STATS 2107
Statistical Modelling and Inference II
Lecture notes
Chapter 6: Bayesian statistics

Jono Tuke

School of Mathematical Sciences, University of Adelaide

Semester 2 2017

Introduction

Frequentist statistics

The purpose of statistical inference is to make conclusions about an unknown parameter, θ, on the basis of data $y_1, y_2, \ldots, y_n$, assumed to be observations from some distribution, $f(\boldsymbol{y}; \theta)$.

Some key concepts are

- Estimation,
- Hypothesis tests,
- Confidence intervals.

Bayesian statistics

In Bayesian inference, it is assumed that **prior** to observing the data, we have some knowledge of the parameter θ. This knowledge is expressed as a probability distribution, $(p(\theta)$.

The distribution $p(\theta)$ is called the **prior distribution**.

For example, if θ is the mean height of Australian adult males, measured in cm, what prior distributions could we consider?

Posterior distribution

The conditional distribution

$$p(\theta|\boldsymbol{y})$$

is called the **posterior distribution**.

Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

Bayes' theorem

$$p(\theta|\boldsymbol{y}) = \frac{p(\theta)p(\boldsymbol{y}|\theta)}{\int p(\theta)p(\boldsymbol{y}|\theta)d\theta}$$

Approximation

$$p(\theta|\boldsymbol{y}) \propto p(\theta)p(\boldsymbol{y}|\theta)$$

The fundamental principle of Bayesian inference

All conclusions about θ are made from the posterior distribution $p(\theta|\boldsymbol{y})$.

Example

Haemophilia is a X-chromosome linked, recessive disorder.

Suppose a woman has a haemophiliac brother, her father is normal, and her mother is a carrier.

Let

$$\theta = \begin{cases} 1 & \text{if the woman is a carrier} \\ 0 & \text{otherwise.} \end{cases}$$

It follows from genetic considerations that the prior distribution is

$$p(\theta) = \begin{cases} \frac{1}{2} & \text{if } \theta = 1, \\ \frac{1}{2} & \text{if } \theta = 0. \end{cases}$$

Suppose the woman has two sons, of which neither have haemophilia. Find the probability the woman as a carrier.

Bayesian prediction

Consider the prior distribution, $p(\theta)$, and data, $\boldsymbol{y}$, with likelihood $p(\boldsymbol{y}|\theta)$.

Suppose a new observation Y_0 is to be made.

The predictive distribution for Y_0 is just the conditional distribution,

$$p(y_0|\boldsymbol{y}).$$

If Y_0 and $\boldsymbol{Y}$ are conditionally independent given θ,

$$\begin{aligned}
p(y_0|\boldsymbol{y}) &= \int p(y_0, \theta|\boldsymbol{y}) d\theta \\
&= \int p(y_0|\theta, \boldsymbol{y}) p(\theta|\boldsymbol{y}) d\theta \\
&= \int p(y_0|\theta) p(\theta|\boldsymbol{y}) d\theta.
\end{aligned}$$

Example (continued)

Suppose the woman has a third son. Given that the first two sons are not haemophiliacs, what is the probability that the third son is not a haemophiliac?

Bayesian estimation

Bayesian point estimation

Suppose a point estimate of θ is required. Three possible quantities are

- **Posterior mode:**

$$\hat{\theta} = argmax_{\theta} p(\theta|\boldsymbol{y})$$

- **Posterior mean:**

$$E(\theta|\boldsymbol{y}) = \int \theta p(\theta|\boldsymbol{y}) d\theta$$

- **Posterior median:**

$$\tilde{\theta} \text{ such that } P(\theta \leq \tilde{\theta}|\boldsymbol{y}) = \frac{1}{2}$$

Note: When the posterior mean is used, the posterior variance $var(\theta|\boldsymbol{y})$ can be used as a measure of accuracy.

Bayesian credible interval

An interval ℓ, u is said to be a $100(1-\alpha)\%$ Bayesian credible interval if

$$P(\ell < \theta < u|\boldsymbol{y}) = 1 - \alpha.$$

Note this statement is a simple statement of (posterior) probability rather than the more complicated statement of confidence.

Example: Beta - Binomial

Suppose we wish to make inference about a binomial success probability θ.

The data are $Y|\theta \sim \mathit{Bin}(n, \theta)$.

Consider the prior distribution $\theta \sim \mathit{Beta}(\alpha, \beta)$.

Find the posterior distribution.

Beta distribution

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \text{ for } 0 < \theta < 1.$$

Estimates

- The posterior mode is

$$\hat{\theta} = \frac{\alpha + y - 1}{\alpha + \beta + n - 2}.$$

- The posterior mean is

$$E(\theta|x) = \frac{\alpha + y}{\alpha + \beta + n}.$$

Estimates

The posterior variance is

$$var(\theta|x) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n + 1)(\alpha + \beta + n)^2}.$$

Estimates

- The posterior median can be found numerically. For example, in R,

```
qbeta(0.5,alpha+x,beta+n-x)}.
```

- The Bayesian credible interval can be found numerically. For example, in R:

```
lower  <- qbeta(0.025,alpha+x,beta+n-x)}
upper  <- qbeta(0.975,alpha+x,beta+n-x)}.
```

Notes

- The posterior mean

$$\hat{\theta} = \frac{\alpha + y}{\alpha + \beta + n}$$

  can be seen to lie between the prior mean

$$\frac{\alpha}{\alpha + \beta}$$

  and the ordinary maximum likelihood estimator:

$$\frac{y}{n}$$

- When n is large relative to $\alpha + \beta$, the posterior mode will be close to y/n.
- When n is small relative to $\alpha + \beta$, the posterior mode will be close to $(\alpha - 1)/(\alpha + \beta - 2)$.

Example

In a study of premature births conducted at Johns Hopkins, it was found that of 39 babies, born at 25 weeks gestation, 31 survived for at least 6 months.

Let θ be the probability of survival.

Assuming an uninformative $U(0,1)$ prior for θ, calculate the posterior mode and posterior mean.

Conjugate priors

Definition

A family of prior distributions $\mathcal{P} = \{p(\theta)\}$ is said to be conjugate to a family of likelihoods, $\mathcal{L} = \{p(\boldsymbol{x}|\theta)\}$ if the posterior distribution always satisfies

$$p(\theta|\boldsymbol{x}) \in \mathcal{P}.$$

In the case of observations from a binomial distribution with a Beta prior, the posterior distribution is also a Beta distribution.

Hence the Beta distribution is a conjugate prior for the binomial distribution, since the posterior distribution is a Beta distribution.

Conjugate priors are mathematically convenient but there is no reason that a conjugate prior is scientifically justified.

Normal distribution

Normal data and conjugate prior

Suppose $y_1, y_2, \ldots, y_n$ are IID $N(\mu, \sigma^2)$ observations with σ^2 known.

Consider the prior, $\mu \sim N(\mu_0, \tau^2)$.

Show that the posterior distribution $p(\mu|\boldsymbol{y})$ is

$$\mu|\boldsymbol{y} \sim N\left(\frac{n\tau^2\bar{y} + \sigma^2\mu_0}{n\tau^2 + \sigma^2}, \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}\right).$$

Alternative derivation

If

- $\mu \sim N(\mu_0, \tau^2)$,
- $\bar{Y}|\mu \sim N(\mu, \sigma^2/n)$

it follows that the joint distribution of $(\mu, \bar{X})$ is

$$\begin{pmatrix} \mu \\ \bar{Y} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_0 \\ \mu_0 \end{pmatrix}, \begin{pmatrix} \tau^2 & \tau^2 \\ \tau^2 & \tau^2 + \sigma^2/n \end{pmatrix} \right)$$

Using the formula for the conditional distribution from a multivariate normal, we obtain

$$E(\mu|\bar{y}) = \mu_0 + \frac{\tau^2}{\tau^2 + \sigma^2/n}(\bar{y} - \mu_0) = \frac{n\tau^2\bar{y} + \sigma^2\mu_0}{n\tau^2 + \sigma^2}$$

and

$$var(\mu|\bar{y}) = \tau^2 - \frac{\tau^4}{\tau^2 + \sigma^2/n} = \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2},$$

and hence,

$$\mu|\boldsymbol{y} \sim N\left(\frac{n\tau^2\bar{y} + \sigma^2\mu_0}{n\tau^2 + \sigma^2}, \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}\right).$$

Trade-off between prior and data

The trade-off between prior information and data can be seen explicitly.

- If $n = 0$, the posterior distribution reduces to the prior, $N(\mu_0, \tau^2)$.
- If $\tau^2 \to \infty$, the posterior distribution becomes $N(\bar{y}, \sigma^2/n)$.
- If $\tau^2 \to 0$, the posterior distribution becomes $N(\mu_0, 0)$. So $\mu = \mu_0$ with probability 1.

Estimates as linear combinations

- The posterior mean can also be expressed as a linear combination of the prior mean μ_0 and the sample mean, $\bar{y}$,

$$E(\mu|\boldsymbol{y}) = w\mu_0 + (1-w)\bar{y} \text{ where } w = \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}.$$

- The posterior variance can also be expressed as a linear combination,

$$var(\mu|\boldsymbol{y}) = w^2\tau^2 + (1-w)^2\sigma^2/n.$$

- The prior distribution has a data equivalent interpretation:

The information in the prior specification $\mu \sim N(\mu_0, \tau^2)$ is roughly the same as what would be provided in a sample with sample mean $\bar{y} = \mu_0$ and standard error $se(\bar{y}) = \tau$.