Bayesian Inference

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

Probability

Direct expression of uncertainty about unknown parameters.

- Tells us what we want: what are plausible values for the parameter of interest?
- No (difficult to interpret) confidence intervals: just report, say, central area that contains 95% of distribution
- 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 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 of Quantities of Interest (QOI), e.g. treatment effect, small area estimates, etc.
- point and interval estimates of any function of the parameters
- probability that QOI (e.g. treatment effect) exceeds a critical threshold
- prediction of QOI in a new unit
- 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, missing data, etc.

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

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

Bayesian inference (continued)

As with any analysis, we start by positing a model, $p(x|\theta)$. This is the **likelihood**, which relates all variables into a **full probability model**.

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 quantities of interest given the data:

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

  - This is the **posterior distribution**.

Bayesian inference (continued)

- 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.

I. Bayesian inference on proportions using binomial distribution

Bayesian inference on proportions

- Suppose we observe r positive responses out of n patients.
- Assuming patients are independent, with common unknown response rate θ, leads to a binomial likelihood:

$$p(r|n,\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r} \propto \theta^r(1-\theta)^{n-r}$$

- θ needs to be given a continuous prior distribution.
- Suppose that, before taking account of the evidence from our study, we believe all values for θ are equally likely. $\Rightarrow \theta \sim$ Unif(0, 1), i.e. $p(\theta) = \frac{1}{1-0} = 1$.
- Posterior is then $p(\theta|r,n) \propto \theta^r(1-\theta)^{n-r} \times 1$.
- This has form of the kernel of a Beta$(r+1,\ n-r+1)$ distribution, where

$$\text{Beta}(a,b) \equiv \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

Bayesian inference on proportions

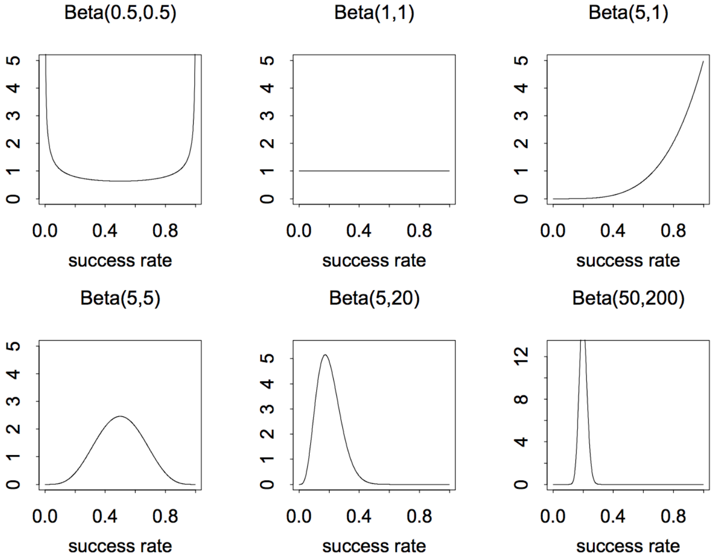
To represent external evidence that some response rates are more plausible than others, it is mathematically convenient to use a Beta(a, b) prior distribution for θ:

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

Combining this with the binomial likelihood gives a posterior distribution

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

Some Beta distributions



Beta(0.5,0.5)

Beta(1,1)

Beta(5,1)

Beta(5,5)

Beta(5,20)

Beta(50,200)

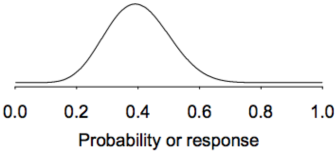
Comments

- When the prior and posterior come from the same family of distributions the prior is said to be conjugate to the likelihood.
  - Occurs when prior and likelihood have the same 'kernel'.
- Beta(a, b) distribution has mean $= a/(a+b)$ and variance $= ab/[(a+b)^2(a+b+1)]$. Hence posterior mean is $E(\theta|r,n) = (r+a)/(n+a+b)$.
- a and b are equivalent to observing a priori $a-1$ successes in $a+b-2$ trials.
- With fixed a and b, as r and n increase, $E(\theta|r,n) \rightarrow r/n$ (the MLE), and the variance tends to zero.
  - This is a general phenomenon: as n increases, posterior distribution gets more concentrated and the likelihood dominates the prior.
- A Beta(1, 1) is equivalent to Uniform(0, 1).

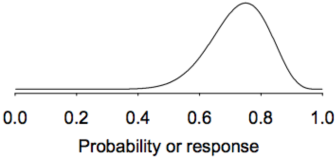
Example: Drug

- Consider a new drug to be given for relief of chronic pain. Experience with similar compounds has suggested that response rates between 0.2 and 0.6 could be feasible. We interpreted this as a distribution with mean = 0.4, standard deviation 0.1 and showed that a Beta(9.2,13.8) distribution has these properties.
- Suppose we now treat n = 20 volunteers with the compound and observe y = 15 positive responses.

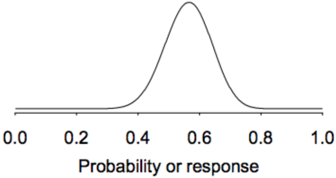
Prior, likelihood and posterior for Drug example



Beta(9.2, 13.8) prior distribution supporting response rates between 0.2 and 0.6



Likelihood arising from a Binomial observation of 15 successes out of 20 cases



Parameters of the Beta distribution are updated to (a+15, b+20-15) = (24.2, 18.8): mean 24.2/(24.2+18.8) = 0.56

Drug (continued): learning about parameters from data using Markov chain Monte-Carlo (MCMC) methods

- In the Drug example so far, we have calculated the posterior distributions in closed form.
  - this is possible because we are using conjugate priors
  - means that we can make use of known properties of the closed-form posterior distribution to make inference, e.g. expected value (posterior mean), tail-area probabilities are known analytically.
- Using MCMC, no need to explicitly specify posterior.
- Can just specify the prior and likelihood separately.
- There are algorithms to evaluate (and summarise) the posterior given (almost) arbitrary specification of prior and likelihood.
  - posterior does not need to be closed form.

II. Bayesian inference using the Normal distribution

Bayesian inference using the Normal distribution

Known variance, unknown mean

- Suppose we have a sample of Normal data $x_i \sim \mathsf{N}(\theta, \sigma^2)$ $i = 1, \ldots, n$.
- For now assume σ^2 is known and θ has a Normal prior $\theta \sim \mathsf{N}(\mu, \sigma^2/n_0)$.
  - 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

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

Bayesian inference using the Normal distribution

- 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 -∞, ∞.
- Posterior mean $(n_0\mu + n\bar{x})/(n_0 + n)$ is a weighted average of the prior mean μ and parameter estimate $\bar{x}$, weighted by their precisions (relative 'sample sizes'), and so is always a compromise between the two.
- 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.
- As $n \to \infty$, $p(\theta|\mathbf{x}) \to \mathsf{N}(\bar{x}, \sigma^2/n)$ which does not depend on the prior.
- Compare with frequentist setting, the MLE is $\hat{\theta} = \bar{x}$ with SE($\hat{\theta}$) $= \sigma/\sqrt{n}$, and sampling distribution

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

Example: THM concentrations

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 average THM concentration in a particular water zone, z. Two independent measurements, x_{z1} and x_{z2} 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.

What should we estimate the mean THM concentration to be in this water zone?

THM example (continued)

Let the mean THM concentration be denoted θ_z. Standard analysis would use sample mean $\bar{x}_z = 130\mu$g/l as an estimate of θ_z, with standard error $\sigma/\sqrt{n} = 5/\sqrt{2} =$ 3.5μg/l.

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

THM example (continued)

Suppose historical data on THM levels in other zones supplied from the same source showed that the mean THM concentration was 120 μg/l with standard deviation 10 μg/l.

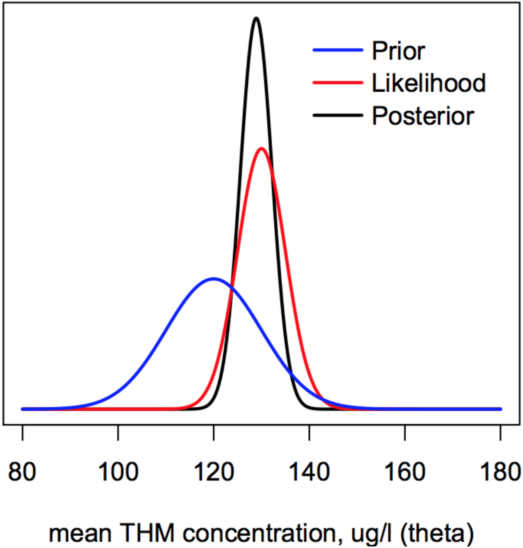
- suggests Normal(120, 10^2) prior for θ_z.
- if we express the prior standard deviation as $\sigma/\sqrt{n_0}$, we can solve to find $n_0 = (\sigma/10)^2 = 0.25$.
- so our prior can be written as $\theta_z \sim$ Normal$(120, \sigma^2/0.25)$.

Posterior for θ_z is then

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

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

Prior, likelihood and posterior for THM example



Prediction

Denoting the posterior mean and variance as $\mu_n = (n_0\mu + n\bar{x})/(n_0 + n)$ and $\sigma_n^2 = \sigma^2/(n_0 + n)$, the **predictive distribution** for a new observation $\tilde{x}$ is

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

which generally simplifies to

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

which can be shown to give

$$p(\tilde{x}|\mathbf{x}) \sim \mathsf{N}(\mu_n, \sigma_n^2 + \sigma^2).$$

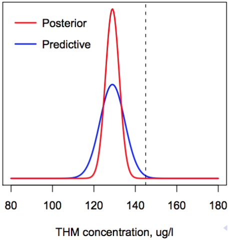
So the predictive distribution is centred around the posterior mean with variance equal to sum of the posterior variance and the sample variance of $\tilde{x}$.

Prediction with THM example

- Suppose the water company will be fined if THM levels in the water supply exceed 145 μg/l.
- Predictive distribution for THM concentration in a future sample taken from the water zone is

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

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



III. Bayesian inference using count data

Bayesian inference using count data

Suppose we have an independent sample of counts $x_1, \ldots, x_n$ which can be assumed to follow a Poisson distribution with unknown mean μ:

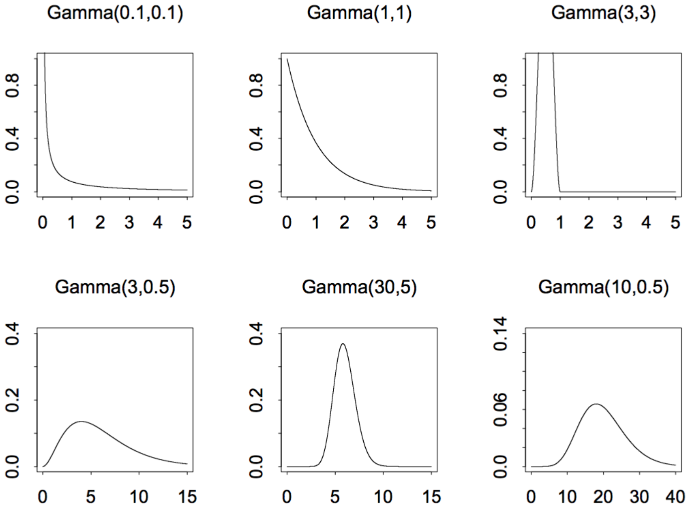
$$p(\mathbf{x}|\mu) = \prod_i \frac{\mu^{x_i} e^{-\mu}}{x_i!}$$

The conjugate prior for the mean of a Poisson distribution is a Gamma distribution:

$$p(\mu) = \text{Gamma}(a, b) = \frac{b^a}{\Gamma(a)} \mu^{a-1} e^{-b\mu}.$$

Gamma(a, b) density has mean a/b and variance a/b^2.

Some Gamma distributions



Bayesian inference using count data (continued)

This implies the following posterior

$$\begin{aligned} p(\mu|\mathbf{x}) &\propto p(\mu)p(\mathbf{x}|\mu) \\ &= \frac{b^a}{\Gamma(a)}\mu^{a-1}e^{-b\mu}\prod_i e^{-\mu}\frac{\mu^{x_i}}{x_i!} \\ &\propto \mu^{a+n\bar{x}-1}e^{-(b+n)\mu} \\ &= \text{Gamma}(a+n\bar{x}, b+n). \end{aligned}$$

The posterior is another (different) Gamma distribution.

$$E(\mu|\mathbf{x}) = \frac{a+n\bar{x}}{b+n} = \bar{x}\left(\frac{n}{n+b}\right) + \frac{a}{b}\left(1-\frac{n}{n+b}\right).$$

So posterior mean is a compromise between the prior mean a/b and the MLE $\bar{x}$.

Example: Estimation of disease risk in a single area

Often interested in estimating the rate or relative risk rather than the mean for Poisson data.

Suppose we observe $x = 5$ cases of leukaemia in one region, with age-sex-standardised expected number of cases $E = 2.8$.

Assume Poisson likelihood for x with mean $\mu = \lambda \times E$, where λ is the unknown relative risk:

$$p(x|\lambda, E) = \frac{(\lambda E)^x e^{-\lambda E}}{x!}.$$

Disease risk example (continued)

Assume Gamma(a, b) prior for the relative risk λ:

$$p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}.$$

Posterior for λ is then

$$\begin{aligned} p(\lambda | x, E) &\propto \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \frac{(\lambda E)^x e^{-\lambda E}}{x!} \\ &\propto \lambda^{a+x-1} e^{-(b+E)\lambda} \propto \text{Gamma}(a + x, b + E). \end{aligned}$$

Disease risk example: Vague prior

Suppose we wish to express vague prior information about λ.

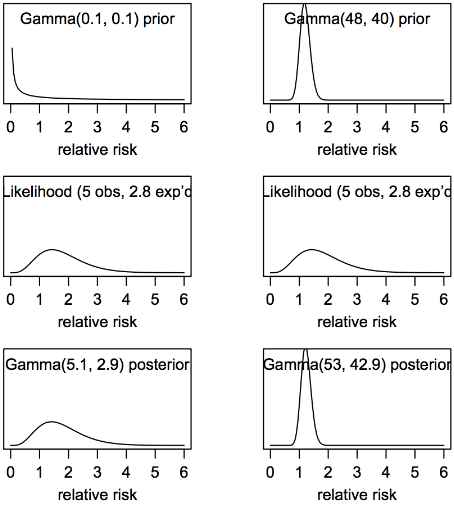
- A Gamma(0.1, 0.1) distribution represents a prior for the relative risk λ with
  - mean 0.1/0.1 $=$ 1
  - variance $0.1/0.1^2 =$ 10
  - 95^{th} percentile $=$ 5.8
- This gives a posterior $p(\lambda|x) =$ Gamma(5.1, 2.9).
- This has posterior mean $=$ 5.1/2.9 $=$ 1.76 (cf MLE $= x/E =$ 5/2.8 $=$ 1.78)

Disease risk example: Informative prior

Alternatively, we may have strong prior information to suggest that the relative risk in the region is probably around 1.2, and has only a 5% probability of being higher than 1.5.

- A Gamma(48, 40) distribution represents a prior for the relative risk λ with
  - mean $48/40 = 1.2$
  - 95^{th} percentile $= 1.5$
- This gives a posterior $p(\lambda|x) =$ Gamma(53, 42.8).
- This has posterior mean $= 53/42.9 = 1.24$

Prior, likelihood and posterior for disease risk example: Vague (left) and Informative (right) priors



Comments

For all these examples, we see that

- the posterior mean is a compromise between the prior mean and the MLE
- the posterior s.d. is less than each of the prior s.d. and the s.e.(MLE)

‘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) As $n \to \infty$,

- the posterior mean → the MLE
- the posterior s.d. → the s.e.(MLE)
- the posterior does not depend on the prior.

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

Choosing prior distributions

When the posterior is in the same family as the prior then we have what is known as conjugacy. This has the advantage that prior parameters can usually be interpreted as a prior sample. Examples include:

Likelihood	Parameter	Prior	Posterior
Normal	mean	Normal	Normal
Normal	precision	Gamma	Gamma
Binomial	success prob.	Beta	Beta
Poisson	rate or mean	Gamma	Gamma

- Conjugate prior distributions are mathematically convenient, but do not exist for all likelihoods, and can be restrictive.
- Computations for non-conjugate priors are harder, but possible using MCMC.