Biostat 602 Winter 2017

Lecture Set 12

Bayesian Estimation

Reading: CB 7.2.3

Frequentist Statistics

Ingredients for a Frequentist Framework

Random Variable $\mathbf{X} = (X_1, \cdots, X_n)$

Data $\mathbf{x} = (x_1, \cdots, x_n)$

Model $\mathcal{P} = \{f_{\mathbf{X}}(\mathbf{x}|\theta) : \theta \in \Omega\}$

Parameter $\theta \in \Omega$

Statistical Inference in a Frequentist Framework

Given $\mathcal{P} = \{f_{\mathbf{X}}(\mathbf{x}|\theta) : \theta \in \Omega\}$

Known $\mathbf{x} = (x_1, \cdots, x_n)$, generated from $f_{\mathbf{X}}(\mathbf{x}|\theta)$.

Unknown $\theta \in \Omega$.

There are no other assumptions of θ, which is an unknown, but a fixed value. Consequently, no probabilistic statement can be attached to the plausible values of θ.

Cancer Screening Example in a Frequentist Framework

Problem: Let $X \in \{0, 1\}$ be a random variable indicating whether an individual has a positive ($X = 1$) or negative ($X = 0$) outcome from a screening test for a particular cancer type. Let $\theta \in \{0, 1\}$ be a variable indicating whether the individual have the cancer ($\theta = 1$) or ($\theta = 0$) not at the time of screening. The distribution of X for each possible θ is given in the following table.

	$\Pr(X = 0 \mid \theta)$	$\Pr(X = 1 \mid \theta)$
$\theta = 0$	0.99	0.01
$\theta = 1$	0.05	0.95

- $\Pr(X = 1 \mid \theta = 1)$ is called the **sensitivity** of the screening test.
- $\Pr(X = 0 \mid \theta = 0)$ is called the **specificity** of the screening test.

Statistical Inference

1. Find the maximum likelihood estimator of θ.
2. What are the bias and MSE of the MLE?

Rephrasing the question

- If the individual does not have the cancer, there is 1% of chance of positive screening results.
- If the individual have the cancer, there is 95% of chance of positive screening results.
- Given X, what are the bias and MSE of MLE of θ?

Solution:

	$X=0$	$X=1$
$L(\theta=0\|X)$	0.99	0.01
$L(\theta=1\|X)$	0.05	0.95
$\hat{\theta}_{MLE}$	0	1

If the individual's screening result was positive, the MLE estimates that (s)he has the cancer. If the screening result was negative, the MLE estimates that (s)he does not have the cancer.

Bias of MLE

$$\begin{aligned} \mathrm{E}\hat{\theta} &= \Pr(\hat{\theta}=0)\cdot 0+\Pr(\hat{\theta}=1)\cdot 1 \\ &= 0.01\ I(\theta=0)+0.95\ I(\theta=1) \\ &= (1-\theta)\times 0.01+\theta\ \times 0.95 = 0.94\ \theta+0.01 \\ \mathrm{Bias}(\hat{\theta}) &= \mathrm{E}(\hat{\theta}-\theta)=0.01-0.06\ \theta \end{aligned}$$

MSE of MLE

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}) &= \mathrm{E}[(\hat{\theta}-\theta)^2]=\Pr(\hat{\theta}\neq\theta) \\ &= 0.05\ I(\theta=1)+0.01\ I(\theta=0)=0.05\ \theta+0.01(1-\theta) \\ &= 0.04\ \theta+0.01 \end{aligned}$$

The MSE is 0.01 if the individual does not have the cancer ($\theta=0$). If (s)he has the cancer ($\theta=1$), MSE is 0.05.

If you're a patient with positive screening results, you may want to ask

- Do I have cancer or not?
- What is the chance that I have cancer now?

Possible Answers

Frequentist I do not know. You're asking a wrong question. Whether you have the cancer or not is not a random variable. It is a fixed value. Therefore, the phrase "chance that you have cancer now" does not make sense.

Bayesian I think you have . . . % chance of having the cancer (How?)

Bayesian Framework

The main distinction from the frequentist framework lies in the fact that

- Parameter θ is considered as a random quantity
- Distribution of θ can be described by probability distribution, referred to as *prior* distribution
- A sample is taken from a population indexed by θ, and the prior distribution is updated using information from the sample to get *posterior* distribution of θ given the sample.

Ingredients

- Prior distribution of θ : $\theta \sim \pi(\theta)$.
- Sample distribution of $\mathbf{X}$ given θ.

$$\mathbf{X}|\theta \sim f(\mathbf{x}|\theta)$$

- Joint distribution $\mathbf{X}$ and θ

$$f(\mathbf{x}, \theta) = \pi(\theta) f(\mathbf{x}|\theta)$$

- Marginal distribution of $\mathbf{X}$

$$m(\mathbf{x}) = \int_{\theta \in \Omega} f(\mathbf{x}, \theta) d\theta = \int_{\theta \in \Omega} f(\mathbf{x}|\theta)\pi(\theta) d\theta$$

- Posterior distribution of θ (conditional distribution of θ given $\mathbf{X}$)

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}, \theta)}{m(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})} \qquad \text{(Bayes' Rule)}$$

- All of the above have discrete counterparts in which integration is replaced by summation.

Inference Under Bayesian Framework

Leveraging Prior Information

Suppose that we know that the chance of the (rare) caner per individual in is 10^{-4} (**Prevalence**).

$$
\begin{aligned}
\Pr(\theta = 1 | X = 1) = \Pr(X = 1|\theta = 1)\frac{\Pr(\theta = 1)}{\Pr(X = 1)} \qquad \text{(Bayes' rule)} \\
= \ & \Pr(X = 1|\theta = 1)\frac{\Pr(\theta = 1)}{\Pr(\theta = 1, X = 1) + \Pr(\theta = 0, X = 1)} \\
= \ & \frac{\Pr(X = 1|\theta = 1)\Pr(\theta = 1)}{\Pr(X = 1|\theta = 1)\Pr(\theta = 1) + \Pr(X = 1|\theta = 0)\Pr(\theta = 0)} \\
= \ & \frac{0.95 \times 10^{-4}}{0.95 \times 10^{-4} + 0.01 \times (1 - 10^{-4})} \approx 0.0094
\end{aligned}
$$

So, even if the screening results were positive, one can conclude that the patient have less than 1% of chance to have the cancer.

Sensitivity: $\Pr(X = 1|\theta = 1)$

Specificity: $\Pr(X = 0|\theta = 0)$

Positive Predictive Value: $\Pr(\theta = 1|X = 1)$

Negative Predictive Value: $\Pr(\theta = 0|X = 0)$

Predictive values are of interest to the patient. Predictive values are affected by the prevalence of the disease. Sensitivity and specificity are more intrinsic to the screening test.

Question: What if the prior information is misleading?

Suppose that, in fact, the cancer is highly heritable, and the patient has a parent died with the same cancer type. It is known that the chance that a child of an affected parent will also have the center is $\Pr(\theta = 1) = 0.1$.

$$\begin{aligned} & \Pr(\theta = 1|X = 1) \\ = & \frac{\Pr(X = 1|\theta = 1)\Pr(\theta = 1)}{\Pr(X = 1|\theta = 1)\Pr(\theta = 1) + \Pr(X = 1|\theta = 0)\Pr(\theta = 0)} \\ = & \frac{0.95 \times 0.1}{0.95 \times 0.1 + 0.01 \times (1 - 0.1)} \approx 0.913 \end{aligned}$$

Even though the patient has 91.3% chance of having cancer, if (s)he did not know that her/his biological parent died of the same type of cancer, (s)he may end up concluding that there are $> 99\%$ chance that this was a false alarm, and doing nothing to get a proper treatment in the early stage.

Advantages and Drawbacks of Bayesian Inference

Advantages over frequentist framework

- Allows making inference on the distribution of θ given data.
- Available information from prior experiment about θ can be utilized.
- Uncertainty of θ can be formally quantified.

Drawbacks of Bayesian Inference

- Bayesian inference is quite sensitive to prior choice. Misleading prior can result in misleading inference.
- Choice of prior distribution can be argued to be highly "subjective".
- Bayesian inference could be sometimes quite complex, requiring high dimensional integration.

Bayes Estimator

Bayes Estimator of θ is defined as the posterior mean of θ.

$$E(\theta|\mathbf{x}) = \int_{\theta\in\Omega} \theta\pi(\theta|\mathbf{x})d\theta$$

Example 1: Let $X_1, \cdots, X_n$ be i.i.d. Bernoulli(p) where $0 \leq p \leq 1$. Assume that the prior distribution of p is Beta(α, β). Find the posterior distribution of p and the Bayes estimator of p, assuming α and β are known.

Solution: Prior distribution of p is

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

Sampling distribution of $\mathbf{X}$ given p is

$$f_{\mathbf{X}}(\mathbf{x}|p) = \prod_{i=1}^{n} \left\{ p^{x_i}(1-p)^{1-x_i} \right\}$$

Joint distribution of $\mathbf{X}$ and p is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}, p) &= f_{\mathbf{X}}(\mathbf{x}|p)\pi(p) \\ &= \prod_{i=1}^{n} \left\{ p^{x_i}(1-p)^{1-x_i} \right\} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \end{aligned}$$

So the marginal distribution of $\mathbf{X}$ is

$$
\begin{aligned}
m(\mathbf{x}) &= \int f(\mathbf{x}, p)dp = \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\sum_{i=1}^n x_i+\alpha-1}(1-p)^{n-\sum_{i=1}^n x_i+\beta-1}dp \\
&= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum x_i+\alpha)\Gamma(n-\sum x_i+\beta)}{\Gamma(\alpha+\beta+n)} \\
&\quad \times \frac{\Gamma(\sum x_i+\alpha+n-\sum x_i+\beta)}{\Gamma(\sum x_i+\alpha)\Gamma(n-\sum x_i+\beta)} p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1}dp \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum_{i=1}^n x_i+\alpha)\Gamma(n-\sum_{i=1}^n x_i+\beta)}{\Gamma(\alpha+\beta+n)} \\
&\quad \times \int_0^1 f_{\mathrm{Beta}(\sum x_i+\alpha, n-\sum x_i+\beta)}(p)dp \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum_{i=1}^n x_i+\alpha)\Gamma(n-\sum_{i=1}^n x_i+\beta)}{\Gamma(\alpha+\beta+n)}
\end{aligned}
$$

The posterior distribution of $p|\mathbf{x}$:

$$
\begin{aligned}
\pi(p|\mathbf{x}) &= \frac{f(\mathbf{x}, p)}{m(\mathbf{x})} \\
&= \frac{\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1}\right]}{\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum x_i+\alpha)\Gamma(n-\sum x_i+\beta)}{\Gamma(\alpha+\beta+n)}\right]} \\
&= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\sum x_i+\alpha)\Gamma(n-\sum x_i+\beta)} p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1} \\
&\sim \mathrm{Beta}\left(\sum x_i+\alpha,\ n-\sum x_i+\beta\right)
\end{aligned}
$$

The Bayes estimator of p is

$$\begin{aligned}
\hat{p} &= \frac{\sum_{i=1}^{n} x_i + \alpha}{\sum_{i=1}^{n} x_i + \alpha + n - \sum_{i=1}^{n} x_i + \beta} = \frac{\sum_{i=1}^{n} x_i + \alpha}{\alpha + \beta + n} \\
&= \frac{\sum_{i=1}^{n} x_i}{n} \times \frac{n}{\alpha + \beta + n} \; + \; \frac{\alpha}{\alpha + \beta} \times \frac{\alpha + \beta}{\alpha + \beta + n} \\
&= [\text{Guess about } p \text{ from data}] \cdot \text{weight}_1 \\
&\quad + [\text{Guess about } p \text{ from prior}] \cdot \text{weight}_2
\end{aligned}$$

Thus the Bayes estimator is a weighted average of the prior mean and sample mean (MLE) of p. As n increases, $\text{weight}_1 = \frac{n}{\alpha+\beta+n} = \frac{1}{\frac{\alpha+\beta}{n}+1}$ becomes bigger and bigger and approaches to 1. In other words, influence of data is increasing, and the influence of prior knowledge is decreasing.

Question: Is the Bayes estimator unbiased?

$$E\left[\frac{\sum_{i=1}^{n} X_i + \alpha}{\alpha + \beta + n}\right] = \frac{np + \alpha}{\alpha + \beta + n} \neq p$$

$$\text{Bias} \;=\; \frac{np + \alpha}{\alpha + \beta + n} - p = \frac{\alpha - (\alpha + \beta)p}{\alpha + \beta + n}$$

As n increases, the bias approaches to zero.

Sufficient statistic and posterior distribution

If $T(\mathbf{X})$ is a sufficient statistic, then the posterior distribution of θ given $\mathbf{X}$ is the same as the posterior distribution of θ given $T(\mathbf{X})$. In other words,

$$\pi(\theta|\mathbf{x}) \;=\; \pi(\theta|T(\mathbf{x}))$$

Proof: The result follows since

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &= \frac{f(\mathbf{x},\theta)}{m(\mathbf{x})} \\
&= \frac{f(\mathbf{x},\theta)}{\int_{\theta\in\Omega} f(\mathbf{x},\theta)d\theta} \\
&= \frac{f(\mathbf{x},T(\mathbf{x}),\theta)}{\int_{\theta\in\Omega} f(\mathbf{x},T(\mathbf{x}),\theta)d\theta} \\
&= \frac{f(\mathbf{x}|T(\mathbf{x}),\theta)f(T(\mathbf{x})|\theta)\pi(\theta)}{\int_{\theta\in\Omega} f(\mathbf{x}|T(\mathbf{x}),\theta)f(T(\mathbf{x})|\theta)\pi(\theta)d\theta} \\
&= \frac{f(\mathbf{x}|T(\mathbf{x}))f(T(\mathbf{x})|\theta)\pi(\theta)}{f(\mathbf{x}|T(\mathbf{x}))\int_{\theta\in\Omega} f(T(\mathbf{x})|\theta)\pi(\theta)d\theta} \\
&= \frac{f(T(\mathbf{x})|\theta)\pi(\theta)}{m(T(\mathbf{x}))} \\
&= \pi(\theta|T(\mathbf{x})
\end{aligned}$$

Conjugate Family

Definition 7.2.15: Let $\mathcal{F}$ denote the class of pdfs or pmfs for $f(x|\theta)$. A class Π of prior distributions is a conjugate family of $\mathcal{F}$, if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, and all priors in Π, and all $x \in \mathcal{X}$.

Example 1 (revisited):

Let $X_1, \cdots, X_n|p \sim$ Bernoulli(p), $\pi(p) \sim$ Beta(α, β) where α, β are known.

The posterior distribution is

$$\pi(p|\mathbf{x}) \sim \text{Beta}\left(\sum_{i=1}^{n} x_i + \alpha,\ n - \sum_{i=1}^{n} x_i + \beta\right)$$

Example 2: Beta-Binomial conjugate

Let $X_1, \cdots, X_n|p \sim$ Binomial(m, p), $\pi(p) \sim$ Beta(α, β) where m, α, β are known.

Then, following the steps of Example 1, the posterior distribution is shown to be

$$\pi(p|\mathbf{x}) \sim \text{Beta}\left(\sum_{i=1}^{n} x_i + \alpha,\ mn - \sum_{i=1}^{n} x_i + \beta\right)$$

Example 3: (Gamma-Poisson) Let $X_1, \cdots, X_n|\lambda \sim \text{Poisson}(\lambda)$, and let $\pi(\lambda) \sim \text{Gamma}(\alpha, \beta)$. Find Bayes estimator of λ.

Solution:

Prior:

$$\pi(\lambda) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \lambda^{\alpha-1} e^{-\lambda/\beta}$$

Sampling distribution

$$\begin{aligned} \mathbf{X}|\lambda \;\; i.i.d. \;\; \frac{e^{-\lambda}\lambda^x}{x!} \\ f_{\mathbf{X}}(\mathbf{x}|\lambda) &= \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \end{aligned}$$

Joint distribution of X and λ

$$\begin{aligned} f(\mathbf{x}|\lambda)\pi(\lambda) &= \left[\prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right] \frac{1}{\Gamma(\alpha)\beta^\alpha} \lambda^{\alpha-1} e^{-\lambda/\beta} \\ &= e^{-n\lambda-\lambda/\beta} \lambda^{\sum x_i + \alpha - 1} \frac{1}{\prod_{i=1}^n x_i!} \frac{1}{\Gamma(\alpha)\beta^\alpha} \end{aligned}$$

Marginal distribution

$$m(\mathbf{x}) = \int f(\mathbf{x}|\lambda)\pi(\lambda) d\lambda = \frac{\Gamma(\sum x_i + \alpha) \left(\frac{1}{n+\frac{1}{\beta}}\right)^{\sum x_i + \alpha}}{\prod_{i=1}^n x_i! \Gamma(\alpha)\beta^\alpha}$$

Posterior distribution

$$\begin{aligned}\pi(\lambda|\mathbf{x}) &= \frac{f(\mathbf{x}|\lambda)\pi(\lambda)}{m(\mathbf{x})} \\ &= e^{-n\lambda-\lambda/\beta}\lambda^{\sum x_i+\alpha-1}\frac{1}{\Gamma(\sum x_i+\alpha)\left(\frac{1}{n+\frac{1}{\beta}}\right)^{\sum x_i+\alpha}}\end{aligned}$$

So, the posterior distribution is Gamma $\left(\sum x_i+\alpha, \left(n+\frac{1}{\beta}\right)^{-1}\right)$.

Bayes Estimator:

$$\hat{\lambda}_B = \frac{\sum x_i+\alpha}{n+\frac{1}{\beta}} = \frac{n}{n+\frac{1}{\beta}}\cdot\frac{\sum x_i}{n}+\frac{1/\beta}{n+\frac{1}{\beta}}\cdot(\alpha\beta)$$

Question: Is it necessary to calculate the marginal distribution?

- Marginal is a normalization constant and its evaluation is typically not needed to identify the posterior distribution if the posterior belongs to a standard distribution family.
- Posterior distribution is "proportional" to the joint distribution. Trick is to write the joint as proportional to the product of sampling and prior distribution omitting constants not depending on the parameter.
- Then identify the structure to a known family of distributions.

Example: Beta-Binomial

$$\pi(p|\mathbf{x}) \propto p^{\sum_{i=1}^n x_i}(1-p)^{mn-\sum_{i=1}^n x_i}\times p^{\alpha-1}(1-p)^{\beta-1} = p^{\sum_{i=1}^n x_i+\alpha-1}(1-p)^{mn-\sum_{i=1}^n x_i+\beta-1}$$

The structure matches that of a Beta distribution. Hence

$$\pi(p|\mathbf{x}) \sim \text{Beta}\left(\sum_{i=1}^n x_i+\alpha,\ mn-\sum_{i=1}^n x_i+\beta\right)$$

Example: Poisson-Gamma