7.4: Bayesian Estimation

Basic Theory

The General 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 T. Of course, our data variable $\boldsymbol{X}$ is almost always vector-valued, so that typically $S \subseteq \mathbb{R}^n$ for some $n \in \mathbb{N}_+$. Depending on the nature of the sample space S, the distribution of $\boldsymbol{X}$ may be discrete or continuous. The parameter θ may also be vector-valued, so that typically $T \subseteq \mathbb{R}^k$ for some $k \in \mathbb{N}_+$.

In *Bayesian analysis*, named for the famous Thomas Bayes, we model the deterministic, but unknown parameter θ with a random variable Θ that has a specified distribution on the parameter space T. Depending on the nature of the parameter space, this distribution may also be either discrete or continuous. It is called the *prior distribution* of Θ and is intended to reflect our knowledge of the parameter θ, *before* we gather data. After observing $\boldsymbol{X} = \boldsymbol{x} \in S$, we then use Bayes' theorem, to compute the conditional distribution of Θ given $\boldsymbol{X} = \boldsymbol{x}$. This distribution is called the *posterior distribution* of Θ, and is an updated distribution, given the information in the data. Here is the mathematical description, stated in terms of probability density functions.

Suppose that the *prior distribution* of Θ on T has probability density function h, and that given $\Theta = \theta \in T$, the conditional probability density function of $\boldsymbol{X}$ on S is $f(\cdot \mid \theta)$. Then the probability density function of the *posterior distribution* of Θ given $\boldsymbol{X} = \boldsymbol{x} \in S$ is

$$h(\theta \mid \boldsymbol{x}) = \frac{h(\theta) f(\boldsymbol{x} \mid \theta)}{f(\boldsymbol{x})}, \quad \theta \in T \tag{7.4.1}$$

where the function in the denominator is defined as follows, in the discrete and continuous cases, respectively:

$$f(\boldsymbol{x}) = \sum_{\theta \in T} h(\theta) f(\boldsymbol{x} \mid \theta), \quad \boldsymbol{x} \in S$$

$$f(\boldsymbol{x}) = \int_T h(\theta) f(\boldsymbol{x} \mid \theta) \, d\theta, \quad \boldsymbol{x} \in S$$

Proof

This is just Bayes' theorem with new terminology. Recall that he joint probability density function of $(\boldsymbol{X}, \Theta)$ is the mapping on $S \times T$ given by

$$(\boldsymbol{x}, \theta) \mapsto h(\theta) f(\boldsymbol{x} \mid \theta) \tag{7.4.2}$$

Then the function in the denominator is the marginal probability density function of $\boldsymbol{X}$. So by definition, $h(\theta \mid x) = h(\theta) f(\boldsymbol{x} \mid \theta) / f(\boldsymbol{x})$ for $\theta \in T$ is the conditional probability density function of Θ given $\boldsymbol{X} = \boldsymbol{x}$.

For $\boldsymbol{x} \in S$, note that $f(\boldsymbol{x})$ is simply the *normalizing constant* for the function $\theta \mapsto h(\theta) f(\boldsymbol{x} \mid \theta)$. It may not be necessary to explicitly compute $f(\boldsymbol{x})$, if one can recognize the functional form of $\theta \mapsto h(\theta) f(\boldsymbol{x} \mid \theta)$ as that of a known distribution. This will indeed be the case in several of the examples explored below.

If the parameter space T has finite measure c (counting measure in the discrete case or Lebesgue measure in the continuous case), then one possible prior distribution is the uniform distribution on T, with probability density function $h(\theta) = 1/c$ for $\theta \in T$. This distribution reflects no prior knowledge about the parameter, and so is called the *non-informative* prior distributioon.

Random Samples

Of course, an important and essential special case occurs when $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the distribution of a basic variable X. Specifically, suppose that X takes values in a set R and has probability density function $g(\cdot \mid \theta)$ for a given $\theta \in T$. In this case, $S = R^n$ and the probability density function $f(\cdot \mid \theta)$ of $\boldsymbol{X}$ given θ is

$$f(x_1, x_2, \ldots, x_n \mid \theta) = g(x_1 \mid \theta) g(x_2 \mid \theta) \cdots g(x_n \mid \theta), \quad (x_1, x_2, \ldots, x_n) \in S \tag{7.4.3}$$

Real Parameters

Suppose that θ is a real-valued parameter, so that $T \subseteq \mathbb{R}$. Here is our main definition.

The conditional expected value $\mathbb{E}(\Theta \mid \boldsymbol{X})$ is the *Bayesian estimator* of θ.

1. If Θ has a discrete distribution on T then
$$\mathbb{E}(\Theta \mid \boldsymbol{X} = \boldsymbol{x}) = \sum_{\theta \in T} \theta h(\theta \mid \boldsymbol{x}), \quad \boldsymbol{x} \in S \tag{7.4.4}$$
2. If Θ has a continuous distribution on T then
$$\mathbb{E}(\Theta \mid \boldsymbol{X} = \boldsymbol{x}) = \int_T \theta h(\theta \mid \boldsymbol{x}) d\theta, \quad \boldsymbol{x} \in S \tag{7.4.5}$$

Recall that $\mathbb{E}(\Theta \mid \boldsymbol{X})$ is a function of $\boldsymbol{X}$ and, among all functions of $\boldsymbol{X}$, is closest to Θ in the mean square sense. Of course, once we collect the data and observe $\boldsymbol{X} = \boldsymbol{x}$, the *Bayesian estimate* of θ is $\mathbb{E}(\Theta \mid \boldsymbol{X} = \boldsymbol{x})$. As always, the term *estimator* refers to a random variable, before the data are collected, and the term *estimate* refers to an observed value of the random variable after the data are collected. The definitions of bias and mean square error are as before, but now conditioned on $\Theta = \theta \in T$.

Suppose that U is the Bayes estimator of θ.

1. The *bias* of U is $\bias(U \mid \theta) = \mathbb{E}(U - \theta \mid \Theta = \theta)$ for $\theta \in T$.
2. The *mean square error* of U is $\mse(U \mid \theta) = \mathbb{E}[(U - \theta)^2 \mid \Theta = \theta]$ for $\theta \in T$.

As before, $\bias(U \mid \theta) = \mathbb{E}(U \mid \theta) - \theta$ and $\mse(U \mid \theta) = \var(U \mid \theta) + \bias^2(U \mid \theta)$. Suppose now that we observe the random variables $(X_1, X_2, X_3, \ldots)$ sequentially, and we compute the Bayes estimator U_n of θ based on $(X_1, X_2, \ldots, X_n)$ for each $n \in \mathbb{N}_+$. Again, the most common case is when we are sampling from a distribution, so that the sequence is independent and identically distributed (given θ). We have the natural asymptotic properties that we have seen before.

Let $\boldsymbol{U} = (U_n : n \in \mathbb{N}_+)$ be the sequence of Bayes estimators of θ as above.

1. $\boldsymbol{U}$ is *asymptotically unbiased* if $\bias(U_n \mid \theta) \to 0$ as $n \to \infty$ for each $\theta \in T$.
2. $\boldsymbol{U}$ is *mean-square consistent* if $\mse(U_n \mid \theta) \to 0$ as $n \to \infty$ for each $\theta \in T$.

Often we cannot construct unbiased Bayesian estimators, but we do hope that our estimators are at least asymptotically unbiased and consistent. It turns out that the sequence of Bayesian estimators $\boldsymbol{U}$ is a martingale. The theory of martingales provides some powerful tools for studying these estimators.

From the Bayesian perspective, the posterior distribution of Θ given the data $\boldsymbol{X} = \boldsymbol{x}$ is of primary importance. Point estimates of θ derived from this distribution are of secondary importance. In particular, the mean square error function $u \mapsto \mathbb{E}[(\Theta - u)^2 \mid \boldsymbol{X} = \boldsymbol{x})$, minimized as we have noted at $\mathbb{E}(\Theta \mid \boldsymbol{X} = \boldsymbol{x})$, is not the only *loss function* that can be used. (Although it's the only one that we consider.) Another possible loss function, among many, is the mean absolute error function $u \mapsto \mathbb{E}(|\Theta - u| \mid \boldsymbol{X} = \boldsymbol{x})$, which we know is minimized at the median(s) of the posterior distribution.

Conjugate Families

Often, the prior distribution of Θ is itself a member of a parametric family, with the parameters specified to reflect our prior knowledge of θ. In many important special cases, the parametric family can be chosen so that the posterior distribution of Θ given $\boldsymbol{X} = \boldsymbol{x}$ belongs to the same family for each $\boldsymbol{x} \in S$. In such a case, the family of distributions of Θ is said to be *conjugate* to the family of distributions 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. Similarly, in the case that the parameter is real valued, we can often compute the Bayesian estimator through a simple formula involving the parameters of the conjugate family.

Special Distributions

The Bernoulli Distribution

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots)$ is sequence of independent variables, each having the Bernoulli distribution with unknown success parameter $p \in (0, 1)$. In short, $\boldsymbol{X}$ is a sequence of Bernoulli trials, given p. In the usual language of reliability, $X_i = 1$ means success on trial i and $X_i = 0$ means failure on trial i. Recall that given p, the Bernoulli distribution has probability density function

$$g(x \mid p) = p^x (1 - p)^{1 - x}, \quad x \in \{0, 1\} \tag{7.4.6}$$

Note that the number of successes in the first n trials is $Y_n = \sum_{i=1}^n X_i$. Given p, random variable Y_n has the binomial distribution with parameters n and p.

Suppose now that we model p with a random variable P that has a prior beta distribution with left parameter $a \in (0, \infty)$ and right parameter $b \in (0, \infty)$, where a and b are chosen to reflect our initial information about p. So P has probability density function

$$h(p) = \frac{1}{B(a, b)} p^{a-1} (1 - p)^{b-1}, \quad p \in (0, 1) \tag{7.4.7}$$

and has mean $a / (a + b)$. For example, if we know nothing about p, we might let $a = b = 1$, so that the prior distribution is uniform on the parameter space $(0, 1)$ (the non-informative prior). 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 prior distribution is unimodal, with mean $\frac{2}{3}$. As a random process, the sequence $\boldsymbol{X}$ with p randomized by P, is known as the beta-Bernoulli process, and is very interesting on its own, outside of the context of Bayesian estimation.

For $n \in \mathbb{N}_+$, the posterior distribution of P given $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ is beta with left parameter $a + Y_n$ and right parameter $b + (n - Y_n)$.

Proof

Fix $n \in \mathbb{N}_+$. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \{0, 1\}^n$, and let $y = \sum_{i=1}^n x_i$. Then

$$f(\boldsymbol{x} \mid p) = g(x_1 \mid p) g(x_2 \mid p) \cdots g(x_n \mid p) = p^y (1 - p)^{n-y} \tag{7.4.8}$$

Hence

$$h(p) f(\boldsymbol{x} \mid p) = \frac{1}{B(a, b)} p^{a-1} (1 - p)^{b-1} p^y (1 - p)^{n-y} = \frac{1}{B(a, b)} p^{a+y-1} (1 - p)^{b+n-y-1}, \quad p \in (0, 1) \tag{7.4.9}$$

As a function of p this expression is proportional to the beta PDF with parameters $a + y$, $b + n - y$. Note that it's not necessary to compute the normalizing factor $f(\boldsymbol{x})$.

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

The Bayesian estimator of p given $\boldsymbol{X}_n$ is

$$U_n = \frac{a + Y_n}{a + b + n} \tag{7.4.10}$$

Proof

Recall that the mean of the beta distribution is the left parameter divided by the sum of the parameters, so this result follows from the previous result.

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

Next let's compute the bias and mean-square error functions.

For $n \in \mathbb{N}_+$,

$$\text{bias}(U_n \mid p) = \frac{a(1-p) - bp}{a+b+n}, \quad p \in (0,1) \tag{7.4.11}$$

The sequence $\boldsymbol{U} = (U_n : n \in \mathbb{N}_+)$ is asymptotically unbiased.

Proof

Given p, Y_n has the binomial distribution with parameters n and p so $E(Y_n \mid p) = np$. Hence

$$\text{bias}(U_n \mid p) = \mathbb{E}(U_n \mid p) - p = \frac{a+np}{a+b+n} - p \tag{7.4.12}$$

Simplifying gives the formula above. Clearly $\text{bias}(U_n \mid p) \to 0$ as $n \to \infty$.

Note also that we cannot choose a and b to make U_n unbiased, since such a choice would involve the true value of p, which we do not know.

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. Note the estimate of p and the shape and location of the posterior probability density function of p on each update. Compare the empirical bias to the true bias.

For $n \in \mathbb{N}_+$,

$$\text{mse}(U_n \mid p) = \frac{p[n - 2\,a(a+b)] + p^2[(a+b)^2 - n] + a^2}{(a+b+n)^2}, \quad p \in (0,1) \tag{7.4.13}$$

The sequence $(U_n : n \in \mathbb{N}_+)$ is mean-square consistent.

Proof

Once again, given p, Y_n has the binomail distribution with parameters n and p so

$$\text{var}(U_n \mid p) = \frac{np(1-p)}{(a+b+n)^2} \tag{7.4.14}$$

Hence

$$\text{mse}(U_n \mid p) = \frac{np(1-p)}{(a+b+n)^2} + \left[\frac{a(1-p) - bp}{a+b+n}\right]^2 \tag{7.4.15}$$

Simplifying gives the result. Clearly $\text{mse}(U_n \mid p) \to 0$ as $n \to \infty$.

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. Note the estimate of p and the shape and location of the posterior probability density function of p on each update. Compare 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 the unknown parameter p:

Let $n \in \mathbb{N}_+$ and let $a = b = \sqrt{n}/2$. Then

$$\text{mse}(U_n \mid p) = \frac{n}{4\left(n + \sqrt{n}\right)^2}, \quad p \in (0,1) \tag{7.4.16}$$

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. Note the estimate of p and the shape and location of the posterior probability density function on each update. Compare 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 (on the interval $(0,1)$) is the sample mean (the proportion of successes):

$$M_n = \frac{Y}{n} = \frac{1}{n} \sum_{i=1}^n X_i \tag{7.4.17}$$

This estimator has mean square error $\mse(M_n \mid p) = \frac{1}{n} p(1-p)$. To see the connection between the estimators, note from (6) that

$$U_n = \frac{a+b}{a+b+n} \frac{a}{a+b} + \frac{n}{a+b+n} M_n \tag{7.4.18}$$

So U_n is a weighted average of $a/(a+b)$ (the mean of the prior distribution) and M_n (the maximum likelihood estimator).

Another Bernoulli Distribution

Bayesian estimation, like other forms of parametric estimation, depends critically on the parameter space. Suppose again that $(X_1, X_2, \ldots)$ is a sequence of Bernoulli trials, given the unknown success parameter p, but suppose now that the parameter space is $\left\{\frac{1}{2}, 1\right\}$. This setup corresponds to the tossing of a coin that is either fair or two-headed, but we don't know which. We model p with a random variable P that has the prior 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. If we are completely ignorant, we might let $a = \frac{1}{2}$ (the non-informative prior). If with think the coin is more likely to be two-headed, we might let $a = \frac{3}{4}$. Again let $Y_n = \sum_{i=1}^n X_i$ for $n \in \mathbb{N}_+$.

The posterior distribution of P given $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ is

1. $h(1 \mid \boldsymbol{X}_n) = \frac{2^n a}{2^n a + (1-a)}$ if $Y_n = n$ and $h(1 \mid \boldsymbol{X}_n) = 0$ if $Y_n < n$
2. $h\left(\frac{1}{2} \mid \boldsymbol{X}_n\right) = \frac{1-a}{2^n a + (1-a)}$ if $Y_n = n$ and $h\left(\frac{1}{2} \mid \boldsymbol{X}_n\right) = 1$ if $Y_n < n$

Proof

Fix $n \in \mathbb{N}_+$. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \{0, 1\}^n$, and let $y = \sum_{i=1}^n x_i$. As before,

$$f(\boldsymbol{x} \mid p) = p^y (1-p)^{n-y} \tag{7.4.19}$$

We adopt the usual conventions (which gives the correct mathematics) that $0^k = 0$ if $k \in \mathbb{N}_+$ but $0^0 = 1$. So from Bayes' theorem,

$$h(1 \mid \boldsymbol{x}) = \frac{h(1) f(\boldsymbol{x} \mid 1)}{h(1/2) f(\boldsymbol{x} \mid 1/2) + h(1) f(\boldsymbol{x} \mid 1)} \tag{7.4.20}$$

$$= \frac{a 1^y 0^{n-y}}{(1-a)(1/2)^n + a 1^y 0^{n-y}} \tag{7.4.21}$$

So if $y < n$ then $h(1 \mid \boldsymbol{x}) = 0$ while if $y = n$

$$h(1 \mid \boldsymbol{x}) = \frac{a}{(1-a)(1/2)^n + a} \tag{7.4.22}$$

Of course, $h\left(\frac{1}{2} \mid \boldsymbol{x}\right) = 1 - h(1 \mid \boldsymbol{x})$. The results now follow after a bit of algebra.

Now let

$$p_n = \frac{2^{n+1} a + (1-a)}{2^{n+1} a + 2(1-a)} \tag{7.4.23}$$

The Bayes' estimator of p given $\boldsymbol{X}_n$ the statistic U_n defined by

1. $U_n = p_n$ if $Y_n = n$
2. $U_n = \frac{1}{2}$ if $Y_n < n$

Proof

By definition, the Bayes' estimator is $U_n = E(P \mid \boldsymbol{X}_n)$. From the previous result, if $Y_n = n$ then

$$U_n = 1 \cdot \frac{2^n a}{2^n a + (1-a)} + \frac{1}{2} \cdot \frac{1-a}{2^n a + (1-a)} \tag{7.4.24}$$

which simplifies to p_n. If $Y_n < n$ then $U = 1 \cdot 0 + \frac{1}{2} \cdot 1 = \frac{1}{2}$.

If we observe $Y_n < n$ then U_n gives the correct answer $\frac{1}{2}$. This certainly makes sense since we know that we do not have the two-headed coin. On the other hand, if we observe $Y_n = n$ then we are not certain which coin we have, and the Bayesian estimate p_n is not even in the parameter space! But note that $p_n \to 1$ as $n \to \infty$ exponentially fast. Next let's compute the bias and mean-square error for a given $p \in \left\{\frac{1}{2}, 1\right\}$.

For $n \in \mathbb{N}_+$,

1. $\text{bias}(U_n \mid 1) = p_n - 1$
2. $\text{bias}\left(U_n \mid \frac{1}{2}\right) = \left(\frac{1}{2}\right)^n \left(p_n - \frac{1}{2}\right)$

The sequence of estimators $(U_n : n \in \mathbb{N}_+)$ is asymptotically unbiased.

Proof

By definition, $\text{bias}(U_n \mid p) = E(U - p \mid p)$. Hence from the previous result,

$$\text{bias}(U \mid p) = (p_n - p)\mathbb{P}(Y = n \mid p) + \left(\frac{1}{2} - p\right)\mathbb{P}(Y < n \mid p) \tag{7.4.25}$$

$$= (p_n - p)p^n + \left(\frac{1}{2} - p\right)(1 - p^n) \tag{7.4.26}$$

Substituting $p = 1$ and $p = \frac{1}{2}$ gives the results. In both cases, $\text{bias}(U_n \mid p) \to 0$ as $n \to \infty$ since $p_n \to 1$ and $\left(\frac{1}{2}\right)^n \to 0$ as $n \to \infty$.

If $p = 1$, the estimator U_n is negatively biased; we noted this earlier. If $p = \frac{1}{2}$, then U_n is positively biased for sufficiently large n (depending on a).

For $n \in \mathbb{N}_+$,

1. $\text{mse}(U_n \mid 1) = (p_n - 1)^2$
2. $\text{mse}\left(U_n \mid \frac{1}{2}\right) = \left(\frac{1}{2}\right)^n \left(p_n - \frac{1}{2}\right)^2$

The sequence of estimators $\boldsymbol{U} = (U_n : n \in \mathbb{N}_+)$ is mean-square consistent.

Proof

By definition, $\text{mse}(U_n \mid p) = \mathbb{E}[(U_n - p)^2 \mid p]$. Hence

$$\text{mse}(U_n \mid p) = (p_n - p)^2\mathbb{P}(Y_n = n \mid p) + \left(\frac{1}{2} - p\right)^2\mathbb{P}(Y_n < n \mid p) \tag{7.4.27}$$

$$= (p_n - p)^2 p^n + \left(\frac{1}{2} - p\right)^2(1 - p^n) \tag{7.4.28}$$

Substituting $p = 1$ and $p = \frac{1}{2}$ gives the results. In both cases, $\text{mse}(U_n \mid p) \to 0$ as $n \to \infty$ since $p_n \to 1$ and $\left(\frac{1}{2}\right)^n \to 0$ as $n \to \infty$.

The Geometric distribution

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots)$ is a sequence of independent random variables, each having the geometric distribution on $\mathbb{N}_+$ with unknown success parameter $p \in (0, 1)$. Recall that these variables can be interpreted as the number of trials between successive successes in a sequence of Bernoulli trials. Given p, the geometric distribution has probability density function

$$g(x \mid p) = p(1 - p)^{x-1}, \quad x \in \mathbb{N}_+ \tag{7.4.29}$$

Once again for $n \in \mathbb{N}_+$, let $Y_n = \sum_{i=1}^n X_i$. In this setting, Y_n is the trial number of the nth success, and given p, has the negative binomial distribution with parameters n and p.

Suppose now that we model p with a random variable P having a prior beta distribution with left parameter $a \in (0, \infty)$ and right parameter $b \in (0, \infty)$. As usual, a and b are chosen to reflect our prior knowledge of p.

The posterior distribution of P given $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ is beta with left parameter $a + n$ and right parameter $b + (Y_n - n)$.

Proof

Fix $n \in \mathbb{N}_+$. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{N}_+^n$ and let $y = \sum_{i=1}^n x_i$. Then

$$f(\boldsymbol{x} \mid p) = g(x_1 \mid p) g(x_2 \mid p) \cdots g(x_n \mid p) = p^n (1 - p)^{y-n} \tag{7.4.30}$$

Hence

$$h(p) f(\boldsymbol{x} \mid p) = \frac{1}{B(a, b)} p^{a-1} (1 - p)^{b-1} p^n (1 - p)^{y-n} = \frac{1}{B(a, b)} p^{a+n-1} (1 - p)^{b+y-n-1}, \quad p \in (0, 1) \tag{7.4.31}$$

As a function of $p \in (0, 1)$ this expression is proportional to the beta PDF with parameters $a + n$ and $b + y - n$. Note that it's not necessary to compute the normalizing constant $f(\boldsymbol{x})$.

Thus, the beta distribution is conjugate to the geometric distribution. Moreover, note that in the posterior beta distribution, the left parameter is increased by the number of successes n while the right parameter is increased by the number of failures $Y - n$, just as in the Bernoulli model. In particular, the posterior left parameter is deterministic and depends on the data only through the sample size n.

The Bayesian estimator of p based on $\boldsymbol{X}_n$ is

$$V_n = \frac{a + n}{a + b + Y_n} \tag{7.4.32}$$

Proof

By definition, the Bayesian estimator is the mean of the posterior distribution. Recall again that the mean of the beta distribution is the left parameter divided by the sum of the parameters, so the result follows from our previous theorem.

Recall that the method of moments estimator of p, and the maximum likelihood estimator of p on the interval $(0, 1)$ are both $W_n = 1 / M_n = n / Y_n$. To see the connection between the estimators, note from (19) that

$$\frac{1}{V_n} = \frac{a}{a + n} \frac{a + b}{a} + \frac{n}{a + n} \frac{1}{W_n} \tag{7.4.33}$$

So $1 / V_n$ (the reciprocal of the Bayesian estimator) is a weighted average of $(a + b) / a$ (the reciprocal of the mean of the prior distribution) and $1 / W_n$ (the reciprocal of the maximum likelihood estimator).

The Poisson Distribution

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots)$ is a sequence of random variable each having the *Poisson distribution* with unknown parameter $\lambda \in (0, \infty)$. Recall that the Poisson distribution is often used to model the number of "random points" in a region of time or space and is studied in more detail in the chapter on the Poisson Process. The distribution is named for the inimitable Simeon Poisson and given λ, has probability density function

$$g(x \mid \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x \in \mathbb{N} \tag{7.4.34}$$

Once again, for $n \in \mathbb{N}_+$, let $Y_n = \sum_{i=1}^n X_i$. Given λ, random variable Y_n also has a Poisson distribution, but with parameter $n\lambda$.

Suppose now that we model λ with a random variable Λ having a prior gamma distribution with shape parameter $k \in (0, \infty)$ and rate parameter $r \in (0, \infty)$. As usual k and r are chosen to reflect our prior knowledge of λ. Thus the prior probability density function of Λ is

$$h(\lambda) = \frac{r^k}{\Gamma(k)} \lambda^{k-1} e^{-r\lambda}, \quad \lambda \in (0, \infty) \tag{7.4.35}$$

and the mean is k / r. The scale parameter of the gamma distribution is $b = 1 / r$, but the formulas will work out nicer if we use the rate parameter.

The posterior distribution of Λ given $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ is gamma with shape parameter $k + Y_n$ and rate parameter $r + n$.

Proof

Fix $n \in \mathbb{N}_+$. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{N}^n$ and $y = \sum_{i=1}^n x_i$. Then

$$f(\boldsymbol{x} \mid \lambda) = g(x_1 \mid \lambda) g(x_2 \mid \lambda) \cdots g(x_n \mid \lambda) = e^{-n\lambda} \frac{\lambda^y}{x_1! x_2! \cdots x_n!} \tag{7.4.36}$$

Hence

$$h(\lambda) f(\boldsymbol{x} \mid \lambda) = \frac{r^k}{\Gamma(k)} \lambda^{k-1} e^{-r\lambda} e^{-n\lambda} \frac{\lambda^y}{x_1! x_2! \cdots x_n!} \tag{7.4.37}$$

$$= \frac{r^k}{\Gamma(k) x_1! x_2! \cdots x_n!} e^{-(r+n)\lambda} \lambda^{k+y-1}, \quad \lambda \in (0, \infty) \tag{7.4.38}$$

As a function of $\lambda \in (0, \infty)$ the last expression is proportional to the gamma PDF with shape parameter $k + y$ and rate parameter $r + n$. Note again that it's not necessary to compute the normalizing constant $f(\boldsymbol{x})$.

It follows that the gamma distribution is conjugate to the Poisson distribution. Note that the posterior rate parameter is deterministic and depends on the data only through the sample size n.

The Bayesian estimator of λ based on $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ is

$$V_n = \frac{k + Y_n}{r + n} \tag{7.4.39}$$

Proof

By definition, the Bayes estimator is the mean of the posterior distribution. Recall that mean of the gamma distribution is the shape parameter divided by the rate parameter.

Since V_n is a linear function of Y_n, and we know the distribution of Y_n given $\lambda \in (0, \infty)$, we can compute the bias and mean-square error functions.

For $n \in \mathbb{N}_+$,

$$\bias(V_n \mid \lambda) = \frac{k - r\lambda}{r + n}, \quad \lambda \in (0, \infty) \tag{7.4.40}$$

The sequence of estimators $\boldsymbol{V} = (V_n : n \in \mathbb{N}_+)$ is asymptotically unbiased.

Proof

The computation is simple, since the distribution of Y_n given λ is Poisson with parameter $n\lambda$.

$$\bias(V_n \mid \lambda) = \mathbb{E}(V_n \mid \lambda) - \lambda = \frac{k + n\lambda}{r + n} - \lambda = \frac{k - r\lambda}{r + n} \tag{7.4.41}$$

Clearly $\bias(V_n \mid \lambda) \to 0$ as $n \to \infty$.

Note that, as before, we cannot choose k and r to make V_n unbiased, without knowledge of λ.

For $n \in \mathbb{N}_+$,

$$\mse(V_n \mid \lambda) = \frac{n\lambda + (k - r\lambda)^2}{(r + n)^2}, \quad \lambda \in (0, \infty) \tag{7.4.42}$$

The sequence of estimators $\boldsymbol{V} = (V_n : n \in \mathbb{N}_+)$ is mean-square consistent.

Proof

Again, the computation is easy since the distribution of Y_n given λ is Poisson with parameter $n\lambda$.

$$\text{mse}(V \mid \lambda) = \text{var}(V_n \mid \lambda) + \text{bias}^2(V_n \mid \lambda) = \frac{n\lambda}{(r+n)^2} + \left(\frac{k - r\lambda}{r+n}\right)^2 \tag{7.4.43}$$

Clearly $\text{mse}(V_n \mid \lambda) \to 0$ as $n \to \infty$.

Recall that the method of moments estimator of λ and the maximum likelihood estimator of λ on the interval $(0, \infty)$ are both $M_n = Y_n / n$, the sample mean. This estimator is unbiased and has mean square error λ / n. To see the connection between the estimators, note from (21) that

$$V_n = \frac{r}{r+n}\frac{k}{r} + \frac{n}{r+n} M_n \tag{7.4.44}$$

So V_n is a weighted average of k/r (the mean of the prior distribution) and M_n (the maximum likelihood estimator).

The Normal Distribution

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots)$ is a sequence of independent random variables, each having the normal distribution with unknown mean $\mu \in \mathbb{R}$ but known variance $\sigma^2 \in (0, \infty)$. Of course, the normal distribution plays an especially important role in statistics, in part because of the central limit theorem. The normal distribution is widely used to model physical quantities subject to numerous small, random errors. In many statistical applications, the variance of the normal distribution is more stable than the mean, so the assumption that the variance is known is not entirely artificial. Recall that the normal probability density function (given μ) is

$$g(x \mid \mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right], \quad x \in \mathbb{R} \tag{7.4.45}$$

Again, for $n \in \mathbb{N}_+$ let $Y_n = \sum_{i=1}^n X_i$. Recall that Y_n also has a normal distribution (given μ) but with mean $n\mu$ and variance $n\sigma^2$.

Suppose now that μ is modeled by a random variable Ψ that has a prior normal distribution with mean $a \in \mathbb{R}$ and variance $b^2 \in (0, \infty)$. As usual, a and b are chosen to reflect our prior knowledge of μ. An interesting special case is when we take $b = \sigma$, so the variance of the prior distribution of Ψ is the same as the variance of the underlying sampling distribution.

For $n \in \mathbb{N}_+$, the posterior distribution of Ψ given $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ is normal with mean and variance given by

$$\mathbb{E}(\Psi \mid \boldsymbol{X}_n) = \frac{Y_n b^2 + a\sigma^2}{nb^2 + \sigma^2} \tag{7.4.46}$$

$$\text{var}(\Psi \mid \boldsymbol{X}_n) = \frac{\sigma^2 b^2}{nb^2 + \sigma^2} \tag{7.4.47}$$

Proof

Fix $n \in \mathbb{N}_+$. Suppose $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}$ and let $y = \sum_{i=1}^n x_i$ and $w^2 = \sum_{i=1}^n x_i^2$. Then

$$f(\boldsymbol{x} \mid \mu) = g(x_1 \mid \mu) g(x_2 \mid \mu) \cdots g(x_n \mid \mu) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\frac{1}{2}\sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2\right] \tag{7.4.48}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\frac{1}{2\sigma^2}(w^2 - 2\mu y + n\mu^2)\right] \tag{7.4.49}$$

On the other hand, of course

$$h(\mu) = \frac{1}{\sqrt{2\pi}b} \exp\left[-\frac{1}{2}\left(\frac{\mu - a}{b}\right)^2\right] = \frac{1}{\sqrt{2\pi}b} \exp\left[-\frac{1}{2b^2}(\mu^2 - 2a\mu + a^2)\right] \tag{7.4.50}$$

Therefore,

$$h(\mu) f(\boldsymbol{x} \mid \mu) = C \exp\left\{-\frac{1}{2}\left[\left(\frac{1}{b^2} + \frac{n}{\sigma^2}\right)\mu^2 - 2\left(\frac{a}{b^2} + \frac{y}{\sigma^2}\right)\mu\right]\right\} \tag{7.4.51}$$

where C depends on n, σ, a, b, $\boldsymbol{x}$, but importantly *not* on μ. So we don't really care what C is. Completing the square in μ in the expression above gives

$$h(\mu) f(\boldsymbol{x} \mid \mu) = K \exp\left[-\frac{1}{2}\left(\frac{1}{b^2} + \frac{n}{\sigma^2}\right)\left(\mu - \frac{a/b^2 + y/\sigma^2}{1/b^2 + n/\sigma^2}\right)^2\right] \tag{7.4.52}$$

where K is yet another factor that depends on lots of stuff, but not μ. As a function of μ, this expression is proportional to the normal distribution with mean and variance, respectively, given by

$$\frac{a/b^2 + y/\sigma^2}{1/b^2 + n/\sigma^2} = \frac{y b^2 + a\sigma^2}{n b^2 + \sigma^2} \tag{7.4.53}$$

$$\frac{1}{1/b^2 + n/\sigma^2} = \frac{\sigma^2 b^2}{\sigma^2 + n b^2} \tag{7.4.54}$$

Once again, it was not necessary to compute the normalizing constant $f(\boldsymbol{x})$, which would have been yet another factor that we do not care about.

Therefore, the normal distribution is conjugate to the normal distribution with unknown mean and known variance. Note that the posterior variance is deterministic, and depends on the data only through the sample size n. In the special case that $b = \sigma$, the posterior distribution of Ψ given $\boldsymbol{X}_n$ is normal with mean $(Y_n + a)/(n+1)$ and variance $\sigma^2/(n+1)$.

The Bayesian estimator of μ is

$$U_n = \frac{Y_n b^2 + a\sigma^2}{n b^2 + \sigma^2} \tag{7.4.55}$$

Proof

This follows immediately from the previous result.

Note that $U_n = (Y_n + a)/(n+1)$ in the special case that $b = \sigma$.

For $n \in \mathbb{N}_+$,

$$\text{bias}(U_n \mid \mu) = \frac{\sigma^2(a - \mu)}{\sigma^2 + n\, b^2}, \quad \mu \in \mathbb{R} \tag{7.4.56}$$

The sequence of estimators $\boldsymbol{U} = (U_n : n \in \mathbb{N}_+)$ is asymptotically unbiased.

Proof

Recall that Y_n has mean $n\mu$ given μ. Hence

$$\text{bias}(U_n \mid \mu) = \mathbb{E}(U_n \mid \mu) - \mu = \frac{n b^2 \mu + a\sigma^2}{n b^2 + \sigma^2} - \mu = \frac{(a - \mu)\sigma^2}{n b^2 + \sigma^2} \tag{7.4.57}$$

Clearly $\text{bias}(U_n \mid \mu) \to 0$ as $n \to \infty$ for every $\mu \in \mathbb{R}$.

When $b = \sigma$, $\text{bias}(U_n \mid \mu) = (a - \mu)/(n+1)$.

For $n \in \mathbb{N}_+$,

$$\text{mse}(U_n \mid \mu) = \frac{n\sigma^2 b^4 + \sigma^4 (a - \mu)^2}{(\sigma^2 + n\, b^2)^2}, \quad \mu \in \mathbb{R} \tag{7.4.58}$$

The sequence of estimators $\boldsymbol{U} = (U_n : n \in \mathbb{N}_+)$ is mean-square consistent.

Proof

Recall that Y_n as variance $n\sigma^2$. Hence

$$\mse(U_n \mid \mu) = \var(U_n \mid \mu) + \bias^2(U_n \mid \mu) = \left(\frac{b^2}{nb^2 + \sigma^2}\right)^2 n\sigma^2 + \left(\frac{(a - \mu)\sigma^2}{nb^2 + \sigma^2}\right)^2 \tag{7.4.59}$$

Clearly $\mse(U_n \mid \mu) \to 0$ as $n \to \infty$ for every $\mu \in \R$.

When $b = \sigma$, $\mse(U \mid \mu) = [n\sigma^2 + (a - \mu)^2]/(n + 1)^2$. Recall that the method of moments estimator of μ and the maximum likelihood estimator of μ on $\R$ are both $M_n = Y_n / n$, the sample mean. This estimator is unbiased and has mean square error $\var(M) = \sigma^2 / n$. To see the connection between the estimators, note from (25) that

$$U_n = \frac{\sigma^2}{nb^2 + \sigma^2} a + \frac{nb^2}{nb^2 + \sigma^2} M_n \tag{7.4.60}$$

So U_n is a weighted average of a (the mean of the prior distribution) and M_n (the maximum likelihood estimator).

The Beta Distribution

Suppose that $\bs{X} = (X_1, X_2, \ldots)$ is a sequence of independent random variables each having the beta distribution with unknown left shape parameter $a \in (0, \infty)$ and right shape parameter $b = 1$. The beta distribution is widely used to model random proportions and probabilities and other variables that take values in bounded intervals (scaled to take values in $(0, 1)$). Recall that the probability density function (given a) is

$$g(x \mid a) = a\, x^{a-1}, \quad x \in (0, 1) \tag{7.4.61}$$

Suppose now that a is modeled by a random variable A that has a prior gamma distribution with shape parameter $k \in (0, \infty)$ and rate parameter $r \in (0, \infty)$. As usual, k and r are chosen to reflect our prior knowledge of a. Thus the prior probabiltiy density function of A is

$$h(a) = \frac{r^k}{\Gamma(k)} a^{k-1} e^{-ra}, \quad a \in (0, \infty) \tag{7.4.62}$$

The mean of the prior distribution is k/r.

The posterior distribution of A given $\bs{X}_n = (X_1, X_2, \ldots, X_n)$ is gamma, with shape parameter $k + n$ and rate parameter $r - \ln(X_1 X_2 \cdots X_n)$.

Proof

Fix $n \in \N_+$. Let $\bs{x} = (x_1, x_2, \ldots, x_n) \in (0, 1)^n$ and let $z = x_1 x_2 \cdots x_n$ Then

$$f(\bs{x} \mid a) = g(x_1 \mid a) g(x_2 \mid a) \cdots g(x_n \mid a) = a^n z^{a-1} = \frac{a^n}{z} e^{a \ln z} \tag{7.4.63}$$

Hence

$$h(a) f(\bs{x} \mid a) = \frac{r^k}{z \Gamma(k)} a^{n+k-1} e^{-a(r - \ln z)}, \quad a \in (0, \infty) \tag{7.4.64}$$

As a function of $a \in (0, \infty)$ this expression is proportional to the gamma PDF with shape parameter $n + k$ and scale parameter $r - \ln z$. Once again, it's not necessary to compute the normalizing constant $f(\bs{x})$.

Thus, the gamma distribution is conjugate to the beta distribution with unknown left parameter and right parameter 1. Note that the posterior shape parameter is deterministic and depends on the data only through the sample size n.

The Bayesian estimator of a based on $\bs{X}_n$ is

$$U_n = \frac{k + n}{r - \ln(X_1 X_2 \cdots X_n)} \tag{7.4.65}$$

Proof

The mean of the gamma distribution is the shape parameter divided by the rate parameter, so this follows from the previous theorem.

Given the complicated structure, the bias and mean square error of U_n given $a \in (0, \infty)$ would be difficult to compute explicitly. Recall that the maximum likelihood estimator of a is $W_n = -n / \ln(X_1 \, X_2 \cdots X_n)$. To see the connection between the estimators, note from (29) that

$$\frac{1}{U_n} = \frac{k}{k+n} \frac{r}{k} + \frac{n}{k+n} \frac{1}{W_n} \tag{7.4.66}$$

So $1/U_n$ (the reciprocal of the Bayesian estimator) is a weighted average of r/k (the reciprocal of the mean of the prior distribution) and $1/W_n$ (the reciprocal of the maximum likelihood estimator).

The Pareto Distribution

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots)$ is a sequence of independent random variables each having the Pareto distribution with unknown shape parameter $a \in (0, \infty)$ and scale parameter $b = 1$. The Pareto distribution is used to model certain financial variables and other variables with heavy-tailed distributions, and is named for Vilfredo Pareto. Recall that the probability density function (given a) is

$$g(x \mid a) = \frac{a}{x^{a+1}}, \quad x \in [1, \infty) \tag{7.4.67}$$

Suppose now that a is modeled by a random variable A that has a prior gamma distribution with shape parameter $k \in (0, \infty)$ and rate parameter $r \in (0, \infty)$. As usual, k and r are chosen to reflect our prior knowledge of a. Thus the prior probabiltiy density function of A is

$$h(a) = \frac{r^k}{\Gamma(k)} a^{k-1} e^{-ra}, \quad a \in (0, \infty) \tag{7.4.68}$$

For $n \in \mathbb{N}_+$, the posterior distribution of A given $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ is gamma, with shape parameter $k + n$ and rate parameter $r + \ln(X_1 X_2 \cdots X_n)$.

Proof

Fix $n \in \mathbb{N}_+$. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in [1, \infty)^n$ and let $z = x_1 x_2 \cdots x_n$ Then

$$f(\boldsymbol{x} \mid a) = g(x_1 \mid a) g(x_2 \mid a) \cdots g(x_n \mid a) = \frac{a^n}{z^{a+1}} = \frac{a^n}{z} e^{-a \ln z} \tag{7.4.69}$$

Hence

$$h(a) f(\boldsymbol{x} \mid a) = \frac{r^k}{z \Gamma(k)} a^{n+k-1} e^{-a(r + \ln z)}, \quad a \in (0, \infty) \tag{7.4.70}$$

As a function of $a \in (0, \infty)$ this expression is proportional to the gamma PDF with shape parameter $n + k$ and scale parameter $r + \ln z$. Once again, it's not necessary to compute the normalizing constant $f(\boldsymbol{x})$.

Thus, the gamma distribution is conjugate to Pareto distribution with unknown shape parameter. Note that the posterior shape parameter is deterministic and depends on the data only through the sample size n.

The Bayesian estimator of a based on $\boldsymbol{X}_n$ is

$$U_n = \frac{k + n}{r + \ln(X_1 X_2 \cdots X_n)} \tag{7.4.71}$$

Proof

Once again, the mean of the gamma distribution is the shape parameter divided by the rate parameter, so this follows from the previous theorem.

Given the complicated structure, the bias and mean square error of U given $a \in (0, \infty)$ would be difficult to compute explicitly. Recall that the maximum likelihood estimator of a is $W_n = n / \ln(X_1 \, X_2 \cdots X_n)$. To see the connection between the estimators, note from (31) that

$$\frac{1}{U_n} = \frac{k}{k+n} \frac{r}{k} + \frac{n}{k+n} \frac{1}{W_n} \tag{7.4.72}$$

So $1/U_n$ (the reciprocal of the Bayesian estimator) is a weighted average of r/k (the reciprocal of the mean of the prior distribution) and $1/W_n$ (the maximum likelihood estimator).
