Statistics 3858b : Bayesian Methods

So far in our course we have viewed parameters as given numbers in a parameter space and the distribution of the observable random variables as coming from a distribution $f(\cdot;\theta)$ for one fixed value of θ. This makes sense in most types of experiments and observational studies.

Another approach to estimation is the so called Bayesian method or approach. In this we view Θ as a random variable on the parameter space; note the change in notation so we will need a new name for the parameter space when it is needed. In this setting we treat the conditional distribution of X given $\Theta = \theta$ as $f(\cdot|\theta)$. When $X_1, \ldots, X_n$ are conditionally iid, given $\Theta = \theta$ this conditional distribution is

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f_{X|\Theta}(x_i|\theta)$$

where $f_{X|\Theta}$ is interpreted as the conditional pdf or pmf of X_i given $\Theta = \theta$.

There is a prior or initial distribution of Θ which we write as f_Θ. From this and the rule of total probability we can calculate the marginal distribution of $X_1, \ldots, X_n$ as either

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \int f_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\theta) f_\Theta(\theta) d\theta$$

in the case of continuous r.v.s or

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \sum f_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\theta) f_\Theta(\theta)$$

in the case of discrete r.v.s. The corresponding integral or sum is over the set of possible values of the r.v. Θ. We can also calculate the joint distribution of $X_1, \ldots, X_n, \Theta$, written as

$$f_{X_1,\ldots,X_n,\Theta}(x_1, \ldots, x_n, \theta)$$

Sometimes we need to mix discrete X with continuous Θ or the other way round. Even though our course has not covered this the analogous and natural formula will hold. This means for the first formula we integrate over θ even if the r.v.s $X_1, \ldots, X_n$ are discrete, so the LHS is a pmf or pdf accordingly. For the second formula we sum over the discrete support of θ, and the resulting LHS is a pmf (if X's are discrete) or pdf (if the X's are continuous).

Can this type of statistical model make sense? One data type, of many types of natural data types, is the following. Different counties or regions of the province have local environmental variation. For a given region let X be the lifetime of a randomly chosen person (animal, insect ...). If we let Θ be the random *environment* then X has lifetime distribution, conditional on $\Theta = \theta$ given by $f(\cdot|\theta)$. We can then take iid observations from this region.

In some animal (usually mice or rates) experiments, an inbreed line of mice have a genetic characteristic (eg resistance to disease) determined by a random variable Θ. For randomly chosen mice from this genetic line the r.v. X (lifetime or disease effect) conditional on $\Theta = \theta$ has a distribution $f(\cdot|\theta)$. Different lines of mice will have a different *enivornmental or genetic* effect.

Another setting for which this Bayesian method is appropriate is the following. For a binomial experimental setting, one observes r.v.s $Y_i, \ldots, X_n$ iid Bernoulli, θ, or $Y = \sum_{i=1}^n X_i \sim \text{Binomial}(n, \theta)$. If n is not large, there is a large change that $\hat{\theta} = \frac{Y}{n}$ equals 0 or 1. If we consder estimation of θ using for example MLE the estimate $\hat{\theta}$ (observed value of the MLE) is then 0 or 1. Is this estimate good. Yes, but is it physically meaningful or believable? Based on many previous similar experiments this may not be a reasonable result, in the sense that it may be physically impossible for θ to be 0 or 1. It is physically more believable that $\theta \in (0, 1)$. One way to deal with this is to use a Bayesian framework and a prior distribution on a r.v. Θ.

We caan just view the Bayes estimators as another estimator, usually different from either the method of moments estimator or the MLE.

In this Bayesian framework, we have r.v.s $X_1, \ldots, X_n, \Theta$. We have a *statistical model* giving the conditional distribution of $X_1, \ldots, X_n$ given $\Theta = \theta$, and *prior* distribution on Θ. We can then obtain the marginal distribution of $X_1, \ldots, X_n$ and the conditional distribution of Θ given $X_1 = x_1, \ldots, X_n = x_n$.

We can also calculate, using Bayes Theorem, the conditional distribution of Θ given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ as

$$\begin{aligned} f_{\Theta|X}(\theta|x_1, \ldots, x_n) &= \frac{f_{X_1,\ldots,X_n,\Theta}(x_1, \ldots, x_n, \theta)}{f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)} \\ &\propto f_{X_1,\ldots,X_n,\Theta}(x_1, \ldots, x_n, \theta) \end{aligned}$$

This is called the *posterior* distribution of Θ given the data $X_1 = x_1, \ldots, X_n = x_n$. In the first line of the formula above for $f_{\Theta|X}(\theta|x_1, \ldots, x_n)$ the denominator is also the number (depending of the data that we are conditioning upon) so this formula or function integrates (or sums) to 1. It is the *normalizing constant* to make this a pdf or pmf. Recall this notion from earlier in the previous semester. This is sometimes useful if for example we recognize the function in the numerator as the *kernel* of a known distribution. The kernel is the pdf or pmf except for the normalizing constant. When this happens and *we recognize it as a kernel* we then may *know* the normalizing constant without having to go through

the integration.

To elaborate on this point a little consider the function

$$cf_{X_1,\ldots,X_n,\Theta}(x_1,\ldots,x_n,\theta)$$

where we treat θ are the argument and consider $(x_1,\ldots,x_n)$ as depending on the given numbers $x_1,\ldots,x_n$. Notice we *effectively* treating $x_1,\ldots,x_n$ as *parameters* for this function with argument θ. We are then looking for the number c, which depends on the values of the parameters, in this case $(x_1,\ldots,x_n)$ so that $c = c(x_1,\ldots,x_n)$, and this function integrates (or sums) to 1. We used this *property* when constructing examples of pdf's. We also used this in obtaining moments and the MGF for Gamma distributions, moments for the Beta distribution and also for obtaining the conditional distributions for bivariate normals.

There is a question of *how to choose* the prior distribution. For our purposes we will take this as given in a problem and postpone that question till later. In some cases, including the only ones we consider here, there is a family of prior distributions on the parameter Θ, called the *conjugate* prior which makes the calculation of the posterior relatively easy, in the sense that the normalizing constant is easy to determine as mentioned above. In an numerical example later we will consider a non-conjugate prior, but will also see there is not easy algebraic way to obtain the *normalizing constant*.

Conjugate priors are related to the conditional joint distribution of $X_1,\ldots,X_n$ given $\Theta = \theta$, in that the prior and posterior are of the same *family* of distributions. This means that the calculation of the posterior is done by simply finding the *update* rule for the parameters for this family of distributions.

Examples with Conjugate Priors

Poisson Example :

X_i, given $\Lambda = \lambda$ are iid Poisson, λ.

Suppose that Λ has Gamma(α,ν) distribution, that is

$$f_\Lambda(\lambda) = \frac{\nu^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\nu\lambda}\mathrm{I}(\lambda > 0)\ . \tag{1}$$

The values α,ν are specified, for example $(1,\frac{1}{2})$, so that Λ has a specific distribution, in this case with mean 2, and variance 4. If Λ we "known" more precisely with mean 2 we might choose $(\alpha,\nu) = (3,\frac{3}{2})$, so the mean variance of Λ are 2 and $3*(4/3)^2 = \frac{4}{3}$.

The marginal distribution of $X_1,\ldots,X_n$ is

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) \quad = \quad P(X_1 = x_1,\ldots,X_n = x_n)$$

$$
\begin{aligned}
&= \int_0^\infty \prod_{i=1}^n \left\{ \frac{\lambda^{x_i}}{x!} e^{-\lambda} \right\} \frac{\nu^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\nu\lambda} d\lambda \\
&= \frac{\nu^\alpha}{\Gamma(\alpha) \prod_{i=1}^n x_i!} \int_0^\infty \lambda^{\alpha+\sum_{i=1}^n x_i - 1} e^{-(\nu+n)\lambda} d\lambda
\end{aligned}
$$

This can be simplified, since the integrand is actually a Gamma pdf except for the normalizing constant.

Given data $X_i = x_i$, $i = 1, \ldots, n$ the posterior distribution of Λ is given by

$$
\begin{aligned}
f_{\Lambda|X}(\lambda|x_1, \ldots, x_n) &\propto f_{X_1,\ldots,X_n,\Lambda}(x_1, \ldots, x_n, \lambda) \\
&\propto \prod_{i=1}^n \left\{ \frac{\lambda^{x_i}}{x!} e^{-\lambda} \right\} \frac{\nu^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\nu\lambda} \\
&\propto \lambda^{\alpha+\sum_{i=1}^n x_i - 1} e^{-(\nu+n)\lambda}
\end{aligned}
$$

Notice this last expression has argument λ, and since the support is $\lambda > 0$, it is the kernel of a Gamma distribution with *parameters*

$$
\alpha' = \alpha + \sum_{i=1}^n x_i \ , \ \nu' = \nu + n \ . \tag{2}
$$

Thus without doing the integration to find the marginal distribution of $X_1, \ldots, X_n$ we know the posterior distribution of Λ given data $X_1 = x_1, \ldots, X_n = x_n$ is Gamma with parameters given by (2). That means that the posterior of Λ is

$$
\begin{aligned}
&f_\Lambda(\lambda|X_1 = x_1, \ldots, X_n = x_n) \\
&= \frac{(\nu')^{(\alpha'}}{\Gamma(\alpha')} \lambda^{\alpha'-1} e^{-\nu'\lambda} \mathrm{I}(\lambda > 0) \\
&= \frac{(\nu+n)^{(\alpha+t(\underline{x}))}}{\Gamma((\alpha + t(\underline{x})))} \lambda^{(\alpha+t(\underline{x}))-1} e^{-(\nu+n)\lambda} \mathrm{I}(\lambda > 0)
\end{aligned}
$$

where $t(\underline{x}) = \sum_{i=1}^n x_i$.

Aside : If we change the prior (1) to anything else we would not have a conjugate prior. If we *knew* for certain that $0 < \lambda < 100$ and used the prior

$$
f_\Lambda(\lambda) = \frac{\nu^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\nu\lambda} \mathrm{I}(0 < \lambda < 100)
$$

then normalizing constant to obtain the posterior is no longer so simple, as it would now involve an incomplete Gamma function instead of the Gamma function.

If we used a prior with support $(, \infty)$, such as the distribution of the absolute value of a normal, then we would not have a conjugate prior.

Normal Conjugate Prior

For normal we use conjugate prior : μ , σ^2 independent Normal and Gamma distributions

Aside : Ξ and ξ are the Greek letters capital lower case of *xi*.

Example : Normal, mean =0, variance = σ^2. For a more convenient notation we write $\xi = \frac{1}{\sigma^2}$

Prior for Ξ : Gamma(α, λ)

Data $X_1, \ldots, X_n$ conditionally upon $\Xi = \xi$ are iid $N(0, \frac{1}{\xi})$

Posterior

$$\begin{aligned} f_{\Xi|X_1,\ldots,X_n}(\xi|x_1,\ldots,x_n) &\propto \xi^{n/2} e^{-\frac{\xi}{2}\sum_{i=1}^n x_i^2} \xi^{\alpha-1} e^{-\lambda\xi} \\ &\propto \xi^{\alpha+\frac{n}{2}-1} e^{-\left(\lambda+\frac{1}{2}\sum_{i=1}^n x_i^2\right)\xi} \end{aligned}$$

We can recognize this is the kernel for a Gamma distribution with parameters

$$\alpha' = \alpha + \frac{n}{2} \text{ and } \lambda' = \lambda + \frac{1}{2}\sum_{i=1}^n x_i^2$$

and thus

$$f_{\Xi|X_1,\ldots,X_n}(\xi|x_1,\ldots,x_) = \frac{\left(\lambda + \frac{1}{2}\sum_{i=1}^n x_i^2\right)^{\alpha+\frac{n}{2}}}{\Gamma(\alpha+\frac{n}{2})} \xi^{\alpha+\frac{n}{2}-1} e^{-\left(\lambda+\frac{1}{2}\sum_{i=1}^n x_i^2\right)\xi}$$

Bernoulli or Binomial with Conjugate Prior

For Binomial we use conjugate prior : Θ having a Beta distribution

$X|\Theta = \theta \sim$ Binomial(m, θ)

pmf is

$$f(x|\theta) = \binom{m}{x}(1-\theta)^{m-x}\theta^x$$

If we use a Beta(α, β) prior the posterior is

$$\begin{aligned} f_{\text{Posterior}}(\theta|X=x) &\propto \binom{m}{x}(1-\theta)^{m-x}\theta^x(1-\theta)^{\alpha-1}\theta^{\beta-1} \\ &\propto (1-\theta)^{\alpha+m-x-1}\theta^{\beta+x-1} \\ &\propto (1-\theta)^{\alpha'-1}\theta^{\beta'-1} \end{aligned}$$

where

$$\alpha' = \alpha + m - x \ , \ \beta' = \beta + x \ .$$

Since $0 < \theta < 1$ we see this is the kernel of a Beta distribution with parameters α', β'. Thus (see AppendixA3 or formula sheet given with the exams)

$$f_{\text{Posterior}}(\theta|X = x) = \frac{\Gamma(\alpha' + \beta')}{\Gamma(\alpha')\Gamma(\beta')}(1-\theta)^{\alpha'-1}\theta^{\beta'-1} \ .$$

Bayes Estimator and Bayes Estimate

In our Bayesian calculations we obtain the *posterior* distribution

$$f_{\Theta|X_1=x_1,\ldots,X_n=x_n}(\theta) = f_{\Theta|X_1,\ldots,X_n}(\theta|x_1,\ldots,x_n) = f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \ .$$

The Bayes estimator is

$$\hat{\Theta} = \mathrm{E}(\Theta|\mathbf{X}) = \mathrm{E}(\Theta|X_1,\ldots,X_n)$$

which is the conditional expectation of Θ conditional on $\mathbf{X} = (X_1,\ldots,X_n)$. It is sometimes referred to as the posterior mean, that is the Bayes estimator is the posterior mean.

The Bayes estimate is the observed value of this r.v. with the obsered data $x_1,\ldots,x_n$. Notice we will in general have to calculate this integral (or sum). It is

$$\hat{\Theta} = \mathrm{E}(\Theta|\mathbf{X}) = \mathrm{E}(\Theta|X_1 = x_1,\ldots,X_n = x_n) \ .$$

Normal with Conjugate Prior continued

The posterior is a Gamma distribution, with parameters

$$\alpha' = \alpha + \frac{n}{2} \ \text{ and } \lambda' = \lambda + \frac{1}{2}\sum_{i=1}^{n} x_i^2 \ .$$

We of course have earlier calculated the mean of a Gamma distribution and using this can easily obtain the Beayes estimator or Bayes estimate.

The Bayes estimate or posterior mean is therefore

$$\hat{\Xi}_{Bayes} = \frac{\alpha + \frac{n}{2}}{\lambda + \frac{1}{2}\sum_{i=1}^{n} x_i^2} = \frac{\frac{2\alpha}{n} + 1}{\frac{2\lambda}{n} + \frac{1}{n}\sum_{i=1}^{n} x_i^2} \ .$$

How does this compare with the MLE, which we studied earlier? If we consider the corresponding MLE in our classical or frequentist method, we have

$$\hat{\Xi}_{MLE} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \frac{1}{\hat{\mu}_{2,n}}$$

It then follows that

$$\hat{\Xi}_{MLE} - \hat{\Xi}_{Bayes} = -\frac{1}{n}\frac{2\alpha}{\frac{2\lambda}{n} + \hat{\mu}_{2,n}}$$

Thus the two estimators are the same up to a difference of order $\frac{1}{n}$.

Return to the Poisson example

See also R code for this example

Conjugate Prior Gamma : $\alpha = 2$ and $\nu = \frac{\alpha}{2}$, so that $\mathrm{E}_{Prior}(\Lambda) = 2$.

Plot this pdf

Now suppose that the "observed" or given value of Λ is $\lambda_{given} = 3$. The posterior of Λ is Gamma with parameters given by (2), that is

$$\alpha' = \alpha + \sum_{i=1}^{n} x_i \ , \ \nu' = \nu + n \ .$$

The Bayes estimate is

$$\mathrm{E}_{Posterior}(\Lambda) = \frac{\alpha'}{\nu'} = \frac{\frac{\alpha}{n} + \bar{x}_n}{\frac{\nu}{n} + 1}$$

Take a sample of size $n = 2$. Plot the posterior pdf, and see it is similar to the prior.

Now consider the same experiment with a sample of size $n = 50$. Now the posterior is quite different from the prior and is centred much closer to $\lambda_{given} = 3$. Here the posterior is a Gamma with shape parameter $\alpha' = \alpha + \sum_{i=1}^{n} x_i$; recall it is a conditional distribution, conditioned upon the given data $X_1 = x_1, \ldots, X_n = x_n$.

Aside : In the fall semester we used moment generating functions to study the distribution of $Y \sim \mathrm{Gamma}(\alpha', \nu')$, specifically a limit distribution of the normalized sequence from Y. Using MGF, and properties from the LLN so that conditional upon $\lambda_{given} = 3$, we will obtain as an approximation for the posterior distribution

$$\sqrt{n}\frac{(\Lambda - \lambda_{given})}{\sqrt{\lambda_{given}}} \to N(0,1) \text{ in distribution as } n \to \infty$$

This tells us the posterior distribution of the r.v. Λ is approximately normal and centred at the unknown given value $\lambda_{given} = 3$ at the beginning of the experiment. This is a conditional Central Limit Theorem, and is somewhat different form the CLT studied in the previous term.

We can use the posterior distribution to give a *prediction* interval of the random variable Λ. This is the standard prediction interval idea from the fall semester or introductory statistics courses. This is a different notion than *confidence intervals* which are a consistency of parameter values with respect to the observed data.

If Θ has posterior distribution $f_{\Theta|\underline{X}_n=\underline{x}}$ then we may find the central $1-\alpha$ region, that is find $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles, say $q_{\frac{\alpha}{2}}$ and $q_{1-\frac{\alpha}{2}}$, for this posterior distribution. The $1-\alpha$ prediction interval for Θ is

then given by

$$q_{\frac{\alpha}{2}} \leq \Theta \leq q_{1-\frac{\alpha}{2}}$$

For some posterior distributions it may be easier to work $\Theta - \mathrm{E}_{\mathrm{Posterior}}(\Theta)$ and so we would find quantiles for this centred distribution. Interpreting $q_{\frac{\alpha}{2}}$ and $q_{1-\frac{\alpha}{2}}$ accordingly then we obtain our prediction interval from

$$q_{\frac{\alpha}{2}} \leq \Theta - \mathrm{E}_{\mathrm{Posterior}}(\Theta) \leq q_{1-\frac{\alpha}{2}}$$

or equivalently

$$\mathrm{E}_{\mathrm{Posterior}}(\Theta) + q_{\frac{\alpha}{2}} \leq \Theta \leq \mathrm{E}_{\mathrm{Posterior}}(\Theta) + q_{1-\frac{\alpha}{2}} \ .$$