PH 240C/STATS 245:
Clinical Trial Design (2)

Jingshen Wang

October 27, 2021

1 Bayesian statistics

The *traditional frequentist* approaches are usually conducted in following steps:

Step 1. Traditional inference (questions are pre-specified): We start with a scientific question

Step 2. Collect data (clinical trials, survey sampling, etc...)

Step 3. These data were generated randomly (by nature, by measurements, by designing a survey, etc...)

Step 4. We make assumption on the data generating processes. For example, i.i.d, parametric modelling assumption, smooth density, Gaussian errors.

Step 5. The generating process is associated to some object of interest (e.g., a parameter, a density, etc...)

Step 6. This object was unknown but *fixed* and we wanted to find it: we either estimated it or tested a hypothesis about this object, etc...

In Bayesian statistics, we still observe data, but we assume the data are randomly generated by some process, which we have prior beliefs about. Using the observed data, we want to update that belief and transform it into a posterior belief. Before we talk about Bayesian statistics in a rigorous sense, let's look into one example:

Example 1. *Suppose p is the proportion of women in the population. We then sample n people randomly with replacement in the population, and denote $X_1, \ldots, X_n$ as their genders:*

$$X_i = \begin{cases} 1 & \textit{woman} \\ 0 & \textit{otherwise} \end{cases}$$

In the frequentist approach, we estimatd p (using the MLE) and constructe some confidence interval for p. Then we form a hypothesis testing

$$H_0 : \ p = 0.5 \ \textit{v.s.} \ H_1 : p \neq 0.5.$$

Before analyzing the data, we may believe that p is likely to be close to $\frac{1}{2}$. Bayesian Statistics is a tool to: (1) include our prior belief in statistical procedures mathematically, and (2) update our prior belief using

the observed data. Suppose we are 90% sure that the proportion p is between 0.4 and 0.6, 95% that it is between 0.3 and 0.8. Hence, we can model our prior belief using a distribution for p as if p was **random**. Note that in the Frequentist frameworks, the true parameter is not random. We might thus assume that $p \sim \mathcal{B}(a, a)$ (Beta distribution) for some $a > 0$. Such a distribution is called the prior distribution.

In our statistical example that $X_1, \ldots, X_n$ are assumed to be i.i.d. Bernoulli r.v. with parameter p condition on p. After observing the available sample $X_1, \ldots, X_n$, we can update our belief about p by taking its distribution conditional on the data. The distribution of p conditional on the data is called the posterior distribution. Here, the posterior distribution is thus

$$\mathcal{B}\Big(a + \sum_{i=1}^{n} X_i, a + n - \sum_{i=1}^{n} X_i\Big).$$

The reason for us to choose Beta distribution[1] is clear: the prior distribution and the posterior distribution share the same probability distribution family. Such a prior is also called a conjugate prior.

Formally, consider a probability distribution on a parameter space Θ with some pdf $\pi(\cdot)$: the prior distribution. Let $X_1, \ldots, X_n$ be a sample of n random variables. Use $p_n(\cdot|\theta)$ to denote the joint pdf of $X_1, \ldots, X_n$ conditional on θ, where $\theta \sim \pi$. Usually, one assumes that $X_1, \ldots, X_n$ are i.i.d. condition on θ. The conditional distribution of θ given $X_1, \ldots, X_n$ is called the **posterior distribution**, denoted by

$$\pi(\cdot|X_1, \ldots, X_n).$$

Bayes' formula states that:

$$\begin{aligned}\pi(\theta|X_1, \ldots, X_n) &= \frac{\pi(\theta, X_1, \ldots, X_n)}{\pi(X_1, \ldots, X_n)} = \frac{\pi(\theta) p_n(X_1, \ldots, X_n|\theta)}{\pi(X_1, \ldots, X_n)} \\ &\propto \pi(\theta) p_n(X_1, \ldots, X_n|\theta), \quad \forall \theta \in \Theta.\end{aligned}$$

Note that the denominator does not depend on θ and its sole purpose is to make sure that the posterior distribution is a non-degenerate one as

$$\pi(\theta|X_1, \ldots, X_n) = \frac{\pi(\theta) p_n(X_1, \ldots, X_n|\theta)}{\pi(X_1, \ldots, X_n)} = \frac{\pi(\theta) p_n(X_1, \ldots, X_n|\theta)}{\int_\Theta \pi(X_1, \ldots, X_n|t) \mathrm{d}\pi(t)}, \forall \theta \in \Theta.$$

The Bayes' formula tells us that the posterior distribution is the product of the likelihood function and the prior distribution. Intuitively, the posterior is a "weighted" version of the likelihood and the weights are chosen based on my prior belief. Thus, similar to the MLE, we can also construct the Maximum a posteriori (MAP) estimation that

$$\widehat{\theta}_{\mathrm{MLE}}(X_1, \ldots, X_n) = \arg\max_{\theta \in \mathbb{R}} \pi(\theta|X_1, \ldots, X_n).$$

What is the difference between the frequentist MLE and the MAP estimator? When will they become very close?

[1] The pdf of the beta distribution is that, for $x \in [0, 1]$, given shape parameters $\alpha, \beta > 0$,

$$f(x; \alpha, \beta) = \text{normalizing constant} \cdot x^{\alpha-1}(1-x)^{\beta-1} = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\mathrm{d}u} \triangleq \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

Note that in the previous example: $\pi(p) \propto p^{a-1}(1-p)^{a-1}$, $p \in (0,1)$. Given p, $X_1, \ldots, X_n \sim$ Bernoulli(p), we have

$$p_n(X_1, \ldots, X_n|\theta) = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n}.$$

Thus,

$$\pi(p|X_1, \ldots, X_n) \propto p^{a-1+\sum_{i=1}^n X_i}(1-p)^{a-1+n-\sum_{i=1}^n X_i}.$$

The posterior distribution is written as

$$\mathcal{B}\Big(a + \sum_{i=1}^n X_i, a + n - \sum_{i=1}^n X_i\Big).$$

Non-informative prior. In case of ignorance, or lack of prior information, one may want to use a prior that is as little informative as possible. Consider our posterior distribution which is the weighted version of likelihood function, when we have no preference, a good candidate: $\pi(\theta) \propto 1$, i.e. constant pdf on Θ. If Θ is bounded, this is the uniform prior on Θ. If Θ is unbounded, this does not define a proper pdf on Θ. An improper prior on Θ is a measurable, non-negative function $\pi(\cdot)$ defined on Θ that is not integrable. In general, one **can** still define a posterior distribution using an improper prior, using Bayes' formula. Several examples of uninformative priors:

1. If $p \sim$ Unif$(0,1)$, and given p, $X_1, \ldots, X_n \sim$ Bernoulli(p), i.i.d.

   $$\pi(p|X_1, \ldots, X_n) \propto p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i},$$

   i.e. the posterior distribution is

   $$\mathcal{B}\Big(1 + \sum_{i=1}^n X_i, 1 + n - \sum_{i=1}^n X_i\Big).$$

2. If $\pi(\theta) = 1$, $\forall \theta \in \mathbb{R}$ and given θ, $X_1, \ldots, X_n \sim N(\theta, 1)$, i.i.d:

   $$\pi(\theta|X_1, \ldots, X_n) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^n (X_i - \theta)^2\right),$$

   and the posterior distribution for θ is $N(\bar{X}_n, \frac{1}{n})$.

3. A famous non-informative prior is the Jefferey's prior:

   $$\pi_J(\theta) \propto \sqrt{\det I(\theta)},$$

   where $I(\theta)$ is the Fisher information matrix of the statistical model associated with $X_1, \ldots, X_n$ in the frequentist approach (provided it exists). Thus in the previous examples:

   (a) Ex. 1: $\pi_J(\theta) \propto \frac{1}{\sqrt{p(1-p)}}$, $p \in (0,1)$.

   (b) Ex. 2: $\pi_J(\theta) \propto 1$, $\theta \in \mathbb{R}$ is an improper prior.

Jefferey's prior satisfies a reparametrization invariance principle. If η is a reparametrization of θ (i.e., $\eta = \phi(\theta)$), then the pdf $\widetilde{\pi}(\cdot)$ of η satisfies :

$$\widetilde{\pi}(\eta) \propto \sqrt{\det \widetilde{I}(\eta)},$$

where $\widetilde{I}(\eta)$ is the Fisher information of the statistical model parametrized by η instead of θ.

Bayesian confidence regions. For $\alpha \in (0,1)$, a Bayesian confidence region with level α is a random subset S of the parameter space Θ, which depends on the sample $X_1, \ldots, X_n$, such that

$$\mathbb{P}\left[\theta \in S | X_1, \ldots, X_n\right] = 1 - \alpha.$$

Note that S depends on the prior $\pi(\cdot)$. "Bayesian confidence region" (credible interval) is different from confidence interval. The credible interval contains a range of values that represent a given level of plausibility, based on the posterior distribution. In contrast, the frequentist confidence interval does not tell us whether the true value is likely to fall into the given interval.

The frequentist confidence interval defined by Neyman: If we test a large number of different null hypotheses at one critical level, say 5%, then we can collect all of the rejected null hypotheses into one set. This set usually forms a continuous interval that can be derived mathematically and Neyman described the limits of this set as confidence limits that bound a confidence interval. If the critical level (probability of incorrectly rejecting the null hypothesis) is 5% then the interval is 95%. Any values of the treatment effect that lie outside the confidence interval are regarded as "unreasonable" in terms of hypothesis testing at the critical level.

2 Thompson sampling

Thompson sampling is an algorithm for online decision problems where actions are taken sequentially in a manner that must balance between exploiting what is known to maximize immediate performance and investing to accumulate new information that may improve future performance.

Given we have developed some understanding of Bayesian Statistics. Let's look into Thompson Sampling with a simple example. Suppose there are two treatments available to us. The first treatment A has a success rate of p_A, and the second treatment B has a success rate of p_B. Both rates are unknown, but are fixed overtime while patients are enrolling in the trial.

The medical practitioner has some prior belief about success rates and assume

$$p_k \sim \text{Beta}(\alpha_k, \beta_k), \quad k \in \{A, B\}.$$

Then, on the first day when the trial starts, there are N_1 patients enrolled in the trial. For a patient i, we sample

$$\widehat{p}_{ki} \sim \text{Beta}(\alpha_k, \beta_k), \quad k \in \{A, B\}, \quad i = 1, \ldots, N_1.$$

We assign the patient i with the treatment that has a higher chance of "success:"

$$D_i = \arg\max_k \Big\{ \widehat{p}_{ki},\ k \in \{A, B\} \Big\}.$$

We observe the patient treatment assignments and associated outcomes, denoted as $\{Y_i, D_i\}_{i=1}^{N_1}$ where

$$Y_i = \begin{cases} 1 & \text{Treatment assigned to ith patient is successful} \\ 0 & \text{Treatment assigned to ith patient fails} \end{cases}$$

Therefore, the "observed rewards" for treatment A and B are defined the sum of patients who positively responded to the treatment:

$$\widehat{r}_A^1 = \sum_{i=1}^{N_1} Y_i \mathbf{1}(D_i = A), \quad \widehat{r}_B^1 = \sum_{i=1}^{N_1} Y_i \mathbf{1}(D_i = B).$$

The posterior distribution can then be updated:

$$p_k \big| \{Y_i, D_i\}_{i=1}^{N_1} \sim \text{Beta}\Big(\alpha_k + \widehat{r}_k^1, \beta_k + N_1 - \widehat{r}_k^1\Big), \quad k \in \{A, B\}.$$

From the second day onwards, we repeat the above procedure until the last day of trial enrolment. Notice here compared to the patients who arrive on the first day, second onwards-patients have a higher chance on average to receive the treatment that has a higher success rate.

Compared to the Greedy algorithm, from the second day onwards, there is no "randomness" introduced into the treatment assignment strategy. Because all patients who arrive on the second day are going to be assigned to the more successful treatment, in the sense that

$$D_i = \arg\max_k \Big\{ \widehat{r}_k^1,\ k \in \{A, B\} \Big\}.$$

Clearly, when the success treatment we identify at the first day singled out solely by chance, the second day patients are all enrolled in the wrong arm. This issue is quite intuitive: Think about choosing your favourite restaurant problem?

Although the above procedure serves our goal to maximize the patients welfare, but it brings in new issues in constructing valid confidence intervals in a frequentist sense. Think about the simple difference in mean estimator:

$$\frac{\sum_{i=1}^{N_1+N_2} Y_i D_i}{\sum_{i=1}^{N_1+N_2} D_i} - \frac{\sum_{i=1}^{N_1+N_2} Y_i (1 - D_i)}{\sum_{i=1}^{N_1+N_2} (1 - D_i)},$$

is this estimator still a "good" estimator? See Xu et al. (2013) for more discussion.

References

Min Xu, Tao Qin, and Tie-Yan Liu. Estimation bias in multi-armed bandit algorithms for search advertising. *Advances in Neural Information Processing Systems*, 26:2400–2408, 2013.