SOME TOPICS IN BAYESIAN STATISTICS

1. Introduction

So far in the course, we only encountered the Bayesian viewpoint (prior and posterior probabilities) in testing a simple hypothesis against a simple alternative. More generally, we may have priors and posteriors for a continuous parameter θ.

2. Definition of priors and posteriors for a continuous θ

In this handout Θ will be a parameter space included in a Euclidean space $\mathbb{R}^k$. For example, for the family of normal distributions, Θ is the open half-plane $\{(\mu, \sigma) : -\infty < \mu < \infty,\ 0 < \sigma < \infty\} \subset \mathbb{R}^2$. On Θ, $d\theta$ will mean $d\theta_1 \cdots d\theta_k$.

Assume given a likelihood function $f(X, \theta)$ defined for $\theta \in \Theta$ and X a vector in $\mathbb{R}^n$. In Bayesian statistics, one assumes before taking any observations that θ has a prior probability density $\pi(\theta)$ with respect to $d\theta$. Then $\pi(\theta) \geq 0$ and $\int_\Theta \pi(\theta)d\theta = 1$. If $\theta = p$ with $0 \leq p \leq 1$ is the success probability in a binomial distribution, a simple and natural choice for its prior (in the absence of any particular information about p) is a $U[0, 1]$ distribution with $\pi(p) = 1$ for $0 \leq p \leq 1$. The earliest works in Bayesian statistics, Bayes (1764) and Laplace (1774), made this choice.

Let $f(x, \theta)$ be a likelihood function for one observation, which may be either a probability mass function if x is discrete or a density function if x is continuous. If we have i.i.d. observations $X = (X_1, ..., X_n)$ we get a likelihood function $f(X, \theta) = \prod_{j=1}^n f(X_j, \theta)$.

However $f(X, \theta)$ is obtained, the *posterior* density $\pi_X(\theta)$ is gotten by multiplying the likelihood function by the prior and then normalizing it,

$$\pi_X(\theta) = \frac{f(X, \theta)\pi(\theta)}{\int_\Theta f(X, \phi)\pi(\phi)d\phi}. \tag{1}$$

To show that (1) makes sense we can use the following:

Theorem 1. *Let $\Theta \subset \mathbb{R}^k$ be a parameter space. Let $\pi(\theta) \geq 0$ be a prior probability density for θ. Suppose that for each $\theta \in \Theta$, $f(X, \theta)$ is a probability density with respect to $X \in \mathbb{R}^n$, so that $\int f(X, \theta)dX = 1$*

Date: 18.650, Nov. 4, 2015.

where $dX = dx_1 dx_2 \cdots dx_n$. *Let* $q(X,\theta) = \pi(\theta)f(X,\theta)$ *for all* $\theta \in \Theta$ *and all* X. *Then*
(a) q *is a probability density for* (X,θ), $X \in \mathbb{R}^m$, $\theta \in \mathbb{R}^k$, *with respect to* $dXd\theta$, *for a joint probability distribution* Q *of* (X,θ),
(b) the marginal density of q *with respect to* θ *is* π,
(c) and for each $\theta \in \Theta$ *the conditional density of* X *given* θ *is* $q(X|\theta) = f(X,\theta)$.
(d) Letting

$$\tau(X) = \int_\Theta q(X,\theta)d\theta,$$

τ *is a probability density and is the marginal density of* Q *with respect to* X.
(e) With probability 1 with respect to Q, *or with respect to its marginal density* τ,

$$0 < \tau(X) < +\infty. \tag{2}$$

(f) For all X *such that (2) holds, a conditional density of* θ *given* X *(posterior density) exists and is given by* $q(\theta|X) = \pi_X(\theta)$ *in (1) where the denominator in (1) is* $\tau(X)$.
(g) We have for Q*-almost all* (x,θ),

$$q(X,\theta) = \pi(\theta)f(X,\theta) = \tau(X)\pi_X(\theta). \tag{3}$$

Proof. For (a), since the integrand is nonnegative we can do an iterated integral in either order. If we integrate first with respect to X we get $\pi(\theta)$ which has integral 1 with respect to θ. This also proves (b), and the rest of the statements are known facts about marginal and conditional densities from probability theory. Since part (e) is crucial in showing that π_X is well-defined with probability 1, let's prove it in detail, assuming part (d). We have

$$\Pr(\tau(X) = 0) = \int_{\tau(X)=0} \tau(X)\, dX = \int 0\, dX = 0.$$

On the other hand let $A = \{X : \ \tau(X) = +\infty\}$. Then

$$\Pr(A) = \int_A \tau(X)\, dX = \int_A +\infty\, dX = +\infty,$$

which is impossible since $\Pr(A) \le 1$, unless $\int_A 1 \cdot dX = 0$, in which case $\int_A +\infty\, dX$ is 0 by definition of (Lebesgue) integral, so $\Pr(A) = 0$, and (e) follows, i.e. (2) holds with probability 1.

For part (g), in (3), the first equation holds by definition of $q(X,\theta)$, and the second by the definitions and parts (d) and (e). $\square$

3. Bayes least-squares estimation

First here is a very simple, probably familiar, fact.

Proposition 1. *For any random variable Y with $E(Y^2) < +\infty$, the unique constant c that minimizes $E((Y-c)^2)$ is $c = EY$.*

Proof. $E((Y-c)^2) = E(Y^2) - 2cEY + c^2$ is a quadratic polynomial in c which goes to $+\infty$ as $c \to \pm\infty$, so it's minimized where its derivative with respect to c equals 0, namely at $c = EY$. $\square$

Suppose we want to estimate a function $g(\theta)$. Then for an estimator $V(X)$, the mean-square error (MSE) for a given θ is $E_\theta[(V(X)-g(\theta))^2]$. For a prior π, the *risk* is the expectation of the MSE with respect to that prior, namely

$$r(V,\pi) := \int_\Theta E_\theta[(V(X) - g(\theta))^2]\pi(\theta)d\theta. \tag{4}$$

A *Bayes* estimator for $g(\theta)$ for the given prior is one that minimizes the risk, provided its risk is finite.

Theorem 2. *For a given likelihood function $f(X,\theta)$ for $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^k$ for some $k \geq 1$, and prior density π, if there exists some estimator $U(X)$ of the given $g(\theta)$ that has finite risk for the given π, then there exists a Bayes estimator T, given by the expectation of $g(\theta)$ with respect to the posterior distribution,*

$$T(X) \;=\; \int_\Theta g(\theta)\pi_X(\theta)d\theta. \tag{5}$$

The Bayes estimator is essentially unique, in the sense that any Bayes estimator must equal this $T(X)$ with probability 1.

Proof. We have

$$r(U,\pi) = \int_\Theta \int [(U(X) - g(\theta))^2 f(X,\theta)dX]\pi(\theta)d\theta < \infty.$$

Interchanging integrals for a nonnegative integrand, this gives by (3)

$$\int\int [(g(\theta) - U(X))^2 \pi_X(\theta)d\theta\, \tau(X)dX < \infty.$$

It follows that with probability 1 with respect to the marginal distribution $\tau(X)dX$ of X,

$$\int ((g(\theta) - U(X))^2 \pi_X(\theta)d\theta < \infty.$$

Since $U(X)$ is constant with respect to θ, and a constant plus a square-integrable function is square-integrable, it further follows that with probability 1 with respect to $\tau(X)dX$, g is square-integrable with respect to the posterior distribution π_X:

$$\int g(\theta)^2 \pi_X(\theta) d\theta < \infty. \tag{6}$$

We would like to minimize (4). Let's write out the E_θ. Recall that $\int \cdots dX$ is a shorthand for

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \cdots dx_1\, dx_2 \cdots dx_n,$$

where the integral(s) are replaced by sums in case X is discrete. The method of proof is essentially the same. Then (4) becomes

$$\int_\Theta \int [(V(X) - g(\theta))^2] f(X, \theta)\, dX\, \pi(\theta) d\theta. \tag{7}$$

Since the integrand is nonnegative and the integrals are well-defined (possibly infinite) we can interchange the two integrals, and (7) becomes

$$\int \int_\Theta [(V(X) - g(\theta))^2] f(X, \theta) \pi(\theta) d\theta\, dX. \tag{8}$$

Then applying (3), the factor $\tau(X)$ doesn't depend on θ so we can take it outside the integral with respect to θ, and (8) becomes

$$\int \int_\Theta [(V(X) - g(\theta))^2] \pi_X(\theta) d\theta\, \tau(X)\, dX. \tag{9}$$

In the inner integral with respect to θ in (9), X is fixed and $g(\theta)$ is a random variable with respect to the posterior density $\pi_X(\theta)$, having finite mean-square by (6). To minimize this inner integral we need to choose $V(X)$, which would be constant for fixed X. By Proposition 1, the correct constant is given by $V(X) = T(X)$ in (5). Since the risk is finite for an estimator U by assumption, the minimum risk must be finite, so $T(X)$ in (5) indeed gives a Bayes estimator. The essential uniqueness follows from the uniqueness in Proposition 1. □

In case of a Gamma(a, c) prior density for a Poisson parameter λ, where the posterior density will also be in the gamma family, the expectation of λ for the posterior density is easy to calculate, see below under "Conjugate priors." Similarly, we have an easy calculation for the posterior expectation of a binomial parameter p using a Beta(a, b) prior.

Some texts give a different formulation of Theorem 2 in which they say that the Bayes estimator is the conditional expectation of $g(\theta)$ given

X, $T(X) = E(g(\theta)|X)$. That is correct in case $\int |g(\theta)|\pi(\theta)d\theta < +\infty$ but integrals with respect to the posterior distributions may be finite even if they are not with respect to the prior. There is more about that in Section 2.6 of the 18.466 OCW notes, but that section is far from self-contained, so otherwise it isn't recommended reading in this course.

4. Admissibility

Recall that a statistic $T(X)$ is said to be *inadmissible* as an estimator of a function $g(\theta)$ of a parameter θ if there exists another estimator $V(X)$ such that $E_\theta((V(X)-g(\theta))^2) \le E_\theta((T(X)-g(\theta))^2)$ for all θ and $E_\theta((V(X)-g(\theta))^2) < E_\theta((T(X)-g(\theta))^2)$ for some θ. Then $T(X)$ is *admissible* if it is not inadmissible. Let's call $T(X)$ *strongly inadmissible* if we add to the the definition that $E_\theta[(V(X)-g(\theta))^2] < E_\theta[(T(X)-g(\theta))^2]$ for all θ in a non-empty open set U, namely, a set such that: for some θ_0 in U and $r > 0$, also θ is in U for all θ such that $|\theta - \theta_0| < r$. In one dimension this would just say that U includes a non-degenerate interval.

If π is a prior density with $\pi(\theta) > 0$ for almost all θ, i.e. if A is the set of θ for which $\pi(\theta) = 0$, then $\int 1_A(\theta)d\theta = 0$, and if T is a Bayes estimator for $g(\theta)$, namely the integral of $g(\theta)$ times the posterior density $\pi_X(\theta)$, then T cannot be strongly inadmissible, or there would be an estimator with smaller overall risk (integrating mean-square error times $\pi(\theta)$), contradicting the Bayes property of T.

Recall that when we considered unbiased estimation earlier in the course, in the handout on mean-squred error, it was pointed out that the usual unbiased estimator

$$s_X^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$$

of the variance σ^2 for $n \ge 2$ is inadmissible for any i.i.d. X_j with $\sigma > 0$ and $E(X_1^4) < +\infty$, using $V(X) = \frac{(n+2)(n-1)}{n(n+1)} s_X^2$ (Yatracos's estimator). In fact s_X^2 is strongly inadmissible as just defined.

5. Unbiasedness

The Bayes property for squared-error loss turns out to be virtually incompatible with unbiasedness. Let's begin with two examples.

Example 1. For $0 \le \theta \le 1$ let δ_θ be the point mass at θ. Suppose θ is unknown in advance and has prior density $U[0,1]$. Suppose that the true $\theta = \theta_0$ for some θ_0. If we have even one observation X_1 from δ_θ, then $\Pr(X_1 = \theta_0) = 1$. So X_1 is both an unbiased estimator and a Bayes estimator of θ for the given prior (or any prior on $[0,1]$). That shows that this combination of properties of an estimator is possible, but it's a rather extreme and impractical case.

Example 2. Let p be the success probability in a binomial(n,p) distribution, and let $\pi(p) > 0$ for $0 < p < 1$ be a prior density for p. Then for any observation X, an integer with $0 \le X \le n$, the likelihood function is proportional to $p^X(1-p)^{n-X}$, which is a bounded function of p, and a posterior density $\pi_X(p)$ defined by (1) exists. Moreover, since p is bounded, the integral $T(X) = \int_0^1 p\pi_X(p)dp$ is always finite, in fact satisfies $0 \le T(X) \le 1$, and so has finite risk as an estimator of p with squared-error loss. So by Theorem 2, T is a Bayes estimator for p for the given π. Now suppose the true $p = 0$. Then we will have $\Pr(X = 0) = 1$ and the likelihood function will be $(1-p)^n$. Then $\pi_X(p) > 0$ for $0 < p < 1$ because $\pi(p) > 0$, and so $\Pr(T(X) > 0) = 1$ and $E_0T(X) > 0$, so T is not an unbiased estimator of p.

Similarly, when the true $p = 1$, $E_1(T(X))$ will be less than 1. For $0 < p_0 < 1$ there do exist priors π of p such that for the Bayes estimators $T(X)$ for π, $E_{p_0}T(X) = p_0$. But it is not possible to find π such that for the Bayes estimator $T(X)$ for π, $E_p(T(X)) = p$ for all p with $0 < p < 1$, as a special case of the following theorem.

Theorem 3. *Let $f(X,\theta)$, $\theta \in \Theta$, be a parametric family of densities for X in n-dimensional Euclidean space $\mathbb{R}^n$, with respect to $dX = dx_1dx_2\cdots dx_n$, where θ is in any parameter space Θ included in a Euclidean space $\mathbb{R}^k$. For any prior density π on Θ and real-valued function g on Θ which is a random variable with respect to π, an unbiased estimator T of g is Bayes for π and squared-error loss if and only if it has risk $r(T,\pi) = 0$, so that $T(x) = g(\theta)$ with Q-probability 1.*

Remark. The theorem shows that an estimator can be both Bayes and unbiased only when $g(\theta)$ can be estimated exactly without error, as in Example 1.

Proof. "If" is clear. To prove "only if," by definition of Bayes estimator, T must have finite risk, and by the proof of Theorem 2, with probability 1 in X, $\int g(\theta)^2\pi_X(\theta)d\theta < \infty$. Let τ be the marginal density of Q for X given by Theorem 1(d). By (3), $q(X,\theta) = \pi_X(\theta)\tau(X)$ with probability 1. We have

$$\begin{aligned} r(T,\pi) &= \int\int (T(X)-g(\theta))^2 q(X,\theta)dX\,d\theta \\ &= \int\int T(X)^2 - 2T(X)g(\theta) + g(\theta)^2\,\pi_X(\theta)\tau(X)dXd\theta. \qquad (10) \end{aligned}$$

As the integrand in (10) is nonnegative we can do the integral in either order. The proof will work by finding two different expressions of the integral of the cross term $-2T(X)g(\theta)$, in (11) and (13), which together will give (14). First, by the Bayes property and equation (5),

$$T(X) = \int g(\theta)\pi_X(\theta)d\theta.$$

Doing the integral in (10) in the order $d\theta\,dX$ and doing the integral with respect to θ of the cross term, X is fixed and we get $-2T(X)\tau(X)T(X) = -2T(X)^2\tau(X)$. It then follows that

$$(11)\quad r(T,\pi) = \int \left[T(X)^2 - 2T(X)^2 + \int g(\theta)^2\,\pi_X(\theta)d\theta\right]\tau(X)dX.$$

Since $r(T,\pi) < +\infty$, and for fixed X, $-T(X)^2$ is also fixed, we have $\int g(\theta)^2\pi_X(\theta)d\theta < +\infty$ for τ-almost all X, and

$$(12)\qquad r(T,\pi) = \int\int \left[g(\theta)^2 - T(X)^2\right] q(X,\theta)d\theta\,dX.$$

On the other hand, doing the integral in (10) in the stated order, we know by unbiasedness that for fixed θ, $E_\theta T(X) = \int T(X)f(X,\theta)dX = g(\theta)$. By (3), $f(X,\theta)\pi(\theta) \equiv \pi_X(\theta)\tau(X)$. As θ is fixed in the inner integral dX, we can take $\pi(\theta)$ outside the integral. We then have

$$\int -2g(\theta)T(X)f(X,\theta)dX = -2g(\theta)^2$$

and so

$$(13)\quad r(T,\pi) = \int\left[\int T(X)^2 f(X,\theta)dX - 2g(\theta)^2 + g(\theta)^2\right]\pi(\theta)d\theta.$$

Next, $r(T,\pi) < +\infty$ implies $\int T(X)^2 f(X,\theta)dX < +\infty$ for π-almost all θ, and

$$(14)\qquad r(T,\pi) = \int \left[T(X)^2 - g(\theta)^2\right] q(X,\theta)dX\,d\theta = -r(T,\pi)$$

from (12), so $r(T,\pi) = 0$, finishing the proof. $\square$

6. Conjugate priors

A conjugate prior for a given parametric family of distributions with a likelihood function is one such that the posterior distributions all belong to the same parametric family. For example, if $\theta = \lambda$ is a Poisson parameter with $0 < \lambda < +\infty$ and the prior $\pi(\theta)$ is a gamma density, then the posterior $\pi_X(\theta)$ is also in the gamma family. Specifically, if λ has prior density Gamma(a, c), where $a > 0$ and $c > 0$, so that for $\lambda > 0$, $\pi(\lambda) = c^a \lambda^{a-1} \exp(-c\lambda)/\Gamma(a)$, and we observe $X_1, ..., X_n$ i.i.d. Poisson(λ) with $S_n := X_1 + \cdots + X_n$, then the likelihood function is proportional to $e^{-n\lambda}\lambda^{S_n}$ and so the posterior density is Gamma$(a + S_n, c+n)$ (it is proportional to this as a function of λ, and a probability density has a unique normalizing constant). As the expectation for $\Gamma(a, c)$ is a/c, the expectation of λ for the posterior distribution (the Bayes estimate of λ by Theorem 2) is $\frac{S_n+a}{n+c}$. With probability 1, this is asymptotic as $n \to \infty$ to the maximum likelihood estimate $\overline{X}_n = S_n/n$, in other words

$$\frac{S_n + a}{n + c} \cdot \frac{n}{S_n} \to 1,$$

because $n/(n + c) \to 1$ and $(S_n + a)/S_n = 1 + (a/S_n) \to 1$ because by the law of large numbers, $S_n/n \to \lambda$ and so $S_n \sim n\lambda \to +\infty$.

Likewise, beta densities give conjugate priors for the binomial probability p, as will be seen in a problem in PS10.

6.1. Normal-inverse-gamma distributions; conjugate for normals.

Let $Y > 0$ be a random variable having a distribution function F and a density $f = f_Y$. Let $V := 1/Y$. Then for any $x > 0$, $\Pr(V \le x) = \Pr(1/Y \le x) = \Pr(Y \ge 1/x) = 1-F(1/x)$, and so by the chain rule $1/Y$ has a density $f_{1/Y}(x) = -f(1/x)\cdot(-1/x^2) = x^{-2}f(1/x)$. Thus if Y has a Gamma(α, β) density $f(y) = \beta^\alpha y^{\alpha-1} \exp(-\beta y)/\Gamma(\alpha)$ then $1/Y$ has the density

$$\beta^\alpha y^{1-\alpha} \exp(-\beta/y)/(y^2\Gamma(\alpha)) = \beta^\alpha y^{-1-\alpha} \exp(-\beta/y)/\Gamma(\alpha)$$

where $\alpha > 0$, $\beta > 0$, and $y > 0$, which is called an inverse gamma (α, β) density.

Parameters of prior or posterior distributions are called hyperparameters. The family of all normal distributions $N(\mu, \sigma^2)$ on the real line has a conjugate prior for the parameter $\theta = (\mu, \sigma)$ called the "normal-inverse-gamma distribution" and given by

$$(15)\quad \pi_{\alpha,\beta,\nu,\lambda}(\mu, \sigma) = \frac{\sqrt{\nu}}{\sigma\sqrt{2\pi}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left(-\frac{2\beta + \nu(\mu - \lambda)^2}{2\sigma^2}\right)$$

with four hyperparameters $\alpha > 0$, $\beta > 0$, $\nu > 0$, and $\lambda \in \mathbb{R}$. For (15), given the hyperparameters, the marginal density of σ^2 is inverse gamma (α, β), or equivalently $1/\sigma^2$ has Gamma(α, β), and the conditional density of μ given σ is $N(\lambda, \sigma^2/\nu)$. If σ is fixed, then the normal distributions give a conjugate prior family for μ, which is much simpler, but it's usually unrealistic to assume σ is known. Likewise if μ is fixed, the gamma distributions for $1/\sigma^2$ give a conjugate prior family, but for μ to be fixed is also usually unrealistic. For the joint conjugate prior density (15) of μ and σ^2, μ and σ^2 are not independent: the density is not a product $f(\mu)$ times $g(\sigma)$ for any functions f and g. So the joint conjugate prior is a bit complicated. There seem to be no rules for choosing hyperparameters. It is not proved here that it is actually a conjugate prior family for normals.

Why do Bayes estimation of normal parameters?

If one observes $X_1, ..., X_n$ assumed to be i.i.d. $N(\mu, \sigma^2)$ for some unknown μ and some $\sigma > 0$, then the sample mean $\overline{X}$ is a natural and usual estimator of μ and is the MLE of μ. It cannot be the Bayes estimator for any non-trivial prior distribution for μ, because it is unbiased (Theorem 3).

For σ^2, we've considered estimators $c_n \sum_{j=1}^n (X_j - \overline{X})^2$, which are unbiased for $c_n = 1/(n-1)$, MLEs for $c_n = 1/n$, and minimize mean-squared error for $c_n = 1/(n+1)$. The three factors c_n differ from each other by amounts of order $O(1/n^2)$, which is relatively small for n large. The unbiased choice $c_n = 1/(n-1)$ gives an inadmissible estimator: Yatracos showed that $c_n = (n+2)/[n(n+1)]$ gives smaller mean-square error, not only for normal distributions but for all distributions with $E(X_1^4) < \infty$ and variance $\sigma^2 > 0$.

In Section 2.7 of the 18.466 (2003) OCW notes, it's pointed out that $\overline{X}$ is an admissible estimator of μ. Likewise for 2-dimensional i.i.d. random vectors X_i with distribution $N(\mu, I)$ where I is the identity covariance matrix, the sample mean vector $\overline{X}$ is still an admissible estimator of the mean vector μ. But in dimension $d \geq 3$, $\overline{X}$ is no longer admissible, as Charles Stein proved in 1956, with a proof given in Lehmann (1983, 1991; Lehmann and Casella, 1998) and also in the OCW notes. The inadmissibility for $d \geq 3$ is sometimes called "Stein's phenomenon." Bayes estimators $T(\cdot)$ are generally admissible, as noted above, but for the normal μ they may not be "minimax," specifically, for large $|\mu|$, $E(|T(X) - \mu|^2)$ may be substantially larger than $E(|\overline{X} - \mu|^2) = d$.

In applied statistics, it seems that the sample mean vector $\overline{X}$ continues to be used unless there is a reason to think that it may be influenced by outliers.

7. Credible intervals

These are the Bayesian counterparts of confidence intervals. A $100(1-\alpha)\%$ credible interval for a real parameter θ is one that has posterior probability $1-\alpha$ of containing θ. A two-sided 95% credible interval for θ, for example, would be the interval with endpoints the 0.025 and 0.975 quantiles of the posterior distribution. For any family of distributions which R handles, one can compute quantiles by "qfamily(β, θ)" where θ is a parameter or a list of parameters and we want the β quantile, for example qbeta$(0.025, a, b)$ would give the 0.025 quantile of a Beta(a, b) distribution. There are cases where a maximum likelihood estimate (MLE) is unbiased, as with the sample mean $\overline{X}$ for the normal mean μ or the Poisson parameter λ. In such cases, typically a Bayes estimator will be somewhere between the MLE and the mean of the prior distribution, becoming asymptotic to the MLE as $n \to \infty$.

8. Historical Notes

These notes are based on Stigler (1986), pp. 359-362. The field of "Bayesian" statistics is named for Thomas Bayes, who wrote a paper about the method in 1764. But the work by the leading mathematician and scientist Laplace (1774) attracted more attention. Stigler (p. 361) wrote that Bayes's article "was ignored until after 1780 and played no important role in scientific debate until the twentieth century." Laplace used the $U[0, 1]$ prior distribution for a binomial parameter p and noted that the posterior distributions are beta distributions. Stigler (p. 359) writes "we can be reasonably certain Laplace was unaware of Bayes's earlier work."

REFERENCES

Bayes, T. (1764), An essay toward solving a problem in the doctrine of chances, *Philos. Trans. Roy. Soc. London* **53**, 370-418; repr. in *Biometrika* **45** (1958) 293-315.

Dudley, R. M. (2003). *Mathematical Statistics*, 18.466 lecture notes, Spring 2003. On MIT OCW (OpenCourseWare) website, 2004.

Laplace, P. S. (1774), "Memoir on the probability of the causes of events," *Mémoires de Math. et de Physique, Présentés à l'Académie*

Royale des Sciences, par divers Savans & lûs dans ses Assemblées, **6**, 621-646; transl. by S. M. Stigler in Stigler (1986), pp. 364–378.

Lehmann, Erich, *Theory of Point Estimation*; originally published by Wiley (1983), then by Wadsworth and Brooks/Cole (1991); Second Edition by E. L. Lehmann and G. Casella, Springer, 1998. (It seems that the 1991 version was a reprinting, not a new edition.)

Stigler, S. M. (1986), Laplace's 1774 memoir on inverse probability, *Statistical Science* **1**, 359-378.