4. Bayes' Estimators

The Method

Suppose again that we have an observable random variable $\boldsymbol{X}$ for an experiment, that takes values in a set S. Suppose also that distribution of $\boldsymbol{X}$ depends on a parameter θ taking values in a parameter space Θ. We will denote the probability density function of $\boldsymbol{X}$ by $g(\boldsymbol{x}|\theta)$ for $\boldsymbol{x} \in S$ and $\theta \in \Theta$. Of course, our data variable $\boldsymbol{X}$ is almost always vector-valued. The parameter θ may also be vector-valued.

In Bayesian analysis, we treat the parameter θ as a random variable, with a given probability density function $h(\theta)$, $\theta \in \Theta$. The corresponding distribution is called the **prior distribution** of θ and is intended to reflect our knowledge (if any) of the parameter, *before* we gather data. After observing $\boldsymbol{x} \in S$, we then use Bayes' theorem, named for Thomas Bayes, to compute the conditional probability density function of θ given $\boldsymbol{X} = \boldsymbol{x}$:

$$h(\theta|\boldsymbol{x}) = \frac{h(\theta)\, g(\boldsymbol{x}|\theta)}{g(\boldsymbol{x})}, \quad \theta \in \Theta, \boldsymbol{x} \in S$$

where g is the (marginal) probability density function of $\boldsymbol{X}$. Recall that for fixed $\boldsymbol{x} \in S$,

$$g(\boldsymbol{x}) = \sum_{\theta \in \Theta} h(\theta)\, g(\boldsymbol{x}|\theta)$$

if the parameter has a discrete distribution, or

$$g(\boldsymbol{x}) = \int_{\Theta} h(\theta)\, g(\boldsymbol{x}|\theta) d\theta$$

if the parameter has a continuous distribution. Equivalently, $g(\boldsymbol{x})$ is simply the *normalizing constant* for $h(\theta)\, g(\boldsymbol{x}|\theta)$ as a function of θ. The conditional distribution of θ given $\boldsymbol{X} = \boldsymbol{x}$ is called the **posterior distribution**, and is an updated distribution, given the information in the data.

If θ is a real parameter, the conditional expected value $\mathbb{E}(\theta|\boldsymbol{X})$ is the **Bayes' estimator** of θ. Recall that $\mathbb{E}(\theta|\boldsymbol{X})$ is a function of $\boldsymbol{X}$ and, among all functions of $\boldsymbol{X}$, is closest to θ in the mean square sense.

Special Distributions

Conjugate Families

In many important special cases, we can find a parametric family of distributions with the following property: If the prior distribution of θ belongs to the family, then so does the posterior distribution of θ given $\boldsymbol{X} = \boldsymbol{x}$. The family is said to be **conjugate** for the distribution of $\boldsymbol{X}$. Conjugate families are nice from a

computational point of view, since we can often compute the posterior distribution through a simple formula involving the parameters of the family, without having to use Bayes' theorem directly.

The Bernoulli Distribution

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the Bernoulli distribution with unknown success parameter $p \in [0, 1]$. In the usual language of reliability, $X_i = 1$ means success on trial i and $X_i = 0$ means failure on trial i. For specific example, suppose that we have a coin with an unknown probability of heads p. We toss the coin n times, defining heads as success and tails as failure. In any event, the number of successes in the n trials is

$$Y = \sum_{i=1}^{n} X_i$$

Suppose now that we give p a prior beta distribution with left parameter a and right parameter b, where a and b are chosen to reflect our initial information about p. For example, if we know nothing, we might let $a = b = 1$, so that the prior distribution of p is uniform on the parameter space $[0, 1]$. On the other hand, if we believe that p is about $\frac{2}{3}$, we might let $a = 4$ and $b = 2$ (so that the mean of the prior distribution is $\frac{2}{3}$).

1. Show that the posterior distribution of p given $\boldsymbol{X}$ is beta with left parameter $a + Y$ and right parameter $b + (n - Y)$.

Thus, the beta distribution is conjugate for the Bernoulli distribution. Note also that the posterior distribution depends on the data vector $\boldsymbol{X}$ only through the number of successes Y. This is true because Y is a sufficient statistic for p. In particular, note that the right beta parameter is increased by the number of successes and the left beta parameter is increased by the number of failures.

2. In the beta coin experiment, set $n = 10$ and $p = 0.7$, and set $a = b = 1$ (the uniform prior). Run the simulation 100 times, updating after each run. Note the shape and location of the posterior probability density function of p on each run.

3. Show that the Bayes' estimator of p given X is

$$U = \frac{a + Y}{a + b + n}$$

4. In the beta coin experiment, set $n = 20$ and $p = 0.3$, and set $a = 4$ and $b = 2$. Run the simulation 100 times, updating after each run. Note the estimate of p and the shape and location of the posterior probability density function of p on each run.

5. Verify the bias of U given p in the following formula. Show that U is asymptotically unbiased.

$$\text{bias}(U|p) = \frac{a(1 - p) - b\,p}{a + b + n}$$

Note also that we cannot choose a and b to make U unbiased, since such a choice would involve the true value

of p, which we do not know.

6. In the beta coin experiment, vary the parameters and note the change in the bias. Now set $n = 20$ and $p = 0.8$, and set $a = 2$ and $b = 6$. Run the simulation 1000 times, updating every 10 runs. Note the estimate of p and the shape and location of the posterior probability density function of p on each update. Note the apparent convergence of the empirical bias to the true bias.

7. Verify the mean square error of U given p in the following formula. Show that U is consistent.

$$\mathrm{MSE}(U|p) = \frac{p(n - 2a(a+b)) + p^2\left((a+b)^2 - n\right) + a^2}{(a+b+n)^2}$$

8. In the beta coin experiment, vary the parameters and note the change in the mean square error. Now set $n = 10$ and $p = 0.7$, and set $a = b = 1$. Run the simulation 1000 times, updating every 10 runs. Note the estimate of p and the shape and location of the posterior probability density function of p on each update. Note the apparent convergence of the empirical mean square error to the true mean square error.

Interestingly, we can choose a and b so that U has mean square error that is independent of p:

9. Show that if $a = b = \frac{\sqrt{n}}{2}$ then for any p,

$$\mathrm{MSE}(U|p) = \frac{n}{4\left(n + \sqrt{n}\right)^2}$$

10. In the beta coin experiment, set $n = 36$ and $a = b = 3$. Vary p and note that the mean square error does not change. Now set $p = 0.8$ and run the simulation 1000 times, updating every 10 runs. Note the estimate of p and the shape and location of the posterior probability density function on each update. Note the apparent convergence of the empirical bias and mean square error to the true values.

Recall that the method of moments estimator and the maximum likelihood estimator of p is the sample mean (the proportion of heads):

$$M = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

This estimator has mean square error $\mathrm{MSE}(M|p) = \frac{1}{n}p(1-p)$.

11. Sketch the graphs of $\mathrm{MSE}(U|p)$ in Exercise 7 and $\mathrm{MSE}(M|p)$ as functions of p, on the same set of axes.

The Bernoulli Distribution Revisited

Consider the coin interpretation of Bernoulli trials in the preceding section, but suppose now that the coin is either fair or two-headed. We give p the prior distribution with probability density function h given by $h(1) = a$, $h\left(\frac{1}{2}\right) = 1 - a$, where $a \in (0, 1)$ is chosen to reflect our prior knowledge of the probability that the coin is two-headed.

12. Show that the posterior distribution of p given $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is as follows. Interpret the result.

$$h(1|\boldsymbol{X}) = \begin{cases} \dfrac{2^n a}{2^n a + (1-a)}, & Y = n \\ 0, & Y < n \end{cases}$$

$$h\left(\frac{1}{2}\middle|\boldsymbol{X}\right) = 1 - h(1|\boldsymbol{X}) = \begin{cases} \dfrac{1-a}{2^n a + (1-a)}, & Y = n \\ 1, & Y < n \end{cases}$$

13. Show that the Bayes' estimator of p is

$$U = \begin{cases} p_n, & Y = n \\ \dfrac{1}{2}, & Y < n \end{cases}, \quad \text{where } p_n = \frac{2^{n+1} a + (1-a)}{2^{n+1} a + 2(1-a)}$$

14. Verify the bias of U given p in the following formula. Show that U is asymptotically unbiased.

$$\text{bias}(U|p) = \begin{cases} 1 - p_n, & p = 1 \\ \left(\dfrac{1}{2}\right)^n \left(\dfrac{1}{2} - p_n\right), & p = \dfrac{1}{2} \end{cases}$$

15. Verify the mean square error of U given p in the following formula. Show that U is consistent.

$$\text{MSE}(U|p) = \begin{cases} (1 - p_n)^2, & p = 1 \\ \left(\dfrac{1}{2}\right)^n \left(\dfrac{1}{2} - p_n\right)^2, & p = \dfrac{1}{2} \end{cases}$$

The Geometric distribution

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the geometric distribution with unknown success parameter p. As usual, we will denote the sum of the sample values by

$$Y = \sum_{i=1}^n X_i$$

Recall that the sample variables can be interpreted as the number of trials between successive successes in a sequence of Bernoulli trials. Thus, Y is the trial number of the n^{th} success. Given p, Y has the negative binomial distribution with parameters n and p. Suppose now that we give p a prior beta distribution with left parameter $a > 0$ and right parameter $b > 0$ as before.

16. Show that the posterior distribution of p given $\boldsymbol{X}$ is beta with left parameter $a + n$ and right parameter $b + (Y - n)$.

Thus, the beta distribution is conjugate to the geometric distribution.

17. Show that the Bayes' estimator of p is

$$V = \frac{a + n}{a + b + Y}$$

The Poisson Distribution

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the Poisson distribution with parameter θ. Moreover, suppose that θ has a prior gamma distribution with shape parameter $k > 0$ and scale parameter $b > 0$. As usual, we will denote the sum of the sample values by

$$Y = \sum_{i=1}^{n} X_i$$

18. Show that the posterior distribution of θ given $\boldsymbol{X}$ is gamma with shape parameter $k + Y$ and scale parameter $\frac{b}{n\,b+1}$.

It follows that the gamma distribution is conjugate to the Poisson distribution.

19. Show that the Bayes' estimator of θ is

$$V = \frac{(k + Y)\,b}{n\,b + 1}$$

20. Show that the bias of V is given by the following formula, and hence that V is asymptotically unbiased:

$$\text{bias}(V|\theta) = \frac{k\,b - \theta}{n\,b + 1}$$

Note that, as before, we cannot choose k and b to make V unbiased, without knowledge of θ

21. Show that the mean square error of V is given by the following formula, and hence V is consistent:

$$\text{MSE}(V|\theta) = \frac{\theta\left(n\,b^2 - 2\,k\,b\right) + \theta^2 + k^2\,b^2}{(n\,b + 1)^2}$$

The Normal Distribution

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the normal distribution with unknown mean μ and known variance $\sigma^2 \in (0, \infty)$. Moreover, suppose that μ is given a prior normal distribution with mean $a \in \mathbb{R}$ and variance $b^2 \in (0, \infty)$, both known of course. Denote the sum of the sample values by

$$Y = \sum_{i=1}^{n} X_i$$

22. Show that the posterior distribution of μ given $\boldsymbol{X}$ is normal with mean and variance given by.

$$\mathbb{E}(\mu|\boldsymbol{X}) = \frac{Y b^2 + a\sigma^2}{\sigma^2 + n b^2}, \quad \text{var}(\mu|\boldsymbol{X}) = \frac{\sigma^2 b^2}{\sigma^2 + n b^2}$$

Therefore, the normal distribution is conjugate for the normal distribution with unknown mean and known variance. Moreover, it follows that the Bayes' estimator of μ is

$$U = \frac{Y b^2 + a\sigma^2}{\sigma^2 + n b^2}$$

23. Show that the bias of U is given by the following formula, and hence U is asymptotically unbiased:

$$\text{bias}(U|\mu) = \frac{\sigma^2 (a - \mu)}{\sigma^2 + n b^2}$$

24. Show that the mean square error of U is given by the following formula and hence U is consistent:

$$\text{MSE}(U|\mu) = \frac{n\sigma^2 b^4 + \sigma^4 (a - \mu)^2}{\left(\sigma^2 + n b^2\right)^2}$$