}^{d(y)-\epsilon} 1 \cdot p(\theta \mid y)d\theta + \int_{d(y)-\epsilon}^{d(y)+\epsilon} 0 \cdot p(\theta \mid y)d\theta \\ & + \int_{d(y)+\epsilon}^{\infty} 1 \cdot p(\theta \mid y)d\theta \\ = & P(\theta < d(y) - \epsilon \mid y) + P(\theta > d(y) + \epsilon \mid y) \\ = & 1 - P(d(y) - \epsilon \leq \theta \leq d(y) + \epsilon \mid y) \end{aligned}$$

This is minimised when
$P(d(y) - \epsilon \leq \theta \leq d(y) + \epsilon \mid y)$ is maximised.
When ϵ is very small, it is maximised when $d(y)$ is the posterior mode of θ.

4. Comparison of Bayesian and classical inferences

Point estimation

Many methods of point estimation have been proposed in classical statistics, the most important being *maximum likelihood*, *least squares* and *method of moments*. All are concerned with constructing functions of the data that are 'good' estimators of a parameter θ.

How to compare estimators?

Criteria used to compare classical (frequentist) estimators to decide which are 'good' ones include:

- *Unbiasedness*: Over repeated sampling, the mean/expectation of a parameter estimate should equal its true value.
- *Efficiency*: An efficient estimator is a function of data y that estimates θ with the lowest mean squared error (MSE). For an unbiased estimator, efficiency means attaining the minimum variance.
- *Consistency*: As the sample size increases, the estimator converges to the true value of the parameter.
- *Admissibility*: An estimator is said to be *inadmissible* if its MSE is uniformly higher than the MSE of another estimator. Otherwise, it is said to be *admissible*.

Comments

- Classical criteria are based on the properties of estimators under repeated sampling (of Y) $\Rightarrow$ all expectations are with respect to the sampling distribution of the data $p(y \mid \theta)$, implying unobserved data are also considered.

  In Bayesian inference, only the observed data y are considered, and θ is a random variable $\Rightarrow$ expectations are taken with respect to the posterior $p(\theta \mid y)$.
- Asymptotic arguments (ie as sample size $\to \infty$) are often used in classical inference. E.g. the MLE is always asymptotically unbiased, even though it may be biased for finite sample sizes.

  Bayesian inference is only concerned with estimating and summarising the posterior distribution $p(\theta \mid y)$. Does not (generally) bother with asymptotic arguments or classical criteria.

- Recall that $p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$. When $p(\theta) \propto k$ (non-informative), it follows that $p(\theta \mid y) \propto p(y \mid \theta)$ approximately (posterior $\propto$ likelihood). So, in this case, posterior mode $\approx$ MLE.
- Also,

  $$\log p(\theta \mid y) = \log p(y \mid \theta) + \log p(\theta) + \text{const}$$

  As the sample size $n \to \infty$, $\log p(\theta)$ becomes negligible compared to $\log p(y \mid \theta)$. So,

  $$\begin{aligned} \log p(\theta \mid y) &\approx \log p(y \mid \theta) + \text{const} \\ \Rightarrow p(\theta \mid y) &\propto p(y \mid \theta) \text{ approximately} \end{aligned}$$

  So, for a sufficiently large sample, posterior mode $\approx$ MLE.

Classical point estimator's mean squared error (with the estimator denoted by $T = T(Y)$):

$$E_{Y|\theta}(T-\theta)^2 = \text{var}_{Y|\theta}(T) + \{E_{Y|\theta}(T) - \theta\}^2$$

$$\begin{aligned} & E_{Y|\theta}(T-\theta)^2 \\ = \; & E_{Y|\theta}\{T - E_{Y|\theta}(T) + E_{Y|\theta}(T) - \theta\}^2 \\ = \; & E_{Y|\theta}\{T - E_{Y|\theta}(T)\}^2 + \{E_{Y|\theta}(T) - \theta\}^2 + \\ & 2E_{Y|\theta}\{T - E_{Y|\theta}(T)\}\{E_{Y|\theta}(T) - \theta\} \\ = \; & \text{var}_{Y|\theta}(T) + \{E_{Y|\theta}(T) - \theta\}^2 \; . \end{aligned}$$

Bayesian point estimator's posterior expected loss using squared-error loss function (with the estimator denoted by $t = T(y)$):

$$E_{\theta|y}(t-\theta)^2 = \text{var}_{\theta|y}(\theta) + \{E_{\theta|y}(\theta) - t\}^2$$

$$\begin{aligned} & E_{\theta|y}(t-\theta)^2 \\ = \; & E_{\theta|y}(\theta - t)^2 \\ = \; & E_{\theta|y}\{\theta - E_{\theta|y}(\theta) + E_{\theta|y}(\theta) - t\}^2 \\ = \; & E_{\theta|y}\{\theta - E_{\theta|y}(\theta)\}^2 + \{E_{\theta|y}(\theta) - t\}^2 + \\ & 2E_{\theta|y}\{\theta - E_{\theta|y}(\theta)\}\{E_{\theta|y}(\theta) - t\} \\ = \; & \text{var}_{\theta|y}(\theta) + \{E_{\theta|y}(\theta) - t\}^2 \; . \end{aligned}$$

What is an optimal Bayes estimator t of θ?

Interval estimation

- In Bayesian statistics, θ is a random variable, and Bayesian credible intervals are direct statements about the *probability* that θ lies within a fixed interval (given the data)

- In classical statistics, θ is fixed and the data are regarded as random, so a 100(1-α)% confidence interval is a *random* interval which contains the true *fixed* value of θ with probability 1-α.
  - In classical inference, confidence intervals are probability statements about the random interval, not about θ.
  - In classical inference, probability statements about θ are not possible because θ is not allowed to be a random variable.

In general, Bayesian and classical intervals can be quite different. However, Bayesian credible interval and classical confidence interval for the same problem can be very close.

Why is this?

In classical inference, let $\widehat{\theta}$ denote MLE of θ. For a large n, $\widehat{\theta}$ is approximately distributed Normal$(\theta, \text{SE}(\widehat{\theta})^2)$. The approximate 95% CI for θ is then

$$\widehat{\theta} \pm 1.96 \times \text{SE}(\widehat{\theta}) \ .$$

In Bayesina inference, when n is large (assuming the likelihood dominates the prior), $\theta|y$ is approximately distributed Normal$(\widehat{\theta}, \text{SE}(\widehat{\theta})^2)$ (see Gelman et al. chapter 4). Thus,

$$\widehat{\theta} \pm 1.96 \times \text{SE}(\widehat{\theta})$$

is also an approx 95% Bayesian credible interval for θ.

Examples

1. On p5, we noted that, for unknown mean θ and known precision τ when the sample size n is large, or the prior precision ϕ_0 is small,

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

   approximately.

   Therefore, a 95% Bayesian credible interval for θ is

   $$\bar{y} \pm \frac{1.96}{\sqrt{n\tau}} \ .$$

   This is the same as the classical CI.

2. On p8, we saw that, for known mean θ and unknown precision τ,
   if $\tau \sim \text{Gamma}(\alpha, \beta)$, then

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

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

Suppose we take improper prior $\tau \sim \text{Gamma}(0, 0)$, equivalent to $p(\tau) \propto \tau^{-1}$. Then

$$\begin{aligned} \tau \mid y \ &\sim \ \text{Gamma}\left(\frac{n}{2}, \ \frac{ns^2_{(n)}}{2}\right) \\ \Rightarrow ns^2_{(n)}\tau \mid y \ &\sim \ \text{Gamma}\left(\frac{n}{2}, \ \frac{1}{2}\right) \equiv \chi^2_n \ . \end{aligned}$$

So, a 95% Bayesian credible interval for τ is

$$\left(\frac{\chi^2_{n,0.025}}{ns^2_{(n)}}, \ \frac{\chi^2_{n,0.975}}{ns^2_{(n)}}\right) \ .$$

Subsequently, a 95% Bayesian credible interval for $\sigma^2 = \tau^{-1}$ is

$$\left(\frac{ns^2_{(n)}}{\chi^2_{n,0.975}}, \ \frac{ns^2_{(n)}}{\chi^2_{n,0.025}}\right) \ ,$$

which is the same as the classical CI (derived from $ns^2_{(n)}/\sigma^2 \sim \chi^2_n$).

The Likelihood Principle

An important philosophical difference between Bayesian and classical inference is that Bayesian inference obeys the *Likelihood Principle* while classical inference does not.

The Likelihood Principle states that the observed value of the likelihood function $p(y \mid \theta)$ summarises all the information provided by the data y about θ.

Bayesian inference obeys the LP since the data only affect the posterior via the observed likelihood: $p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$.

Classical (frequentist) inference violates the LP since inference about θ is based on the sampling distribution of Y which takes into account all possible (even unobserved) values of Y.

(Note: MLE obeys the LP).

Example 2.3: Likelihood Principle

You carry out an experiment to test whether a coin is fair, by repeatedly tossing it and recording whether it landed heads or not.

Let θ be the probability of heads.

Design 1: You decide to toss the coin 12 times, and record 9 heads and 3 tails. Here the random variable Y is the number of heads in $n = 12$ trials, giving a binomial likelihood

$$p_1(y \mid \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = \binom{12}{9}\theta^9(1-\theta)^3$$

Design 2: You decide to toss the coin until the third tail is recorded, and this occurs after 9 heads have been recorded. Here Y is the number of heads before $r = 3$ tails occur, which has a negative binomial likelihood

$$p_2(y \mid \theta) = \binom{y+r-1}{y}\theta^y(1-\theta)^r = \binom{11}{9}\theta^9(1-\theta)^3$$

Likelihoods p_1 and p_2 differ only by a constant of proportionality; so, in Bayesian inference, it will lead to the same posterior for θ, from which we could easily evaluate, say, $P(\theta > 0.5)$.

However, a classical test of the hypothesis H_0: $\theta = 0.5$ vs H_1: $\theta > 0.5$ gives p-values of:

$$P(Y \geq 9 \mid \theta = 0.5) = \sum_{i=9}^{12}\binom{12}{i}\theta^i(1-\theta)^{12-i} = 0.075$$

under the binomial likelihood, and

$$P(Y \geq 9 \mid \theta = 0.5) = \sum_{i=9}^{\infty}\binom{2+i}{i}\theta^i(1-\theta)^3 = 0.0325$$

under the negative binomial likelihood.

Summary: *The LP implies that it matters only what was observed, and not what might have been observed.* However, the frequentist approach depends not only on what was observed, but also on the design of the study (that is, in this example, how the experiment was stopped).

5. Advantages and disadvantages of Bayesian approaches

Some advantages

- Bayesian inference is theoretically sound and coherent (and often theoretically superior to classical inference)
- Bayesian inference follows directly from the posterior; no need to worry about statistical 'principles' such as bias, efficiency, consistency, asymptotics, and no need for separate theories of estimation, testing etc.
- Bayesian inference uses all the available information (with the ability to include prior knowledge)

Some disadvantages

- Inferences need to be justified to an outside world (reviewers, regulatory bodies, the public and so on).
  - Where did the prior come from?
- Although huge progress has been made, computational problems can still be considerable.

Outline revisited

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

Next week: Prior Distributions