STA 250: Statistics

Notes 14. Bayes Testing

Book chapters: 9.8

1 Bayes testing

The basic goal of testing is to provide a summary of evidence toward/against a hypothesis of the kind $H_0 : \theta \in \Theta_0$ (against $H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$), for some scientifically important subset $\Theta_0 \subset \Theta$.

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

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

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

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

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

2 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 $\theta \in \Theta_1$". Then we incur a penalty if we go for d_0 and the observed θ turns out to be in Θ_1, 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. Let $\text{loss}(d, \theta)$ denote 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 and suppose the following "loss table" gives the possible values of $\text{loss}(d, \theta)$:

		Follow-up result	
		$\theta \in \Theta_0$	$\theta \in \Theta_1$
Decision	d_0	0	w_0
	d_1	w_1	0

Define the posterior expected loss of a decision d as

$$r(d) = \mathrm{E}\{\mathrm{loss}(d,\theta)|X = x\} = \int \mathrm{loss}(d,\theta)\xi(\theta \mid x)d\theta.$$

Then the posterior expected loss for the two decisions are:

$$r(d_0) = w_0 P(\theta \in \Theta_1 \mid X = x), \quad \text{and} \quad r(d_1) = w_1 P(\theta \in \Theta_0 \mid X = x)$$

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

$$r(d_1) < r(d_0) \iff \frac{P(\theta \in \Theta_0 \mid X = x)}{P(\theta \in \Theta_1 \mid X = x)} < \frac{w_0}{w_1} \iff P(\theta \in \Theta_0 \mid X = x) < \frac{w_0}{w_0 + w_1}$$

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.

This Bayesian approach is entrely different 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 $\xi(\theta \mid x)$, the actual decisions are entirely based on expected costs associated with the two decisions where expectations are evaluated via $\xi(\theta \mid x)$. Unlike the classical setting, there is no frequentist guarantee that's sought here.

3 Issues with testing point nulls

Consider the statistical analysis done by Laplace on female birthrate. X = number of female births among n births is modeled as $X \sim \mathsf{Binomial}(n, p)$ with $p \sim \mathsf{Uniform}(0, 1) = \mathsf{Beta}(1, 1)$. The observed data of $n = 493472$ and $X = 241945$ lead to the posterior pdf $\mathsf{Beta}(249146, 251528)$. For testing $H_0 : p \geq 0.5$ against $H_1 : p < 0.5$ we would report $P(p \geq 0.5) = 10^{-42}$.

Arguably, 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 $P(p = 0.5 \mid X = x) = P(p = 0.5) = 0$. 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.

3.1 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 $\xi(\theta \mid x)$.

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

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

If this summary is close to zero, it reflects that θ_0 is far out in the tails of the $\xi(\theta \mid x)$ 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.]

3.2 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 $d > 0$.

3.3 Formal testing

A third and formal approach to testing a point null $H_0 : \theta = \theta_0$ is to reflect the special status of θ_0 in the prior specification. This approach says we must use 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:

$$P(p = 0.5) = p_0, \quad p \mid [p \neq 0.5] \sim \xi_1(p).$$

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:

$$\xi(p) = p_0 \delta_{0.5}(p) + (1 - p_0)\xi_1(p)$$

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"

$$\begin{aligned} \xi(p|x) &= \text{const} \times p^x (1-p)^{n-x} \times \xi_1(p) \\ &= p_0(x)\delta_{0.5}(p) + (1 - p_0(x))\xi_1(p|x) \end{aligned}$$

where $\xi_1(p|x) = \text{const} \times p^x (1-p)^{n-x} \times \xi_1(p)$ and

$$p_0(x) = \frac{1}{1 + \frac{1-p_0}{p_0} \frac{\int_0^1 p^x (1-p)^{n-x} \xi_1(p) dp}{(0.5)^n}}.$$

Notice that $P(p = 0.5|X = x)$ 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 $P(H_0|x)$.

The formal approach and the Bayes tail area approach may lead to very different answers even when using apparently similar prior distributions. 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 $X \sim \mathsf{Binomial}(n, p)$, with $n = 98451$ and $p \in [0, 1]$. For the non-informative choice $\xi(p) = \mathsf{Uniform}(0, 1)$ we get

$P(p \geq 0.5|X = 48870, n = 98451) = 0.012$, 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 $\xi_1(p) = \mathsf{Uniform}(0, 1)$, we get $p_0(x) = 0.95$, indicating rather strong evidence toward H_0.

This difference manifests in many scenarios and points to additional care needed for dealing with point null hypotheses. The general consensus is that we should think really hard about how special the value $\theta = \theta_0$ is. If it is only a convenient abstraction of values near θ_0, then a tail area probability approach is a reasonable choice. But if $\theta = \theta_0$ is special in a singular way, then we must use the form approach and use a mixture prior that assigns a positive probability to $\theta = \theta_0$.