STA 250: Statistics

Notes 7. Bayesian Approach to Statistics

Book chapters: 7.2

1 From calibrating a procedure to quantifying uncertainty

We saw that the central idea of classical testing is to provide a rigorous calibration of how a testing rule "reject H_0 if $T(x) > c$" will perform under different scenarios. Once a rule with good performance (small size, large power at alternatives) has been chosen, it is applied to the data in question to make a accept/reject decision on H_0. No quantification is provided as to how certain we are about H_0 being correct.

The p-value may seem to provide one such calibration, but it is **wrong** to interpret the p-value as the probability of H_0 being true. For observed data x, the p-value may be interpreted as the maximum probability of $\{T(X) > T(x)\}$ under the null. If the p-value $= 0.02$, then we are right to say "either H_0 is false or something rare (only 2% chance) has happened". Don't be fooled in thinking that this statement means H_0 has 2% chance of being true. In fact, the statement says nothing as to what odds we are willing to assign to its "either" and "or" parts.

A similarly structured statement can be made in the following situation. Suppose you flip a coin 3000 times and count 1498 heads. Then it is correct to state "either the coin is biased or something rare (only 1.5% chance) has happened" (because, when $X \sim \mathsf{Binomial}(3000, 0.5)$, $P(X = 1498) = 0.015$.). But in this case it is quite insane to suspect that the coin maybe biased (certainly not with a 985-to-15 odds).

At the other end of the spectrum of statistical practice is the Bayesian approach where one cares more about quantifying uncertainties about statements made regarding the data and problem at hand. Below is a simple example you might be familiar with.

Example (Clinical diagnosis)**.** A clinical test for a relatively rare disease (only 1% of population affected) is tested to have a 99% accuracy rate on patients who have the disease, and a 2% failure rate on patients who do not have it. A patient takes the test and gets a positive. What are the chances that he has the disease?

Let D denote the event that this patient has the disease. Then $P(D) = 0.01$. Let T_+ denote the event that the test results in a positive. Then $P(T_+|D) = 0.99$ and $P(T_+|D^c) = 0.02$ where D^c is the complement of D, i..e., the event that the patient does not have this disease. We want to evaluate $P(D|T_+)$.

This is an instance of "inverse probability" calculation that we learn as the Bayes theorem in our probability course:

$$P(D|T_+) = \frac{P(D)P(T_+|D)}{P(D)P(T_+|D) + P(D^c)P(T_+|D^c)} = \frac{1}{3}.$$

So the patient has a one in three chance of having the disease. In other words, his odds of not having the disease is two to one. □

2 Inference via inverse probability reasoning

The formal components of the above analysis are

1. Plausibility sores attached to the two possible states of the unknown disease status ($P(D) = 0.01$, and consequently, $P(D^c) = 0.99$).

2. Plausibility scores attached to test outcomes for each state of disease status ($P(T_+|D) = 0.99 = 1 - P(T_-|D)$, where T_- is the event that the test gives a negative; similarly, $P(T_+|D^c) = 0.02 = 1 - P(T_-|D^c)$).

3. Combining the above two via the Bayes theorem to update the plausibility scores of D and D^c once an outcome of the test has been observed.

Component 2 above is like a probability model $X \sim f(x|\theta)$, with $\theta \in \Theta$ an unobservable quantity of interest (disease status of the patient with $\Theta = \{D, D^c\}$), while $X \in S$ is data to be observed (outcome of the clinical test, $S = \{T_+, T_-\}$). One learns $f(x|\theta)$ through experience and laboratory experimentation.

Component 1 is the novel feature of the Bayesian approach, where one needs to attach plausibility scores to the possible states of the unobservable quantity of interest before any observation is made. This plausibility scores represent one's prior belief – the belief that precedes the observation process.

Component 3 is pure mathematics, and results straight out of the Bayes theorem once components 2 and 3 have been specified and an observation has been made for the observable quantity.

Prior belief is not a singular quantity and cannot be learned. Prior belief combines current understanding of the unknown quantity of interest with what one is willing to assume about it. It may vary from one person to another. It may require more than a single set of plausibility scores to represent one's prior belief.

Example (Clinical diagnosis (contd.))**.** For our clinical diagnosis example, the fact that the patient has been recommended to take the test may persuade us to put $P(D)$ between 1% to 3%. In this case, $P(D|T_+)$ ranges between 33% to 61%.

3 Bayesian analysis: from prior to posterior

In the general setting, a Bayesian analysis of data X combines a statistical model $X \sim f(x|\theta)$, $x \in S$, $\theta \in \Theta$, with a **prior pdf/pmf** $\xi(\theta)$ on Θ. This pdf/pmf represents "pre-observation" (or *a priori*) plausibility scores of the parameter values. The function $h(x, \theta) = f(x|\theta)\xi(\theta)$ is simply a pd/mf on $S \times \Theta$ giving a "pre-observation" joint description of X and θ. So, once $X = x$ is observed, the "post-observation" description of θ conditional on the observed data

on X is simply the conditional pdf/pmf which can be written in the form of Bayes rule as:

$$\xi(\theta|x) = \begin{cases} \frac{f(x|\theta)\xi(\theta)}{\int_A f(x|\theta')\xi(\theta')d\theta'} & \theta \in A, \text{if } f(x|\theta)\xi(\theta) > 0 \text{ on some interval } A \\ \frac{f(x|\theta)\xi(\theta)}{\sum_{\theta\in B} f(x|\theta')\xi(\theta')} & \text{if } f(x|\theta)\xi(\theta) > 0 \text{ on some discrete set } B. \end{cases}$$

We shall call this pdf/pmf $\xi(\theta|x)$ **the posterior pdf/pmf** of θ (on Θ) based on the model $X \sim f(x|\theta)$, the prior $\xi(\theta)$ and the observation $X = x$.

4 Likelihood function and posterior pdf/pmf

Note that post the observation $X = x$, the relative plausibility of $\theta = \theta_1$ against $\theta = \theta_2$ is given by

$$\frac{\xi(\theta_1|x)}{\xi(\theta_2|x)} = \frac{f(x|\theta_1)\xi(\theta_1)}{f(x|\theta_2)\xi(\theta_2)} = \frac{L_x(\theta_1)}{L_x(\theta_2)} \times \frac{\xi(\theta_1)}{\xi(\theta_2)}.$$

Therefore the scores given by the posterior combine the scores given by the prior (pre-observation beliefs) with scores given by the likelihood function (evidence/support from observation).

$$\xi(\theta|x) = \frac{L_x(\theta)\xi(\theta)}{\int_\Theta L_x(\theta')\xi(\theta')d\theta'} = \frac{a(\theta)\xi(\theta)}{\int_\Theta a(\theta')\xi(\theta')d\theta'}$$

when $\xi(\theta)$ is a pdf, and the same holds when $\xi(\theta)$ is a pmf.

5 An example: female birth rate in 18th century Paris

The great scholar Pierre-Simon, marquis de Laplace (1749-1827) was interested in learning the rate $p \in [0, 1]$ of female birth in Paris in the 18th century. He had access to a large body of birth records in Paris between 1745 to 1770 with n entries. From these he could extract the total number of entires X which recorded a female birth. A reasonable model for X is $X \sim \mathsf{Binomial}(n, p)$, $p \in [0, 1]$.

For a prior pdf on p, Laplace decided that he had no reason to believe that for any two $p_1, p_2 \in [0, 1]$, the case $p = p_1$ was more plausible than the case $p = p_2$. In other words, Laplace believed all possible values of $p \in [0, 1]$ to be equally plausible. A pdf that ensures this is the $\mathsf{Uniform}(0, 1)$ pdf with $\xi(p) = 1$; $p \in [0, 1]$.

For an observations $X = x$, where $x \in \{0, 1, \cdots, n\}$, the likelihood function is $L_x(p) = \text{const.} \times p^x(1-p)^{n-x}$. Therefore, the posterior pdf $\xi(p|x)$ takes the form:

$$\xi(p|x) = \frac{p^x(1-p)^{n-x}}{\int_0^1 q^x(1-q)^{n-x}dq} = \frac{p^x(1-p)^{n-x}}{B(x+1, n-x+1)}, \quad p \in [0, 1].$$

where $B(a, b)$ denotes the beta function, defined for any $a > 0, b > 0$ as

$$B(a, b) = \int_0^1 q^{a-1}(1-q)^{b-1}dq = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

where $\Gamma(a) = \int_0^\infty x^{a-1}\exp(-x)dx$ is the gamma function defined for every $a > 0$.

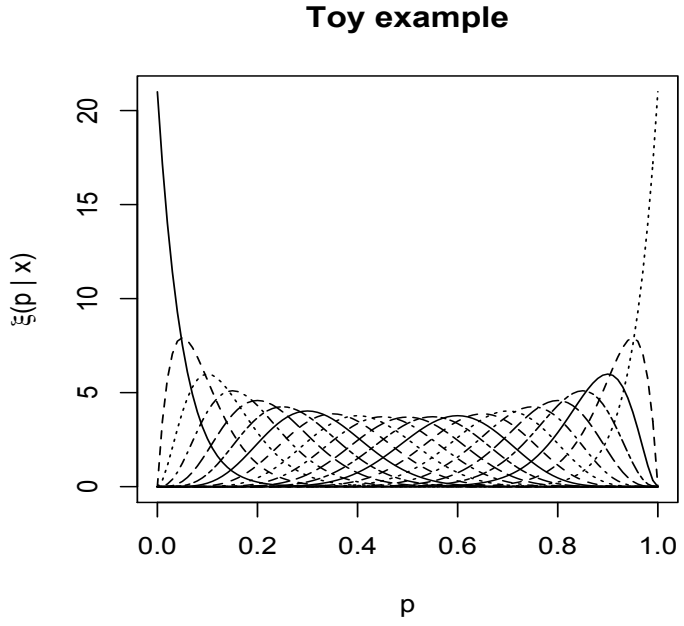

Figure 1: Posterior pdfs $\xi(p|x)$ for the model $X|\theta \sim \mathsf{Binomial}(n, p)$ and $p \sim \mathsf{Uniform}(0, 1)$. Here $n = 20$ and the posterior $\xi(p|x)$ is shown for each of $x = 0, 1, \cdots, 20$.

The pdf $g(y) = y^{a-1}(1-y)^{b-1}/B(a, b)$, $y \in [0, 1]$ is called the beta pdf with parameters a, b (both must be positive), and is denoted $\mathsf{Beta}(a, b)$. Therefore, $\xi(p|x)$ equals $\mathsf{Beta}(x + 1, n-x+1)$. In `R` , you can use `dbeta()`, `pbeta()`, `qbeta()` and `rbeta()` to get, respectively, the density function, the cumulative distribution function, the quantile function and random observations from a beta distribution. Figure 1 shows $\xi(p|x)$ against p for a toy setting with $n = 20$. Each curve on the Figure corresponds to one x in the range $0, 1, 2, \cdots, n$. As x increases from 0 to n, the peak of the $\xi(p|x)$ curve shifts from left to right.

The data Laplace had contained $n = 493472$ records with $x = 241945$ female births. This leads to a $\xi(p|x) = \mathsf{Beta}(241946, 251528)$ posterior distribution for θ. Figure 2 shows this posterior pdf. Below are some of several possible summaries of the posterior.

- Laplace was concerned whether the female birth rate was smaller than the commonly held figure of 50%. The plausibility of this event, based on Laplace's model and observed data, is simply $P(p \leq 0.5|X = x) = \int_0^{0.5} \xi(p|x)dp$, which in `R` can be computed as

  `> pbeta(0.5, 241946, 251528) = 1`. Keep in mind that this number is close to 1, but not exactly 1. In fact, it is more useful to look at converse: $P(p > 0.5|X = x) = 1 - P(p \leq 0.5|x) = \int_{0.5}^1 \xi(p|x)dp$, which equals

  `> pbeta(0.5, 241946, 251528) = 1.146e-42`.

  Laplace concluded that he was 'morally certain' that Θ is smaller than 0.5.

- If we are interested in a single number summary of p, we could try to extract a single number summary of the pdf $\xi(p|x)$. An attractive choice is the mean (expectation)

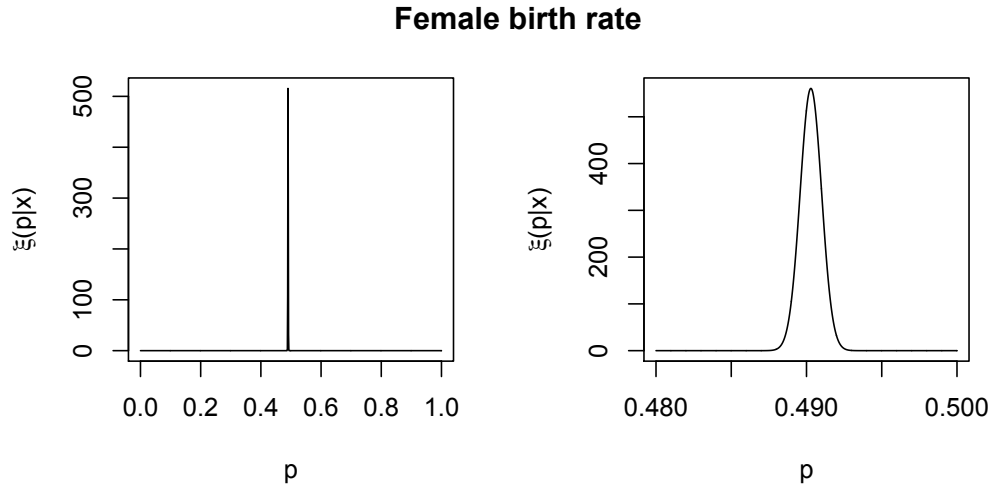

Figure 2: Posterior pdf $\xi(p|x)$ for female birth analysis by Laplace. The 50% rate ($\theta = 0.5$) is highlighted with a dotted vertical line in the middle. The posterior concentrates at a value lower than this mark. The right panel shows the same, but zooms into the range $p \in [0.48, 0.5]$.

under this pdf: $\bar{p}(x) = \mathrm{E}[p|X = x] = \int_0^1 p\xi(p|x)dp$. The mean of a $\mathsf{Beta}(a, b)$ pdf equals $a/(a + b)$, therefore the posterior mean of the female birth rate p is

```
> 241946 / (241946 + 251528) = 0.490
```

- If we are interested in reporting a range of values of p, we can look for an interval such that the pdf $\xi(p|x)$ assigns a small probability outside this interval. This is best represented by the quantiles $p_u(x)$ of $\xi(p|x)$, defined for any $u \in (0, 1)$, as the point a such that $P(\theta \leq a|X = x) = \int_0^a \xi(p|x)dp = u$. In particular, for any $\alpha \in (0, 1)$, the interval $A(x) = [p_{\alpha/2}(x), p_{1-\alpha/2}(x)]$ satisfies:

$$\begin{aligned} P(p \notin A(x)|X = x) &= P(p < p_{\alpha/2}(x)|X = x) + P(p > p_{1-\alpha/2}(x)|X = x) \\ &= \alpha/2 + \alpha/2 \\ &= \alpha. \end{aligned}$$

For $\alpha = 5\%$, the end-points of the interval $[p_{\alpha/2}(x), p_{1-\alpha/2}(x)]$ equal

```
> lower.end <- qbeta(.05 / 2, 241946, 251528) = 0.489
> upper.end <- qbeta(1 - .05 / 2, 241946, 251528) = 0.491
```

and, indeed, we can say the (posterior) probability of $\{0.489 \leq \theta \leq 0.491\}$ equals 95%.