Bayesian Estimation for the Proportions
in a Mixture of Distributions

Javad Behboodian*

Department of Statistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 733

January, 1971
BAYESIAN ESTIMATION FOR THE PROPORTIONS IN A MIXTURE OF DISTRIBUTIONS

Javad Behboodian
University of North Carolina, Chapel Hill
and Pahlavi University, Shiraz, Iran

ABSTRACT

The joint density of a random sample from a mixture of two distributions is expressed as a binomial mixture of conditional densities. Then, the posterior distribution of the proportion, in a mixture of two known and distinct distributions, relative to a beta prior distribution is explored in detail and a conjugate family of prior distributions is introduced. The results are generalized for estimating the proportions in a finite mixture of known distributions.

1. INTRODUCTION

Consider

\[ m(x) = pf(x) + qg(x) \]

where \( f(x) \) and \( g(x) \) are two known and distinct probability density functions with \( 0 < p < 1 \) and \( q = 1-p \). Since \( f(x) \) and \( g(x) \) are distinct from each other, \( m(x) \) is identifiable, i.e., corresponding to two different values of \( p \) we have two different mixtures. The estimation problem of proportions in a finite mixture of distributions has been already investigated by Boes. In [1], he
suggests a class of unbiased estimators which converge with probability 1, and in [2] he gives minimax unbiased estimators for proportions. Malëva [4] introduces moment estimators for proportions, and shows that they are unbiased and asymptotically normal. As far as I know, a Bayesian analysis of this problem has not been so far explored. Here we look at the problem from a subjectivistic Bayesian point of view by considering the experimenter's information about the proportion \( p \) prior to taking a sample. Such information is usually expressed by a prior distribution for \( p \), which reflects subjective beliefs or knowledge about \( p \). The prior distribution is modified, by using the sample information, in accordance with Bayes theorem to yield a posterior distribution of \( p \). A Bayesian analysis of \( p \) would consist in the exploration and interpretation of the posterior distribution. Examples of such analyses are available in many of the references provided by L. J. Savage [5].

It is known that the posterior distribution is proportional to the product of the likelihood function and prior density. Before introducing a prior distribution for \( p \), we first write its likelihood function in a convenient form.

2. THE LIKELIHOOD FUNCTION

Let \( X_1, X_2, \ldots, X_n \) be a random sample from a population with probability density function (1.1). The likelihood function of \( p \), for \( 0 < p < 1 \) and with \( x_1, x_2, \ldots, x_n \) as the experimental values of the random sample, is

\[
L(p) = \prod_{i=1}^{n} [pf(x_i) + qg(x_i)].
\]  

(2.1)

The contribution of any \( x_i \), for which \( f(x_i) = g(x_i) \), to \( L(p) \) does not depend on \( p \), and we can eliminate such non-informative \( x_i \) from our observed sample without any loss. It is quite possible to have \( g(x_i) = f(x_i) \) for
\[ i = 1, 2, \ldots, n. \] For example, this may happen in the case of a mixture of two uniform distributions defined on two different overlapping intervals of equal length. However, due to distinction of \( f(x) \) and \( g(x) \), the chance of such event becomes small as \( n \) becomes large. Therefore, to avoid exceptional cases, we assume that \( f(x_1) \neq g(x_1) \) for all \( x_1 \)'s from the beginning. The logarithm of \( L(p) \) is a concave function of \( p \) with at most one local maximum. It can be shown that, with only light regularity conditions on the mixed densities \( f(x) \) and \( g(x) \), \( L(p) \) has a unique maximum at \( \hat{p} \) in the interval \( (0,1) \) when \( n \) is sufficiently large, with the usual desirable large sample properties. However, for small samples \( \hat{p} \) may be a poor estimate.

To write \( L(p) \) in a suitable form, we can expand the right side of (2.1). But, it is more useful to apply the following probabilistic argument: Let us denote the right side of (2.1), which is in fact the joint density of the random sample \( X_1, X_2, \ldots, X_n \), by \( h(x_1, x_2, \ldots, x_n) \). Using conditional density, we have

\[
h(x_1, x_2, \ldots, x_n) = \sum_{k=0}^{n} h(x_1, x_2, \ldots, x_n | E_k) P(E_k),
\]

(2.2)

where \( E_k \) is the event that exactly \( k \) of the \( X_i \)'s have density \( f(x) \) and the rest have density \( g(x) \). It is clear that

\[
P(E_k) = \binom{n}{k} p^k q^{n-k}.
\]

(2.3)

The event \( E_k \) can happen in \( \binom{n}{k} \) equally likely ways depending on the partition of the random sample. The joint density of \( X_1, X_2, \ldots, X_n \) corresponding to a particular partition is

\[
h_{t_k}(x_1, x_2, \ldots, x_n) = \prod_{a \in A_k} f(x_a) \prod_{b \in B_k} g(x_b),
\]

(2.4)

where \( t_k \) is a partition of the set \( \{1, 2, \ldots, n\} \) into two sets \( A_k \) and \( B_k \) with \( k \) elements in \( A_k \). Denoting the set of all such partitions by \( T_k \) and using conditional density once more, we obtain
\[ S_k(x_1, x_2, \ldots, x_n) = h(x_1, x_2, \ldots, x_n | E_k) = \sum_{t_k \in T_k} h_t(x_1, x_2, \ldots, x_n) \binom{n}{k}, \quad (2.5) \]

where \( S_k(x_1, x_2, \ldots, x_n) \) is a symmetric \( n \)-variate density. Now, from (2.2), (2.3) and (2.5), we have

\[ L(p) = h(x_1, x_2, \ldots, x_n) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} S_k(x_1, x_2, \ldots, x_n). \quad (2.6) \]

Therefore, the joint density of the random sample \( x_1, x_2, \ldots, x_n \) is a binomial mixture of the densities \( S_k(x_1, x_2, \ldots, x_n) \) defined by (2.4) and (2.5).

For example, if \( n = 3 \), then we have four 3-variate densities:

\[
S_0(x_1, x_2, x_3) = g(x_1) g(x_2) g(x_3)
\]
\[
S_1(x_1, x_2, x_3) = [f(x_1) g(x_2) g(x_3) + f(x_2) g(x_1) g(x_3) + f(x_3) g(x_1) g(x_2)]/3
\]
\[
S_2(x_1, x_2, x_3) = [f(x_1) f(x_2) g(x_3) + f(x_2) f(x_3) g(x_1) + f(x_3) f(x_1) g(x_2)]/3
\]
\[
S_3(x_1, x_2, x_3) = f(x_1) f(x_2) f(x_3).
\]

3. **Bayesian Estimation for \( p \)**

Theoretically \( p \) can assume any real number in the interval \((0, 1)\). Suppose the experimenter expresses his information and belief about \( p \), which is now considered as a random variable, by a beta density of the form

\[
\beta(p; u, v) = \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)} p^{u-1}(1-p)^{v-1} \quad (3.1)
\]

for \( 0 < p < 1 \), and zero elsewhere, with parameters \( u > 0 \) and \( v > 0 \). Now, the posterior density of \( p \) becomes

\[
\Pi(p|x_1, x_2, \ldots, x_n) \propto L(p) \beta(p; u, v), \quad (3.2)
\]

where \( \propto \) denotes proportionality. By using (2.6) and (3.1) and omitting the
multiplier of (3.1) which does not involve \( p \), we obtain

\[
\Pi(p|x_1, x_2, \ldots, x_n) = \sum_{k=0}^{n} \binom{n}{k} S_k(x_1, x_2, \ldots, x_n) p^{k+v-1} q^{n-k-v-1} .
\]  

(3.3)

The missing constant of proportionality can be easily found from the fact that the posterior density must integrate to one. Simple calculation shows that

\[
\Pi(p|x_1, x_2, \ldots, x_n) = \sum_{k=0}^{n} w_k \beta(p; k+u, n-k+v),
\]  

(3.4)

where \( \beta(p; k+u, n-k+v) \) is a beta density with parameters \( k+u \) and \( n-k+v \), and the weights \( w_k \) are obtained from

\[
w_k = \gamma_k S_k(x_1, x_2, \ldots, x_n) \sum_{j=0}^{n} \gamma_j S_j(x_1, x_2, \ldots, x_n)
\]  

(3.5)

with

\[
\gamma_k = \binom{n}{k} \frac{n+u+v-1}{k+u-1}.
\]  

(3.6)

The weights \( w_k \) depend on the parameters of the prior and on the observed samples; the contribution of the prior is reflected in \( \gamma_k \) and the contribution of the data in \( S_k(x_1, x_2, \ldots, x_n) \). Thus, we have the following result:

Let \( x_1, x_2, \ldots, x_n \) be the experimental values of a random sample from a mixture of two unknown and distinct distributions with proportion parameter \( p \), where the prior distribution of \( p \) is a beta distribution with parameter \( u \) and \( v \). Then, the posterior distribution of \( p \) is a mixture of \( n+1 \) beta distributions with parameters \( k+u \) and \( n-k+v \) for \( k = 0, 1, \ldots, n \).

Simple calculation, by using (3.4), the mean and variance of a beta distribution, and the formulae for the mean and variance of a mixture [3], shows that the posterior mean and the posterior variance of \( p \) are

\[
E_{\Pi}(p) = \sum_{k=0}^{n} w_k \frac{k+u}{n+u+v}.
\]  

(3.7)

\[
\text{var}_{\Pi}(p) = \sum_{k=0}^{n} \frac{(k+u)(n-k+v)}{(n+u+v)^2(n+u+v+1)} + \sum_{k=0}^{n} w_k \left[ \frac{k+u}{n+u+v} - E_{\Pi}(p) \right]^2.
\]  

(3.8)
These results show that the posterior mean takes values in \((0,1)\), and the posterior variance becomes small as \(n\) becomes large.

We know that when \(u = v = 1\), the beta distribution is a uniform distribution over the unit interval. This might be used to represent a diffuse state of prior knowledge about \(p\). In this case, the maximum likelihood estimate \(\hat{p}\) is the posterior mode and in large samples it is usually near to the posterior mean \(\sum_{k=0}^{n} w_k (k+1)/(n+2)\). On the other hand, if the experimenter believes that the population with density \(f(x)\) is slightly contaminated by the population with density \(g(x)\), i.e., \(p\) is close to 1, then he can express his view by a beta prior with large \(u\) and small \(v\). But it follows from (3.6) that for sufficiently large \(u\) and small \(v\) the coefficient \(\gamma_k\) is an increasing function of \(k\) since

\[
\frac{\gamma_{k+1}}{\gamma_k} = \frac{(k+1)(k+u)}{(n-k)(n-k+v)}. \tag{3.9}
\]

The meaning of the above result can be expressed in the following manner:

When we have a strong opinion about the closeness of \(p\) to 1, the \(\gamma_k\) associated with \(S_k(x_1,x_2,\ldots,x_n)\), the conditional density that exactly \(k\) of the \(x_i\)'s come from the population with density \(f(x)\), becomes larger as \(k\) increases. However, the effect of data and mixed densities which is reflected in \(S_k(x_1,x_2,\ldots,x_n)\), a factor of the weight \(w_k\), may confirm or reject our prior opinion about \(p\).

Actually, the beta family of distributions includes symmetric, skewed, unimodal, U-shaped and J-shaped distributions. But we can have a wider variety of distributions for representing a person's knowledge about \(p\) by using a family of finite mixtures of beta distributions. It is interesting to observe, by an analysis similar to that of Section 3, that if we take a member of this family for the prior density of \(p\), then the posterior density will be again a member.
of the same family. Therefore, a family of finite mixtures of beta distributions is conjugate with respect to the likelihood function (2.7).

4. Estimation in a Finite Mixture of Distributions

Consider, for \( N \geq 2 \),

\[
m(x) = \sum_{i=1}^{N+1} p_i f_i(x)
\]

(4.1)

where \( f_1(x), f_2(x), \ldots, f_{N+1}(x) \) are \( N+1 \) known and distinct probability density functions with \( p_i > 0 \), \( i = 1, 2, \ldots, N+1 \), and \( \sum_{i=1}^{N+1} p_i = 1 \). \( m(x) \) is the probability density function of \( N+1 \) distributions. Since the \( p_i \)'s are linearly dependent, the unknown parameters can be assumed to be \( p_1, p_2, \ldots, p_N \). Moreover, it is assumed that \( m(x) \) is identifiable, i.e., corresponding to different vectors \( (p_1, p_2, \ldots, p_N) \) we obtain different mixtures. As before, let \( x_1, x_2, \ldots, x_n \) be the experimental values of a random sample from a population with probability density function (4.1) and \( L(p_1, p_2, \ldots, p_N) \) the likelihood function of the unknown parameters \( p_1, p_2, \ldots, p_N \).

As an \( N \)-variate analogue of (2.6), we can easily show that the joint density of the random sample is a multinomial mixture of the densities \( S_{k_1, k_2, \ldots, k_N}(x_1, x_2, \ldots, x_n) \), which is the conditional density that \( k_1 \) of the \( X_1 \)'s have density \( f_1(x) \), \( k_2 \) of the \( X_1 \)'s have density \( f_2(x) \), and so on. For the joint prior density of \( p_1, p_2, \ldots, p_N \), we can take the \( N \)-variate analogue of the beta density (3.1), which is the \( N \)-variate Dirichlet density [6]. Simple calculation shows that the joint posterior density of \( p_1, p_2, \ldots, p_N \) becomes a mixture of \( n+1 \) Dirichlet densities. Similarly, we observe that a family of finite mixtures of Dirichlet distributions is conjugate with respect to the likelihood function \( L(p_1, p_2, \ldots, p_N) \).
REFERENCES


