Advanced Topic: Derivation of Bayes Factors for Testing Two Proportions

August 28, 2016

Posterior Distributions and Predictive Distributions

Let's start by reviewing Bayes Theorem for the posterior under a Bernoulli sampling model and a conjugate prior Beta prior. We will start by considering the distributions with one group.

If $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} \mathsf{Ber}(\theta)$ where each of the X_i are independent and identically distributed with $P(X_i = 1 \mid \theta) = \theta$) then the conjugate prior for θ is Beta prior distribution,

$$p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$$

where $B(a,b)$ is known as the Beta function and is the normalizing constant of the Beta distribution [1]

Bayes Theorem says that

$$p(\theta \mid X_1, \ldots, X_n) = \frac{p(\theta)p(X_1, \ldots, X_n \mid \theta)}{p(X_1, \ldots, X_n)}$$

where the denominator is the marginal distribution or prior predictive distribution of the data. Using the definition of the Beta density we can find this without having to use calculus as long as we know the normalizing constants for the beta density. For those that are comfortable with calculus and integration see if you can confirm this using the definition in the footnote.

Starting with Bayes Theorem for the posterior density

$$p(\theta \mid X_1, \ldots, X_n) = \frac{p(\theta)p(X_1, \ldots, X_n \mid \theta)}{p(X_1, \ldots, X_n)}$$

we begin by substituting the distributions of data and prior

$$\begin{aligned} &\propto \frac{\prod_{i=1}^n [\theta^{X_i}(1-\theta)^{1-X_i}]\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \\ &= \frac{\theta^{\sum_{i=1}^n X_i}(1-\theta)^{\sum_{i=1}^n (1-X_i)}\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}. \end{aligned}$$

[1] the Beta function is formally defined as $B(a,b) \equiv \int_0^1 \theta^{a-1}(1-\theta)^{b-1}$ so that

$$\int_0^1 \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} = 1.$$

Letting $Y = \sum_{i=1}^{n} X_i$ and combining terms in the exponent, we have

$$\frac{\theta^{Y+a-1}(1-\theta)^{n-Y+b-1}}{B(a,b)}.$$

Recognizing that the numerator is the 'kernel' of a Beta density, we just need to multiply by $B(a,b)$ and divide by $B(Y+a, n-Y+b)$ so that the result is a Beta density:

$$p(\theta \mid X_1, \ldots, X_n) = \frac{\theta^{Y+a-1}(1-\theta)^{n-Y+b-1}}{B(a,b)} \frac{B(a,b)}{B(Y+a, n-Y+b)}$$

where the $B(A,b)$ term cancels from numerator and denominator to obtain the normalized posterior density for θ.

Since the prior predictive distribution or marginal distribution of the data is the denominator in Bayes Theorem, we have that the marginal distribution is the inverse of the term in red:

$$p(X_1, \ldots, X_n) = \frac{B(\sum X_i + a, n - \sum X_i + b)}{B(a,b)} \tag{1}$$

which is a ratio of Beta functions and the posterior density for θ simplifies to

$$p(\theta \mid X_1, \ldots, X_n) = \frac{\theta^{\sum X_i + a - 1}(1-\theta)^{n - \sum X_i + b - 1}}{B(\sum X_i + a, n - \sum X_i + b)}.$$

From this we have the updating rule for Beta distributions from Week 2: if under the prior,

$$\theta \sim B(a_0, b_0)$$

(using the subscript 0 to suggest no data), then after seeing n observations from a Bernoulli(θ) with Y "successes" and $n - Y$ "failures" the posterior distribution is

$$\theta \mid Y, n \sim B(a_n, b_n)$$

where $a_n = a_0 + Y$ and $b_n = b_0 + n - Y$ adding the prior and observed successes and the prior and observed failures.

Application to Bayes Factors

Recall that the Bayes Factor is defined as the ratio of prior predictive densities under two hypotheses H_1 and H_2:

$$BF[H_1 : H_2] \equiv \frac{p(\text{data} \mid H_1)}{p(\text{data} \mid H_2)}$$

Let's apply this to the case with two groups of Bernoulli observations $X_{A,i} \mid \theta_A \overset{iid}{\sim} \mathsf{Ber}(\theta_A)$ for $i = 1, \ldots, n_A$ and $X_{B,i} \mid \theta_B \overset{iid}{\sim} \mathsf{Ber}(\theta_B)$ for $i = 1, \ldots, n_B$ where we are interested in testing $H_1 : \theta_A = \theta_B$ versus $H_2 : \theta_A \neq \theta_B$

Under H_1 let's denote the common value of the parameter as $\theta = \theta_A = \theta_B$. Our sampling model is that

$$X_{A,i} \mid \theta, H_1 \overset{iid}{\sim} \mathsf{Ber}(\theta) \text{ for } i = 1, \ldots, n_A$$
$$X_{B,i} \mid \theta, H_1 \overset{iid}{\sim} \mathsf{Ber}(\theta) \text{ for } i = 1, \ldots, n_B$$

If additionally the observations are independent across groups we may combine them into a single sample. Using a conjugate Beta prior

$$\theta \mid H_1 \sim B(a, b)$$

then the prior predictive distributions for the data $X_{A,1}, \ldots, X_{A,n_A}, X_{B,1}, \ldots, X_{B,1}, \ldots, X_{B,n_B}$ will be

$$p(\text{data} \mid H_1) = \frac{B(Y_A + Y_B + a, n_A + n_B - Y_A - Y_B + b)}{B(a, b)}$$

where $Y_A = \sum_{i=1}^{n_A} X_{A,i}$ and $Y_B = \sum_{i=1}^{n_B} X_{B,i}$.

Under H_2 we assume that each group has its own probability of success:

$$X_{A,i} \mid \theta_A, H_2 \overset{iid}{\sim} \mathsf{Ber}(\theta_A) \text{ for } i = 1, \ldots, n_A$$
$$X_{B,i} \mid \theta_B, H_2 \overset{iid}{\sim} \mathsf{Ber}(\theta_B) \text{ for } i = 1, \ldots, n_B$$

and as before are independent. If we assign independent Beta priors to the θ's for each group

$$\theta_A \sim \mathsf{Beta}(a_A, b_A)$$
$$\theta_B \sim \mathsf{Beta}(a_B, b_B)$$

then it is straightworward to show that we may apply the result about the predictive distribution to each group separately and that the joint predictive distribution is the product of the predictive distributions within each group:

$$p(\text{data} \mid H_2) = \frac{B(Y_A + a_A, n_A - Y_A + b_A)}{B(a_A, b_A)} \times \frac{B(Y_B + a_B, n_B - Y_B + b_B)}{B(a_B, b_B)}$$

The resulting Bayes factor is

$$\begin{aligned} BF[H_1 : H_2] = &\frac{B(Y_A + Y_B + a, n_A + n_B - Y_A - Y_B + b)}{B(a, b)} \div \\ &\left[\frac{B(Y_A + a_A, n_A - Y_A + b_A)}{B(a_A, b_A)} \times \frac{B(Y_B + a_B, n_B - Y_B + b_B)}{B(a_B, b_B)}\right] \end{aligned}$$

expressed as a function of the summary counts in the two groups and sample sizes. The beta fucntion $B(,)$ is available in most statistical/mathematical programming packages. When sample sizes are large, computing the log Bayes factor is recommended

$$\begin{aligned} \log(BF[H_1 : H_2]) = &\texttt{lbeta}(\texttt{Y}_\texttt{A} + \texttt{Y}_\texttt{B} + \texttt{a}, \texttt{n}_\texttt{A} + \texttt{n}_\texttt{B} - \texttt{Y}_\texttt{A} - \texttt{Y}_\texttt{B} + \texttt{b}) - \texttt{lbeta}(\texttt{a}, \texttt{b}) - \\ &[\texttt{lbeta}(\texttt{Y}_\texttt{A} + \texttt{a}_\texttt{A}, \texttt{n}_\texttt{A} - \texttt{Y}_\texttt{A} + \texttt{b}_\texttt{A}) - \texttt{lbeta}(\texttt{a}_\texttt{A}, \texttt{b}_\texttt{A})] - \\ &[\texttt{lbeta}(\texttt{Y}_\texttt{B} + \texttt{a}_\texttt{B}, \texttt{n}_\texttt{B} - \texttt{Y}_\texttt{B} + \texttt{b}_\texttt{B}) - \texttt{lbeta}(\texttt{a}_\texttt{B}, \texttt{b}_\texttt{B})] \end{aligned}$$

where `lbeta` is the log of the Beta function.

Hyperparameters

Consonni et al (2013) suggest setting $a_A = b_A = a_B = b_B$ (symetric Beta) and that $a_A = a_B = b_A = b_B$ under H_2 and that $a = a_A + a_B$ and $b = b_A + b_B$ under H_1 so that the same amount of prior information is imposed under H_1 and H_2.

One approach for deriving this is to think about the two independent sources of information for θ_A and θ_B under H_1. One way to interpret the hyper parameters of a Beta distribution is as psuedo or imaginary observations. If we started with a_A prior successes and b_A failures and updated this with an improper prior distriubtion of the form

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

then the posterior distribution based on the psuedo observations would be

$$\theta^{a_A-1}\theta^{b_A-1}$$

which hopefully is recognizable as a $B(a_A, b_A)$. If we now use this as our prior and update it with the psuedo observations a_B and b_B, then this new posterior will be proportional to

$$\theta^{a_A+a_B-1}(1-\theta)^{b_A+b_B-1}$$

or a $B(a_A + a_B, b_A + b_B)$ distribution.

For a default prior, we may use the Jeffrey's or reference prior within each group under H_2, then $a_A = a_B = b_A = b_B = 1/2$ resulting in $a = 1, b = 1$ or a Uniform distribution for θ under H_1. If a reference prior under H_1 is desired, then $a_A = a_B = b_A = b_B = 1/4$.

Alternative Priors

Consonni et al (2013) generalize this default Bayes factor to create what they refer to as a "balanced objective" prior that addresses 1) problems with conventional Bayes factors that may inflate the evidence for the smaller model when the prior on the larger model is too diffuse as and 2) issues where the evidence in favor of the smaller model accumulates at a slower rate as the sample size increases when the smaller model is true.

Gûnel and Dickkey (1974) suggest alternative prior distributions in this and the more general problem of contingency tables. These are available in the `R` package `BayesFactor` (see Jamil et al (2016) for more details.)

References

Consonni, G., Forster, J.J. and La Rocca, L. (2013) The Whetstone and the Alum Block: Balanced Objective Bayesian Comparison of Nested Models for Discrete Data. *Statistical SCience* 28: 398–423. `https://arxiv.org/pdf/1310.0661v1.pdf`

Gûnel E. and Dickey, J. (1974) Bayes factors for independence in contingency tables *Biometrika* 61 (3): 545-557. doi: 10.1093/biomet/61.3.545 `http://biomet.oxfordjournals.org/content/61/3/545.short` http://biomet.oxfordjournals.org/content/61/3/545.short

Jamil, T., Ly, A., Morey, R.D. et al. (2016). Behavioral Research Methods. doi:10.3758/s13428-016-0739-8 `http://link.springer.com/article/10.3758%2Fs13428-016-0739-8`