Bayes Testing and More

STA 215. Surya Tokdar

Bayes testing

The basic goal of testing is to provide a summary of evidence toward/against a hypothesis of the kind $H_0 : \theta \in \Theta_0$, for some scientifically important subset Θ_0 of the parameter space Θ.

For a data model $X \sim f(x \mid \theta), \theta \in \Theta$, a Bayesian would start by specifying a prior pdf $\pi(\theta)$ for θ. The prior then combines with the data $X = x$ to produce a posterior pdf $\pi(\theta \mid x)$ for θ. At this stage, we can simply summarize the evidence toward H_0 by

$$P(H_0|x) = \Pr(\theta \in \Theta_0 \mid X = x) = \int_{\Theta_0} \pi(\theta \mid x) d\theta$$

and the evidence against H_0 is simply $1 - P(H_0|x)$.

This probability represents our updated belief about the statement H_0. If a "reject/accept H_0" type decision is indeed warranted, then we could do it by subjecting $Pr(\theta \in \Theta_0 \mid X = x)$ to a cut-off of our choice. That is, we reject H_0 if

$$\Pr(\theta \in \Theta_0 \mid X = x) < k$$

for some (positive) cut-off k. How do we choose this cut-off?

Loss function

To guide the choice of a cut-off, we need to carefully think about the consequences of our decisions. We now have to pretend that θ is going to be observed (in future) and our decision is going to be checked against the observed value. If the decision matches the observed value, we incur no penalty, otherwise we are penalized a positive amount. Let d_0 denote "we decide $\theta \in \Theta_0$" and d_1 denote "we decide $\theta \notin \Theta_0$". Then we incur a penalty if we go for d_0 and the observed θ turns out to be in $\Theta \setminus \Theta_0$, or if we go for d_1 and θ turns out to be in Θ_0. These two penalties can potentially differ in the amount we lose. This is expressed in the following loss table:

	$\theta \in \Theta_0$	$\theta \in \Theta \setminus \Theta_0$
d_0	0	w_0
d_1	w_1	0

If we denote by $\mathsf{loss}(d, \theta)$ the loss incurred when we go for a decision $d \in \{d_0, d_1\}$ and the parameter value is later observed to be θ, then

$$\mathsf{loss}(d_0, \theta) = 0, \theta \in \Theta_0, \quad \mathsf{loss}(d_0, \theta) = w_0, \theta \in \Theta \setminus \Theta_0,$$
$$\mathsf{loss}(d_1, \theta) = w_1, \theta \in \Theta_0, \quad \mathsf{loss}(d_1, \theta) = 0, \theta \in \Theta \setminus \Theta_0.$$

Therefore the posterior expected loss of a decision d

$$r(d) = \mathbb{E}[\mathsf{loss}(d,\theta)|X = x] = \int \mathsf{loss}(d,\theta)\pi(\theta \mid x)d\theta$$

can be simplified to

$$\begin{aligned} r(d_0) &= w_0 \int_{\Theta\setminus\Theta_0} \pi(\theta \mid x)d\theta = w_0 \Pr(\theta \in \Theta \setminus \Theta_0 \mid X = x) \\ r(d_1) &= w_1 \int_{\Theta_0} \pi(\theta \mid x)d\theta = w_1 \Pr(\theta \in \Theta_0 \mid X = x). \end{aligned}$$

If we go for the decision that minimizes our posterior expected loss, then we are committed to reject H_0 if (and only if)

$$r(d_1) < r(d_0) \iff \frac{\Pr(\theta \in \Theta_0 \mid X = x)}{\Pr(\theta \in \Theta \setminus \Theta_0 \mid X = x)} < \frac{w_0}{w_1} \iff \Pr(\theta \in \Theta_0 \mid X = x) < \frac{w_0}{w_0 + w_1}$$

the last equivalence follows from the fact that $\Pr(\theta \in \Theta\setminus\Theta_0 \mid X = x) = 1 - \Pr(\theta \in \Theta_0 \mid X = x)$. Tying back to the preceding section, we see that the cut-off $k = w_0/(w_0 + w_1)$ is determined by the relative gravity of the two possible mistakes we can make.

Notice that the above approach starkly differs from the "controlling errors" foundation of the classical testing procedures. In the Bayesian setting, once the post-data belief about θ is expressed by the posterior $\pi(\theta \mid x)$, the actual decisions are entirely based on expected costs associated with the two decisions where expectations are evaluated via $\pi(\theta \mid x)$. Unlike the classical setting, there is no frequentist guarantee that's sought here.

Issues with testing point nulls

Consider the statistical analysis done by Laplace on female birthrate. He had modeled X =number of female births among n births as $X \sim \mathit{Bin}(n,p)$ with $p \sim \mathit{Unif}(0,1) = \mathit{Be}(1,1)$. The observed data were $n = 493472$ and $X = 241945$ which lead to the posterior pdf $\mathit{Be}(249146, 251528)$. For testing $H_0 : p \geq 0.5$ against $H_1 : p < 0.5$ Laplace would report $\Pr(p \geq 0.5) = 10^{-42}$.

One can argue that what Laplace really wanted to study was whether $H_0 : p = 0.5$ against $H_1 : p \neq 0.5$. This presents a unique challenge. Because p is modeled with a pdf over $[0,1]$, the posterior is also a pdf over $[0,1]$ and hence $\Pr(p = 0.5 \mid X = x) = \Pr(p = 0.5) = 0$. Note that this "zero" does not reflect that the posterior concentrates away from $p = 0.5$. It is simply an artifact of our prior on p which treats p as a continuous random variable, and so the probability of any single value is simply zero. There are a couple of different ways to go about this.

Bayesian tail area probability

The goal of testing a point null $H_0 : \theta = \theta_0$ can be interpreted as judging the plausibility of a special value θ_0 (i.e., for female birth rate $p = 0.5$ is special because it captures equal odds). This can be effectively done by communicating how central θ_0 is to the posterior pdf $\pi(\theta \mid x)$.

We could look at all $100(1-\alpha)\%$, equal-tail, posterior credible intervals for θ [given by the $\alpha/2$ and $(1-\alpha/2)$th posterior quantiles of θ] and check what is the largest value of α for which this includes θ_0. This limiting α value is simply

$$2 \times \min(\Pr(\theta > \theta_0 \mid X = x), P(\theta < \theta_0 \mid X = x)).$$

If this summary is close to zero, it reflects that θ_0 is far out in the tails of the $\pi(\theta \mid x)$ pdf. I refer to the above number a "Bayesian tail area probability" that quantifies evidence in support of H_0 [with obvious analogy to p-values for classical testing.]

Ignorance range

Some statisticians contest the basic premise of a point null, arguing that it gives an extreme abstraction of a range of interesting values. That is, with $H_0 : \theta = \theta_0$ we perhaps want to capture $H_0 : |\theta - \theta_0| < d$ for some small positive number d. Thus one could instead report $P(|\theta - \theta_0| < d \mid X = x)$ for all (interesting) $d > 0$. The best way to report this would be to make a plot $P(|\theta - \theta_0| < d \mid X = x)$ as a function of $d > 0$.

Formal testing

There is in fact one other way to approach the point null testing problem. It requires using a prior distribution that recognizes that θ_0 is a special value and assigns it a positive probability. For female birthrate, this can be achieved if we describe p as follows:

$$\Pr(p = 0.5) = p_0, \quad p \mid [p \neq 0.5] \sim \pi_1(p).$$

The above indeed defines a random variable p which takes values in $[0, 1]$, but it is described by a "mixture" of a point mass at 0.5 and a pdf over $[0, 1]$.

In fact one can write the prior "pdf" of p as:

$$\pi(p) = p_0 \delta_{0.5}(p) + (1 - p_0)\pi_1(p)$$

where $\delta_a(x)$ denotes the Kronecker Delta function ($\delta_a(x) = 1$ if $x = a$, and is zero otherwise). This leads to the following calculation of posterior "pdf"

$$\begin{aligned} \pi(p|x) &= \text{const} \times p^x (1-p)^{n-x} \times \pi_1(p) \\ &= p_0(x)\delta_{0.5}(p) + (1 - p_0(x))\pi_1(p|x) \end{aligned}$$

where $\pi_1(p|x) = \text{const} \times p^x(1-p)^{n-x} \times \pi_1(p)$ and

$$p_0(x) = \frac{1}{1 + \frac{1-p_0}{p_0} \frac{\int_0^1 p^x (1-p)^{n-x} \pi_1(p) dp}{(0.5)^n}}.$$

Notice that $\Pr(p = 0.5|X = x)$ is precisely $p_0(x)$. And therefore we could report $p_0(x)$ as a summary of evidence in support of H_0, as it precisely gives $P(H_0|x)$.

However, such a formal framework for hypothesis testing is not universally accepted. A major concern being the use of a drastically different prior on θ than what one would have used if only a credible interval was to be reported. The difference in the choice of prior can have a pronounced effect on the posterior inference. The difference is often stark when apparently "low-information" priors are used for both cases. See the next example [known as Lindley's paradox].

Example. Imagine a city where 49,581 boys and 48,870 girls have been born over a certain period of time. The number of female births X is modeled with $X \sim \mathit{Bin}(n, p)$, with $n = 98451$ and $p \in [0, 1]$. For the non-informative choice $\pi(p) = \mathit{Unif}(0, 1)$ we get $P(p \geq 0.5|X = 49581, n = 98451) = 0.012$, and so a Bayesian tail area probability is $2 \times 0.012 = 0.024$, indicating moderately strong evidence against H_0. For a "low-information" point-null prior with $p_0 = 0.5$ and $\pi_1(p) = \mathit{Unif}(0, 1)$, we get $p_0(x) = 0.95$, indicating rather strong evidence toward H_0.

Model comparison and Bayes factor

In the point-null approach, we actually considered two different models:

$$\begin{aligned} M_0 &: X \sim f(x|\theta_0) \\ M_1 &: X \sim f(x|\theta), \theta \in \pi_1(\theta) \end{aligned}$$

along with prior model probabilities, $P(M_0) = p_0$ and $P(M_1) = 1 - p_0$. The quantity $p_0(x)$ is precisely $p_0(x) = P(M_0|x)$.

This setting generalizes to a more complex framework with potentially many models:

$$\begin{aligned} M_1 &: X \sim f_1(x|\theta_1), \theta_1 \sim \pi_1(\theta_1), \theta_1 \in \Theta_1 \\ M_2 &: X \sim f_2(x|\theta_2), \theta_2 \sim \pi_2(\theta_2), \theta_2 \in \Theta_2 \\ &\vdots \\ M_k &: X \sim f_k(x|\theta_k), \theta_k \sim \pi_k(\theta_k), \theta_k \in \Theta_k \end{aligned}$$

where each model can have its own distinct family of pdfs/pmfs with different parameters living on different spaces. The specification is completed by attaching prior model probabilities:

$$P(M_1) = p_1, \cdots, P(M_k) = p_k$$

with $p_i \geq 0$ and $\sum_i p_i = 1$.

Bayes rule gives that the posterior probability of model M_j is

$$P(M_j|X = x) = p_j(x) = \frac{p_i \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j}{\sum_{i=1}^{k} p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i}$$

and the conditional posterior distribution of θ_j under model M_j is

$$\pi_j(\theta_j|x) = \frac{f_j(x_j|\theta_j)\pi_j(\theta_j)}{\int_{\Theta_j} f_j(x_j|\theta_j)\pi_j(\theta_j)d\theta_j}.$$

Bayes factor

The posterior odds of model M_i to model M_j is

$$\frac{p_i(x)}{p_j(x)} = \frac{p_i}{p_j} \times \frac{\int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i}{\int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j} = \frac{p_i}{p_j} \times BF_{ij}(x)$$

where $BF_{ij}(x)$, called Bayes factor of M_i to M_j is the ratio of the marginal likelihoods of the two models. Many people prefer reporting the Bayes factor to the posterior odds, as the former does not depend on the prior odds. Any reader can multiply the reported Bayes factor with her prior odds to obtain her odds of posterior probabilities.

Marginal likelihood calculations

If $X \sim f(x|\theta)$, $\theta \sim \pi(\theta)$ is a conjugate model then the marginal likelihood $f(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$ can be calculated in closed form [this is really the normalizing constant in $\pi(\theta|x) = f(x|\theta)\pi(\theta)/f(x)$]. For example, if $X \sim Bin(n,p)$ and $p \sim bet(a,b)$, then

$$f(x) = \binom{n}{x} \int_0^1 p^x(1-p)^{n-x} \frac{p^x(1-p)^{n-x}}{B(a,b)} dp = \binom{n}{x} \frac{B(a+x, b+n-x)}{B(a,b)}.$$

For a non-conjugate model, calculation of the marginal likelihood is a fairly challenging task, usually more challenging than sampling θ from the posterior $\pi(\theta|x)$. Common numerical techniques include quadrature (when $\dim(\theta)$ is small), or stochastic calculation based on importance sampling Monte Carlo frequently coupled with sequential sampling strategies [see Tokdar and Kass (2010).]

Improper prior

In the birthrate example above, with the point-null model, we used a $Unif(0,1)$ prior on p given $p \neq 0.5$. What happens if we used a uniform prior on $\log \frac{p}{1-p}$? Recall that this corresponds to the improper $Be(0,0)$ prior on p with pdf $\pi_1(p) = c/\{p(1-p)\}$, with c arbitrary. For our data with $x = 49581 > 0$ and $n - x = 48870 > 0$ the resulting posterior is a proper $Be(49518, 48870)$ pdf. But,

$$p_0(x) = \frac{1}{1 + c \times 2^{984451} \times B(49581, 48870)}$$

which depends on the choice of c. This is a common problem with using improper priors for comparing models, though some solutions now exist in the literature (see Berger and Pericci 1996).