Lecture 3 - Key

The binomial model

After suspicious performance in the weekly soccer match, 37 mathematical sciences students, staff, and faculty were tested for performance enhancing drugs. Let $Y_i = 1$ if athlete i tests positive and $Y_i = 0$ otherwise. A total of 13 mathletes tested positive.

What is the sampling model $p(y_1, ..., y_{37}|\theta)$?

$p(y_1, ..., y_{37}|\theta) = \binom{N}{y}\theta^y(1-\theta)^{n-y}$

Assume a uniform prior distribution on $p(\theta)$. Write the pdf for this distribution.

$$p(\theta) = \begin{cases} 1, & \text{if } 0 \leq \theta \leq 1 \\ 0, & \text{otherwise} \end{cases} *] \tag{1}$$

In what larger class of distributions does this distribution reside? What are the parameters?

Beta, $\alpha = 1$, $\beta = 1$.

Beta distribution. Recall, $\theta \sim Beta(\alpha, \beta)$ *if:*

$$*p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}*$$

Note that $E[\theta] = \frac{\alpha}{\alpha+\beta}$ and $Var[\theta] = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$ if $\theta \sim Beta(\alpha, \beta)$.

Now compute the posterior distribution, $p(\theta|\boldsymbol{y})$.

$$\begin{aligned} p(\theta|\boldsymbol{y}) &= \frac{\mathcal{L}(\theta|\boldsymbol{y})p(\theta)}{\int_\theta \mathcal{L}(\theta|\boldsymbol{y})p(\theta)d\theta} \\ &= \frac{\mathcal{L}(\theta|\boldsymbol{y})p(\theta)}{p(\boldsymbol{y})} \\ &\propto \mathcal{L}(\theta|\boldsymbol{y})p(\theta) \\ &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{13+1-1}(1-\theta)^{37-13+1-1} \\ &\sim Beta(14, 25) \end{aligned}$$

The posterior expectation,$E[\theta|y] = \frac{\alpha+y}{\alpha+\beta+n}$, *is a function of prior information and the data.*

Conjugate Priors

We have shown that a beta prior distribution and a binomial sampling model lead to a beta posterior distribution. This class of beta priors is **conjugate** for the binomial sampling model.

Def: Conjugate *A class $\mathcal{P}$ of prior distributions for θ is called conjugate for a sampling model $p(y|\theta)$ if $p(\theta) \in \mathcal{P} \to p(\theta|y) \in \mathcal{P}$.*

Conjugate priors make posterior calculations simple,*but might not always be the best representation of prior beliefs.*

Intuition about Prior Parameters

Note the posterior expectation can be written as:

$$\begin{aligned} E[\theta|y] &= \frac{\alpha + y}{\alpha + \beta + n} \\ &= \frac{\alpha + \beta}{\alpha + \beta + n}\left(\frac{\alpha}{\alpha + \beta}\right) + \frac{n}{\alpha + \beta + n}\left(\frac{y}{n}\right) \end{aligned}$$

Now what do we make of:

- α: *is roughly the prior number of 1's*
- β: *is roughly the prior number of 0's*
- $\alpha + \beta$: *is roughly the prior sample size*

If $n >> \alpha + \beta$ *then much of the information in the weighted average comes from the data.*

If $n << \alpha + \beta$ *then much of the information in the weighted average comes from the prior.*

Predictive Distributions

An important element in Bayesian statistics is the predictive distribution, in this case let Y^* be the outcome of a future experiment. We are interested in computing:

$$\begin{aligned} Pr(Y^* = 1|y_1, ..., y_n) &= \int Pr(Y^* = 1|\theta, y_1, ..., y_n) p(\theta|y_1, ..., y_n) d\theta \\ &= \int \theta p(\theta|y_1, ..., y_n) d\theta \\ &= E[\theta|y_1, ..., y_n] \\ &= \frac{\alpha + \boldsymbol{y}}{\alpha + \beta + n}, \text{where } \boldsymbol{y} = \sum_i^n y_i \end{aligned}$$

Note that the predictive distribution does not depend on any unknown quantities, but rather only the observed data. Furthermore, Y^* is not independent of $Y_1, ..., Y_n$ but depends on them through θ.

Posterior Intervals

With a Bayesian framework we can compute **credible intervals**.

Credible Interval: *An interval* $[l(y), u(y)]$ *is an* $1 - \alpha\%$ *credible interval for* θ *if:*

$$*Pr(l(y) < \theta < u(y)|Y = y) = 1 - \alpha* \tag{2}$$

Recall in a frequentist setting

$$Pr(l(y) < \theta < u(y)|\theta) = \begin{cases} *1, & \text{if } \theta \in [l(y), u(y)]* \\ *0, & \text{otherwise}* \end{cases} \tag{3}$$

Note that in some settings Bayesian intervals can also have frequentist coverage probabilities, at least asymptotically.

Quantile based intervals With quantile based intervals, the posterior quantiles are used with $\theta_{\alpha/2}, \theta_{1-\alpha/2}$ such that:

1. *$Pr(\theta < \theta_{\alpha/2}|Y = y) = \alpha/2$ and*
2. *$Pr(\theta > \theta_{1-\alpha/2}|Y = y) = \alpha/2$.*

Quantile based intervals are typically easy to compute.

Highest posterior density (HPD) region: A 100 × (1-α)% HPD region consists of a subset of the parameter space, $s(y) \subset \Theta$ such that

1. *$Pr(\theta \in s(y)|Y = y) = 1 - \alpha$*
2. *If $\theta_a \in s(y)$, and $\theta_b \notin s(y)$, then $p(\theta_a|Y = y) > p(\theta_b|Y = y)$.*

All points in the HPD region have higher posterior density than those not in region. Additionally the HPD region *need not be a continuous interval. HPD regions are typically more computationally intensive to compute than quantile based intervals.*

The Poisson Model

Now assume the National Park Services records daily totals of tourists caught breaking the rules. This data can be modeled with a Poisson model.

Recall, $Y \sim Poisson(\theta)$ if

$$* Pr(Y = y|\theta) = \frac{\theta^y \exp(-\theta)}{y!}.* \tag{4}$$

Properties of the Poisson distribution:

- E[Y]= θ
- Var(Y) = θ
- $\sum_i^n Y_i \sim Poisson(\theta_1 + ... + \theta_n)$ if $Y_i \sim Poisson(\theta_i)$

Conjugate Priors for Poisson

Recall conjugate priors for a sampling model have a posterior model from the same class as the prior. Let $y_i \sim Poisson(\theta)$, then

$$\begin{aligned} p(\theta|y_1, ..., y_n) &\propto p(\theta)\mathcal{L}(\theta|y_1, ..., y_n) && (5) \\ &\propto p(\theta) \times \theta^{\sum y_i} \exp(-n\theta) && (6) \end{aligned}$$

Thus the conjugate prior class will have the form $\theta^{c_1}\exp(c_2\theta)$. *This is the kernel of a gamma distribution.*

A positive quantity θ has a $Gamma(a,b)$ *distribution if:*

$$p(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}\exp(-b\theta), \text{ for } \theta, a, b > 0$$

Properties of *a gamma distribution:*

- $E[\theta] = a/b$

- $Var(\theta) = a/b^2$

Posterior Distribution

Let $Y_1, ..., Y_n \sim Poisson(\theta)$ and $p(\theta) \sim gamma(a,b)$, then

$$\begin{aligned} p(\theta|y_1,...,y_n) &= \frac{p(\theta)p(y_1,...,y_n|\theta)}{p(y_1,...,y_n)} && (7)\\ * &= \{\theta^{a-1}\exp(-b\theta)\} \times \{\theta^{\sum y_i}\exp(-n\theta)\} \times \{c(y_1,...,y_n,a,b)\} && (8)\\ * &\propto \theta^{a+\sum y_i - 1}\exp(-\theta(b+n)) && (9)\\ & && (10) \end{aligned}$$

So $\theta|y_1,...,y_n \sim gamma(a+\sum y_i, b+n)$.

Note that

$$\begin{aligned} E[\theta|y_1,...,y_n] &= \frac{a+\sum y_i}{b+n}\\ * &= \frac{b}{b+n}\frac{a}{b} + \frac{n}{b+n}\frac{\sum y_i}{n} * \end{aligned}$$

So now a bit of intuition about the prior distribution. The posterior expectation of θ is a combination of the prior expectation and the sample average:

- b *is interpreted as the number of prior observations*

- a *is interpreted as the sum of the counts from b prior observations*

When n >> b the information from the data dominates the prior distribution. When n << b the information from the prior distribution dominates the data.

Predictive distribution

The predictive distribution, $p(y^*|y_1, ..., y_n)$, can be computed as:

$$\begin{aligned}
p(y^*|y_1, ..., y_n) &= \int p(y^*|\theta, y_1, ..., y_n) p(\theta|y_1, ..., y_n) d\theta \\
&= \int p(y^*|\theta) p(\theta|y_1, ..., y_n) d\theta \\
&= \int \left\{ \frac{\theta^{y^*} \exp(-\theta)}{y^*!} \right\} \left\{ \frac{(b+n)^{a+\sum y_i}}{\Gamma(a+\sum y_i)} \theta^{a+\sum y_i - 1} \exp(-(b+n)\theta) \right\} \\
&= ... \\
&= ...
\end{aligned}$$

You can (and likely will) show that $p(y^*|y_1, ..., y_n) \sim NegBinom(a + \sum y_i, b + n)$.

Exponentia Families}

The binomial and Poisson models are examples of one-parameter exponential families. A distribution follows a one-parameter exponential family if it can be factorized as:

$$p(y|\theta) = h(y)c(\phi)\exp(\phi t(y)), \tag{11}$$

where ϕ is the unknown parameter and $t(y)$ is the sufficient statistic. The using the class of priors, where $p(\phi) \propto c(\phi)^{n_0} \exp(n_0 t_0 \phi)$, is a conjugate prior. There are similar considerations to the Poisson case where n_0 can be thought of as a “prior sample size” and t_0 is a “prior guess.”

Prior Distribution Choice

A noninformative prior, $p(\theta)$, *contains no information about θ.*

Example 1. Suppose θ is the probability of success in a binomial distribution, then the uniform distribution on the interval $[0, 1]$ is a noninformative prior.

Example 2. Suppose θ is the mean parameter of a normal distribution. What is a noninformative prior distribution for the mean?

- *One option would be a Normal distribution centered at zero with very large variance. However, this will still contain more mass at the center of the distribution and hence, favor that part of the parameter space.*

- *We'd like to place a uniform prior on θ, but $\int_{\infty}^{\infty} c \, dx = \infty$, so the uniform prior on the real line is not a probability distribution. Does this matter? This was actually a common prior used by LaPlace. Sometimes is the answer. Ultimately the inference is based on the posterior, so if an improper prior leads to a proper posterior that is okay. In most simple analyses we will see in this class improper priors will be fine.*

Invariant Priors

Recall ideas of variable transformation (from Casella and Berger): Let X have pdf $p_x(x)$ and let Y = g(X), where g is a monotone function. Suppose $p_X(x)$ is continuous and that $g^{-1}(y)$ has a continuous derivative. Then the pdf of Y is given by

$$p_y(y) = p_x(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

Example. Let $p_x(x) = 1$, for $x \in [0, 1]$ and let $Y = g(x) = -\log(x)$, then

$$\begin{aligned} *p_y(y) &= p_x(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \\ * &= \left| \frac{d}{dy} g^{-1}(y) \right| \\ * &= \exp(-y) \text{ for } y \in [0, \infty)* \end{aligned}$$

Now if $p_x(x)$ had been a prior on X, the transformation to y and $p_y(y)$ *results in an informative prior for y.*

Jeffreys Priors

The idea of invariant priors was addressed by Jeffreys. Let $p_J(\theta)$ be a Jeffreys prior if:

$$p_J(\theta) = [I(\theta)]^{1/2},$$

where $I(\theta)$ is the expected Fisher information given by

$$I(\theta) = -E\left[\frac{\partial^2 \log p(X|\theta)}{\partial\theta^2}\right]$$

Example. Consider the Normal distribution and place a prior on μ when σ^2 is known. Then the Fisher information is

$$\begin{aligned} *I(\theta) &= -E\left[\frac{\partial^2}{\partial\mu^2}\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)\right] \\ * &= \frac{1}{\sigma^2}* \end{aligned}$$

**Hence in this case the Jeffreys prior for μ is a constant.*

A similar derivation for the joint prior $p(\mu, \sigma) = \frac{1}{\sigma}$

Advantages and Disadvantages of Objective Priors

Advantages

- *Objective prior distributions reflect the idea of there being very little information available about the underlying process.*

- *There are sometimes mathematically equivalent results obtained by Bayesian methods using objective prior distributions and results obtained using frequentist methods on the same problem (but the results in the two cases have different philosophical interpretations.)*

- *Objective prior distributions are easy to defend.*

Disadvantages

- *Sometimes improper priors result from objective prior distributions*

- *In multiple parameter situations, the parameters are often taken to be independent*

- *Improper objective prior distributions are problematic in Bayes factor computations and some model selection settings.*

Advantages and Disadvantages of Subjective Priors

Advantages

- *Subjective prior distributions are proper.*

- *Subjective prior distributions introduced informed understanding into the analysis.*

- *Subjective prior distributions can provide sufficient information to solve problems when other methods are not sufficient.*

Disadvantages

- *It is difficult to assess subjective prior beliefs as it is hard to translate prior knowledge into a probability distribution*

- *Result of a Bayesian analysis may not be relevant if the prior beliefs used do not match your own.*

- *A subjective prior may not be computationally convenient*

In many cases weakly-informative prior distributions are used that have some of the benefits of subjective priors without imparting strong information into the analysis.