Lecture 3.
Univariate Bayesian inference: conjugate analysis

Summary

1. Posterior predictive distributions
2. Conjugate analysis for proportions
3. Posterior predictions for proportions
4. Conjugate analysis for Normal
5. Conjugate analysis for Poisson
6. Mixtures of prior distributions

The Bayesian analyst (continuous parameters) needs to

- explicitly state a reasonable opinion concerning the plausibility of different values of the parameters *excluding* the evidence from the study (the **prior distribution**)
- provide the support for different values of the treatment effect based *solely* on data from the study (the **likelihood**),
- weight the likelihood from the study with the relative plausibilities defined by the prior distribution to produce
- a final opinion about the parameters (the **posterior distribution**)

$$p(\theta \mid y) = \frac{p(\theta)\, p(y \mid \theta)}{\int p(\theta)\, p(y \mid \theta)\, d\theta} \propto p(y \mid \theta)\, p(\theta)$$

when considering $p(y \mid \theta)$ as a function of θ: ie the *likelihood*.

posterior $\propto$ likelihood $\times$ prior.

Posterior predictive distributions

When we have observed some data y, the predictive distribution for a new observation $\tilde{y}$ is given by

$$p(\tilde{y}|y) \;=\; \int p(\tilde{y}|y,\theta)p(\theta|y)d\theta.$$

Assuming past and future observations are conditionally independent given θ, this simplifies to

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

The posterior-predictive expectation is

$$\boldsymbol{E}[\tilde{Y}|y] = \int \boldsymbol{E}[\tilde{y}|\theta]p(\theta|y)d\theta$$

Replace integration by summation for discrete parameters

Three coins: continued

We have observed a single head, changing the prior distribution (0.33, 0.33, 0.33) to a posterior distribution (0,17, 0.33, 0.50)

Suppose we want to predict probability that *next* toss is a head. Now

$$\begin{aligned} P(\tilde{Y}=1|y) &= \sum_i P(\tilde{Y}=1|\theta_i)p(\theta_i|y) = \sum_i \theta_i p(\theta_i|y) \\ &= (0.25\times 0.167)+(0.50\times 0.333)+(0.75\times 0.500) \\ &= 7/12 \end{aligned}$$

Inference on proportions using a continuous prior

Suppose we observe y positive responses out of n Bernoulli trials.

Binomial sampling distribution:

$$p(y|\theta,n) = \left(\begin{array}{c} n \\ y \end{array}\right)\theta^y(1-\theta)^{n-y} \propto \theta^y(1-\theta)^{n-y}$$

Suppose that, before taking account of this evidence, we believe all values for θ are equally likely (is this plausible?) $\Rightarrow \theta \sim$ Unif$(0,1)$ i.e. $p(\theta)=\frac{1}{1-0}=1$

Posterior is then

$$p(\theta|y,n) \propto \theta^y(1-\theta)^{n-y}\times 1$$

This has form of the *kernel* of a Beta$(y+1, n-y+1)$ distribution with mean $(y+1)/(n+2)$

To represent external evidence that some response rates are more plausible than others, it is mathematically convenient to use a Beta(a,b) prior distribution for θ

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Combining this with the binomial likelihood gives a posterior distribution

$$\begin{aligned} p(\theta \mid y,n) &\propto p(y \mid \theta,n)p(\theta) \\ &\propto \theta^y(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{y+a-1}(1-\theta)^{n-y+b-1} \\ &\propto \text{Beta}(y+a,\ n-y+b) \end{aligned}$$

$$\boldsymbol{E}(\theta|r,n) = (y+a)/(n+a+b) = w\frac{a}{a+b}+(1-w)\frac{y}{n}$$

where $w=(a+b)/(a+b+n)$; a weighted average of the prior mean and y/n, the standard maximum-likelihood estimator, where the weight w reflects the relative contribution of the prior 'effective sample size' $a+b$.

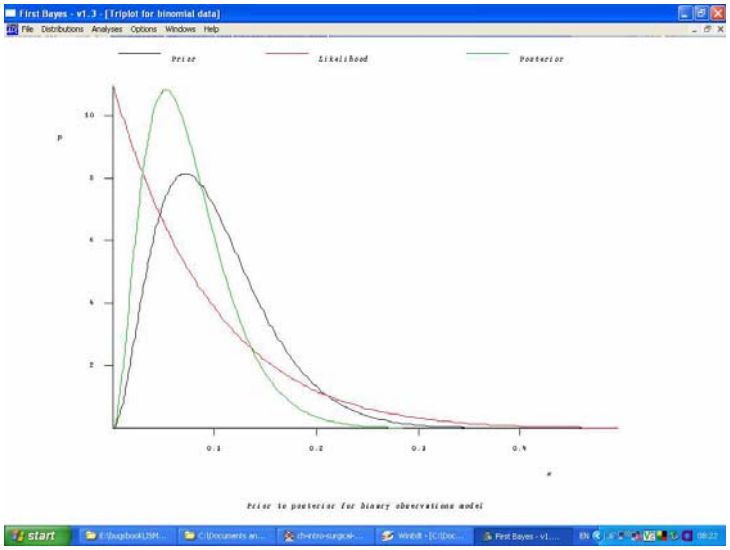
Hence the prior parameters a and b can be interpreted as equivalent to observing a events in $a+b$ trials

Surgery (continued): suppose we now operate on $n = 10$ patients and observe $y = 0$ deaths. What is the current posterior distribution

We used a Beta$(3, 27)$ as a prior distribution for a mortality rate. Plugging in the relevant values of $a = 3$, $b = 27$, $y = 0$ and $n = 10$ we obtain a posterior distribution for the mortality rate θ of $p(\theta|y, n) =$ Beta$(3, 37)$

Can use *First Bayes* : `www.firstbayes.co.uk/`

Written by Tony O'Hagan: not nice to install but fun to use!



Posterior-predictive distributions for Binomial data

Use Beta-binomial distribution, but now with parameters of posterior distribution:

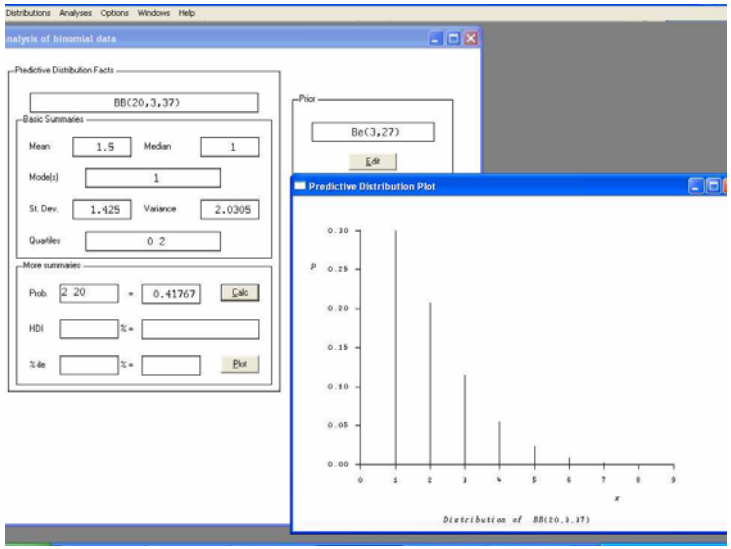
i.e. to get predictive distribution for $\tilde{Y}$ successes out of a further n',

$$p(\tilde{y}) = \frac{\Gamma(a' + b')}{\Gamma(a')\Gamma(b')} \binom{n'}{\tilde{y}} \frac{\Gamma(a' + \tilde{y})\Gamma(b' + n' - \tilde{y})}{\Gamma(a' + b' + n')}.$$

where $a' = a + y, b' = b + n$ are the revised parameters of the beta distribution after having observed y out of n successes.

Surgery (continued): after $n = 10$ patients and observe $y = 0$ deaths, what is the probability that the next patient will survive the operation, and what is the probability that there are 2 or more deaths in the next 20 operations?

The probability of a death at the next operation is simply $\boldsymbol{E}(\theta|y, n) = (y + a)/(n + a + b) = 3/40 = 0.075$. When considering the number $\tilde{y}$ of deaths in the next 20 operations, from the beta-binomial predictive distribution, we can calculate $P(\tilde{y} \geq 2) = 0.418$.



Suppose $y = n$, i.e. the event has happened at every opportunity! What is the chance it will happen next time?

The posterior-predictive expectation is

$$p(\tilde{Y} = 1|y) = \int \theta p(\theta|y) d\theta = \frac{n+1}{n+2}$$

Known as *Laplace's law of succession*

Assumes 'exchangeable events' (see below): ie the same (unknown) θ applies to each

Laplace originally applied to the problem of whether the sun will rise tomorrow. But he recognised the background knowledge should overwhelm simplistic assumptions. *"But this number [i.e., the probability that the sun will rise tomorrow] is far greater for him who, seeing in the totality of phenomena the principle regulating the days and seasons, realizes that nothing at the present moment can arrest the course of it."*

Normal data with unknown mean and known variance

Suppose we have an independent sample of data

$$y_i \sim \text{Normal}(\mu, \sigma^2), \quad i = 1 \ldots n$$

where σ^2 is known and μ is unknown. The conjugate prior for the Normal mean is also Normal

$$\mu \sim \text{Normal}(\gamma, \tau^2)$$

where γ and τ^2 are assumed specified.

It is convenient to write τ^2 as σ^2/n_0, where n_0 represents the 'effective number of observations' in the prior distribution.

Then the posterior distribution for μ is given by

$$\begin{aligned} p(\mu \mid \boldsymbol{y}) &\propto p(\mu) \prod_{i=1}^{n} p(y_i \mid \mu) \\ &\propto \exp\left[-\frac{1}{2}\left\{\frac{(\mu-\gamma)^2}{\sigma^2/n_0}\right\}\right] \exp\left[-\frac{1}{2}\left\{\frac{\sum(y_i-\mu)^2}{\sigma^2}\right\}\right] \end{aligned}$$

Straightforward to show posterior has the form of another Normal density

$$p(\mu \mid \boldsymbol{y}) = \text{Normal}(\gamma_n, \tau_n^2)$$

$$\text{where} \quad \gamma_n = \frac{n_0\gamma + n\overline{y}}{n_0 + n} \quad \text{and} \quad \tau_n^2 = \frac{\sigma^2}{n_0 + n}$$

Posterior precision = prior precision and data precision. Other equivalent expressions for the posterior mean:

$$\begin{aligned} \gamma_n &= w\gamma + (1-w)\overline{y} \quad \text{where } w = \frac{n_0}{n_0+n} \\ \gamma_n &= \gamma + (\overline{y}-\gamma)\frac{n}{n_0+n}; \\ \gamma_n &= \overline{y} - (\overline{y}-\gamma)\frac{n_0}{n_0+n}; \end{aligned}$$

Shows 'shrinkage' towards prior mean.

A sceptic's view:

'A Bayesian is one who, vaguely expecting a horse and catching a glimpse of a donkey, strongly concludes he has seen a mule'
(Senn, 1997)

The posterior predictive distribution can be shown to give

$$p(\tilde{y}|\boldsymbol{y}) \;=\; \text{Normal}\left(\gamma_n, \sigma^2 + \tau_n^2\right).$$

So the predictive distribution is centered at the posterior mean of μ with variance equal to the sum of the posterior variance of μ plus the data (residual) variance.

[Write $\tilde{y} = \mu + \epsilon$ where $\epsilon \sim \text{Normal}\left(0, \sigma^2\right)$, and so $\tilde{y}$ is the sum of two independent normal variables.]

Normal data with unknown variance and known mean

Suppose again $y_i \sim \text{Normal}(\mu, \sigma^2)$ but this time μ is assumed known and σ^2 is unknown.

Convenient to change parameterisation to the precision $\omega = 1/\sigma^2$: this is the reason BUGS uses the precision in the normal distribution.

The conjugate prior for ω is then

$$\omega \sim \text{Gamma}(\alpha, \beta),$$

so that

$$p(\omega) \propto \omega^{\alpha-1} \exp\{-\beta\omega\} :$$

σ^2 is then said to have an inverse-gamma distribution

The posterior distribution for ω takes the form

$$p(\omega|\mu, \boldsymbol{y}) \propto \omega^{\frac{n}{2}} \exp\left\{-\frac{\omega}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right\} \times \omega^{\alpha-1} \exp\{-\beta\omega\}.$$

Collecting terms reveals that

$$p(\omega|\mu, \boldsymbol{y}) \;=\; \text{Gamma}\left(\alpha + \frac{n}{2},\; \beta + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right).$$

- Natural to think of $\alpha = n_0/2$, where n_0 is the 'effective number of observations'
- Since $\sum_{i=1}^{n}(y_i - \mu)^2/n$ is an estimate of $\sigma^2 = 1/\omega$, then we interpret 2β as representing a $n_0\times$ a prior estimate $\hat{\sigma}_0^2$

Alternative, equivalent representations

- Can write our conjugate prior as

$$\omega \sim \text{Gamma}(n_0/2, n_0\hat{\sigma}_0^2/2).$$

- $\sigma^2 = 1/\omega$ therefore has an 'inverse-Gamma' distribution with parameters $a = n_0/2, b = n_0\hat{\sigma}_0^2/2$.
- For an inverse Gamma, $p(\sigma^2) \propto (\sigma^2)^{-(a+1)}e^{-b/\sigma^2}$.
- Gelman et al (2004), p50 point out that σ^2 has a distribution equivalent to that of $n_0\hat{\sigma}_0^2/X$, where X has a $\chi^2_{n_0}$ distribution
- They say that σ^2 has a 'scaled inverse-χ^2' distribution with notation

$$\sigma^2 \sim \text{Inv}{-}\chi^2(n_0, \hat{\sigma}_0^2).$$

- Useful when assessing prior distributions for sample variances

Poisson data

Suppose $Y_i \sim \text{Poisson}(\mu t_i)$: (unknown) μ is the rate per unit of t

$$p(\boldsymbol{y}|\mu) \;=\; \prod_i \frac{(\mu t_i)^{y_i} e^{-\mu t_i}}{y_i!}$$

The kernel of the Poisson likelihood (as a function of μ) has the same form as that of a Gamma(a, b) prior for μ

This implies the following posterior for μ

$$\begin{aligned} p(\mu \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \mu)\, p(\mu) \\ &\propto \prod_{i=1}^{n} \mu^{y_i} e^{-\mu t_i}\, \mu^{a-1} e^{-b\mu} \\ &\propto \mu^{a+y_s-1}\, e^{-(b+t_s)\mu} \\ &= \text{Gamma}(a + y_s,\, b + t_s). \end{aligned}$$

where $y_s = \sum_{i=1}^n y_i, t_s = \sum_{i=1}^n t_i$

$E(\mu \mid \boldsymbol{y}) = \frac{a+y_s}{b+t_s} = \frac{y_s}{t_s}\left(\frac{n}{n+b}\right) + \frac{a}{b}\left(1 - \frac{n}{n+b}\right)$ which we can again see is a compromise between the prior mean a/b and the MLE $\frac{y_s}{t_s}$

Thus b can be interpreted as an 'effective exposure', and a/b as a prior estimate of the Poisson mean

Mixtures of prior distributions

- Suppose we doubt which of two or more prior distributions is appropriate to the data in hand
- *eg* might suspect that *either* a drug will produce similar effect to other related compounds, *or* if it doesn't behave like these compounds we are unsure about its likely effect
- For two possible prior distributions $p_1(\theta)$ and $p_2(\theta)$ the overall prior distribution is then a *mixture*

$$p(\theta) = qp_1(\theta) + (1-q)p_2(\theta),$$

  where q is the assessed probability that p_1 is 'correct'.
- Consider the two priors as hypotheses H_1 and H_2, so that $q = p(H_1)$, and

$$p(\theta) = qp(\theta|H_1) + (1-q)p(\theta|H_2)$$

- If we now observe data y, the posterior for θ is

$$p(\theta|y) = q'p(\theta|y, H_1) + (1 - q')p(\theta|y, H_2)$$

where

$$\begin{aligned} p(\theta|y, H_i) &\propto p(y|\theta)p(\theta|H_i) \\ q' &= p(H_1|y) = \frac{qp(y|H_1)}{qp(y|H_1) + (1-q)p(y|H_2)}, \end{aligned}$$

where $p(y|H_i) = \int p(y|\theta)p(\theta|H_i)d\theta$ is the predictive probability of the data y assuming H_i

- The posterior is a mixture of the respective posterior distributions under each prior assumption, with the mixture weights adapted to support the prior that provides the best prediction for the observed data.