 Journal of Business & Economic Statistics, July 1987, Vol. 5, No. 3

Bayesian Estimation and Prediction for the Beta-Binomial Model

Jack C. Lee and Darius J. Sabavala
Bell Communications Research, Inc., Morristown, NJ 07960-1961

The beta-binomial distribution, which is generated by a simple mixture model, has been widely applied in the social, physical, and health sciences. Problems of estimation, inference, and prediction have been addressed in the past, but not in a Bayesian framework. This article develops Bayesian procedures for the beta-binomial model and, using a suitable reparameterization, establishes a conjugate-type property for a beta family of priors. The transformed parameters have interesting interpretations, especially in marketing applications, and are likely to be more stable. More specifically, one of these parameters is the market share and the other is a measure of the heterogeneity of the customer population. Analytical results are developed for the posterior and prediction quantities, although the numerical evaluation is not trivial. Since the posterior moments are more easily calculated, we also propose the use of posterior approximation using the Pearson system. A particular case (when there are two trials), which occurs in taste testing, brand choice, media exposure, and some epidemiological applications, is analyzed in detail. Simulated and real data are used to demonstrate the feasibility of the calculations. The simulation results effectively demonstrate the superiority of Bayesian estimators, particularly in small samples, even with uniform ("non-informed") priors. Naturally, "informed" priors can give even better results. The real data on television viewing behavior are used to illustrate the prediction results. In our analysis, several problems with the maximum likelihood estimators are encountered. The superior properties and performance of the Bayesian estimators and the excellent approximation results are strong indications that our results will be potentially of high value in small sample applications of the beta-binomial and in cases in which significant prior information exists.

KEY WORDS: Conjugate-type priors; Marketing applications; Pearson Type I approximation; Reparameterization.

1. INTRODUCTION

The beta-binomial model is one of the oldest discrete probability mixture models and is widely applied in the social, physical and health sciences. The model was formally proposed by Skellam (1948), although the idea was suggested earlier by Pearson (1925) in an experimental investigation of Bayes's theorem. In spite of a rich theoretical and applied literature, Bayesian methods for estimation and prediction using the beta-binomial model do not appear to have been developed. This article develops the appropriate Bayesian framework, derives several useful results, and illustrates the application of some results using simulation and a real data set.

The versatility of the beta-binomial model has led to its use in mental testing (Huynh 1979; Lord 1965; Wilcox 1981), toxicological experimentation (Williams 1975), epidemiology (Griffiths 1973), media exposure (Greene 1970), and buying behavior (Massy, Montgomery, and Morrison 1970). In many cases, motivated by the apparent inapplicability of the beta-binomial model, modifications and extensions have been proposed. Some examples are found in Kupper and Haseman (1978), Morrison and Brockway (1979), and Sabavala and Morrison (1981). The estimation procedures that are currently used include the method of moments, minimum chi-squared, and maximum likelihood (ML) (see Griffiths 1973; Kalwani 1980; Kleinman 1973; Morrison 1966; Wilcox 1979).

The development of Bayesian methods appears to be important for at least two reasons. First, although ML estimates have some desirable asymptotic properties and hence may be preferred when sample sizes are large, they may be inappropriate for small samples. Moreover, the regularity conditions cannot be shown to hold in general and, in some cases, ML estimates are not defined. Second, given the extensive application of the beta-binomial model in many fields, the Bayesian framework formally allows prior information to be incorporated in the analysis. In many applications it would be reasonable to assume that the analyst or the decision maker has such information.

The beta-binomial model is generated in the following way. Consider a population in which for each member some event occurs as the outcome of a Bernoulli trial with probability p. Then, k, the number of occurrences in r trials, given p, is binomial,

$$\Pr(k \mid r, p) = \binom{r}{k} p^k (1-p)^{r-k}, \qquad 0 \le p \le 1, \quad k = 0, 1, 2, \ldots, r. \quad (1)$$

Suppose that p varies across the population according to a beta distribution,

$$f(p \mid \alpha, \beta) = p^{\alpha-1}(1-p)^{\beta-1}/B(\alpha, \beta), \quad \alpha > 0, \quad \beta > 0, \quad 0 \le p \le 1, \quad (2)$$

where $B(\alpha, \beta)$ is the complete beta function. Since p is not observable, the probability distribution of k in r trials, given α and β, for a randomly chosen member is

$$\Pr(k \mid r, \alpha, \beta) = \int \Pr(k \mid r, p) f(p \mid \alpha, \beta)\, dp.$$

Using (1) and (2), the model is defined as

$$\Pr_{BB}(k \mid r, \alpha, \beta) = \binom{r}{k} B(\alpha + k, \beta + r - k) \div B(\alpha, \beta), \qquad k = 0, 1, 2, \ldots, r. \quad (3)$$

Note that we are concerned with inference on α and β, the beta-binomial model, rather than that on p, the binomial model, which is a somewhat different problem. It is well known that usually Bayesian analysis of the binomial model involves a beta conjugate prior. For Bayesian analysis of the multinomial model, which includes the binomial as a special case, see Fienberg and Holland (1973), Novick, Lewis, and Jackson (1973), and Leonard (1977).

As the beta-binomial model and its properties were discussed in detail by Johnson and Kotz (1969), we restrict our attention to the estimation and prediction problems. A key initial step in our analysis is to reparameterize the model in terms of $\mu = \alpha/(\alpha + \beta)$ and $\rho = 1/(1 + \alpha + \beta)$. A similar reparameterization, using $\rho/(1 - \rho) = 1/(\alpha + \beta)$ instead of ρ, was used by Griffiths (1973) and Williams (1975). Although α and β can take on any positive values, μ and ρ are restricted to values between 0 and 1. Moreover, μ and ρ have useful interpretations in almost any context, but α and β do not. μ is the mean of the beta distribution or the average probability of the event occurring. ρ is a measure of heterogeneity in the population with respect to p (see Sabavala and Morrison 1977). Another interesting interpretation of ρ is that it is the correlation between the binary outcome vectors across individual members on any two trials.

The main problems with which we will be concerned in this article are estimation of the parameters and prediction of the event frequencies. The typical input data, for a fixed and known number of trials, r, and a sample of size n, consist of the number of occurrences, k, for each of the n units. The data may be summarized as $\{n_k;\, k = 0, 1, 2, \ldots, r\}$, where n_k is the number of units with k occurrences. The estimation problem is to make point and interval estimates of μ and ρ, based on $\{n_k\}$. The prediction problem is to estimate k', the number of event occurrences in some future r' opportunities for a specific member of the population, conditional on having observed k out of r for that same member. In making these predictions, it is assumed that p for each member of the population remains constant. Essentially, for the n sampled units, we predict the events in some time domain that is not observed. It is also possible to pose another prediction problem, that is, for k' for some other units, not observed but believed to belong to the same population.

Section 2 describes the ML approach, and Section 3 describes the Bayesian approach. Results for the estimation and prediction problems for $r = 2$ are presented, and those for $r \ge 3$ are available from us on request. Since much of the rest of the article is based on $r = 2$, Section 4 briefly discusses the importance of this special case. Because of the complicated analytical results for the Bayesian method and the consequent computational difficulties, an approximation using the Pearson Type I distributions is proposed in Section 5. The ML, exact, and approximate Bayesian methods are compared using simulated data in Section 6. Section 7 illustrates an application to the problem of predicting television viewing. The final section summarizes the results, discusses the benefits and limitations of the Bayesian methods for the beta-binomial distribution, and indicates areas for future research.

2. THE MAXIMUM LIKELIHOOD APPROACH

Most applications of the beta-binomial model have estimated the parameters by using either the method of moments or the ML approach by direct search in the parameter space, although the first-order conditions for maximizing the likelihood are known (see Griffiths 1973; Kleinman 1973). The estimators are not available in closed form, but numerically solving these nonlinear likelihood equations must be preferred to direct search because of greater computing efficiency and precision and because the likelihood function may not be smooth and well behaved.

Using (3), the likelihood function is

$$L(\alpha, \beta \mid r, \{n_k\}) = \prod_{k=0}^{r} \{\Pr_{BB}(k \mid r, \alpha, \beta)\}^{n_k}.$$

Taking logs and differentiating with respect to α and β can be shown to yield

$$\sum_{k=1}^{r} n_k A(\alpha, k) - nA(\alpha + \beta, r) = 0, \quad (4a)$$

$$\sum_{k=1}^{r} n_{r-k} A(\beta, k) - nA(\alpha + \beta, r) = 0, \quad (4b)$$

where $A(u, i) = 1/(u) + 1/(u + 1) + \cdots + 1/(u + i - 1)$. The solution to (4) will provide the ML estimates, $\hat{\alpha}_{m1}$ and $\hat{\beta}_{m1}$. Because of the invariance property of ML estimates under one-to-one transformations, $\hat{\mu}_{m1}$ and $\hat{\rho}_{m1}$ are directly calculated. Interval estimates are made by resorting to the asymptotic theory (see Rao 1973). The asymptotic variances and covariance of $\hat{\mu}_{m1}$ and $\hat{\rho}_{m1}$ have not appeared previously and are available from us on request. Practically, it is rarely clear how large n should be for these asymptotic results to be reasonable. Moreover, when parameters are bounded, as μ and ρ are, asymptotic interval estimates could fall outside the valid range.

The usual solution to the first prediction problem using ML estimates is to find $\Pr(k' \mid r', k, r, \{n_k\})$. That is, given the data $\{n_k\}$, we would like to predict k' for the units observed to have k successes in the past. Although this quantity is not known, it can be shown that, given α and β,

$$\Pr(k' \mid r', k, r, \alpha, \beta) = \Pr_{BB}(k' \mid r', \alpha + k, \beta + r - k). \quad (5)$$

The predictive (i.e., conditional) distribution has the same form as (3), but with modified parameters. This appealing result follows from the fact that the beta distribution, $f(p)$, is conjugate with respect to the binomial distribution. A point estimate of k' could be the mean of the conditional distribution, $[r'(\alpha + k)/(\alpha + \beta + r)]$. Although α and β (or μ and ρ) are not known, their ML estimates are substituted in evaluating (5). This "plug-in" approach is in sharp contrast to the Bayesian approach.

The second prediction problem posed previously is also solved by "plugging in" the ML estimates, but in this case the past behavior of the individual units is not observed and (3) may be directly used:

$$\Pr(k' \mid r', \alpha, \beta) = \Pr_{BB}(k' \mid r', \alpha, \beta). \quad (6)$$

3. THE BAYESIAN APPROACH

The Bayesian (B) approach begins by specifying a prior distribution, $g(\mu, \rho)$. Using the likelihood function $L(\mu, \rho \mid r, \{n_k\})$, the posterior distribution of μ and ρ is obtained. When interest centers on one parameter, then the marginal posterior density may be found. The B estimator may be chosen as the mean, median, or mode of this marginal posterior. Interval estimates may be found using the symmetric or the highest posterior density (hpd) interval. This possibility of integrating out "nuisance parameters" does not have a parallel in the ML approach. The prediction problem is approached using (5) and (6) but explicitly integrating out μ and ρ. As indicated earlier, this is quite different from the ML approach, where predictions are made by substituting $\mu = \hat{\mu}_{m1}$ and $\rho = \hat{\rho}_{m1}$; that is, a degenerate posterior is being assumed—with probability 1, $\mu = \hat{\mu}_{m1}$, and $\rho = \hat{\rho}_{m1}$.

We specify the prior with beta marginals and, for mathematical tractability, we assume independence; that is,

$$g(\mu, \rho) = g_1(\mu)g_2(\rho), \quad (7a)$$

$$g_1(\mu) = \mu^{\gamma_1 - 1}(1 - \mu)^{\gamma_2 - 1}/B(\gamma_1, \gamma_2), \quad (7b)$$

$$g_2(\rho) = \rho^{\delta_1 - 1}(1 - \rho)^{\delta_2 - 1}/B(\delta_1, \delta_2). \quad (7c)$$

Although the prior is specified assuming independence, the posterior will reflect any dependence of μ and ρ or vice versa. The prior can be chosen to be noninformative, in the sense that by specifying $\gamma_1 = \gamma_2 = \delta_1 = \delta_2 = 1$, $g(\mu, \rho) = 1$. In addition, the beta marginals make it possible for subjective priors to be elicited. Although we may be able to induce dependence consistent with these marginals, it would require an additional parameter.

The rest of this article will be concerned mostly with the special and important case of $r = 2$. When $r \geq 3$, the derivation and results are quite complicated and, therefore, are available from us on request. Numerical computation is also likely to be burdensome.

For $r = 2$, the likelihood function can be expressed as a finite sum of terms, each of which has the same form as the kernel of the prior, $g(\mu, \rho)$:

$$\begin{aligned} L(\mu, \rho \mid r = 2, \{n_k\}) &= 2^{n_1} \sum_{t_0=0}^{n_0} \sum_{t_2=0}^{n_2} (-1)^{t_0} \binom{n_0}{t_0} \binom{n_2}{t_2} \\ &\quad \times \mu^{n_1 + 2n_2 + t_0 - t_2}(1 - \mu)^{n_0 + n_1 + t_2} \rho^{t_2} (1 - \rho)^{n_1 + t_0}. \end{aligned} \quad (8)$$

Using (7) and (8), the joint posterior is

$$\begin{aligned} \Pr(\mu, \rho \mid r = 2, \{n_k\}) &= \frac{C 2^{n_1}}{B(\gamma_1, \gamma_2) B(\delta_1, \delta_2)} \\ &\quad \times \sum_{t_0=0}^{n_0} \sum_{t_2=0}^{n_2} (-1)^{t_0} \binom{n_0}{t_0} \binom{n_2}{t_2} \\ &\quad \times \mu^{n_1 + 2n_2 + \gamma_1 + t_0 - t_2 - 1}(1 - \mu)^{n_0 + n_1 + \gamma_2 + t_2 - 1} \\ &\quad \times \rho^{\delta_1 + t_2 - 1}(1 - \rho)^{n_1 + \delta_2 + t_0 - 1}, \end{aligned} \quad (9)$$

where

$$\begin{aligned} C^{-1} &= \frac{2^{n_1}}{B(\gamma_1, \gamma_2) B(\delta_1, \delta_2)} \\ &\quad \times \sum_{t_0=0}^{n_0} \sum_{t_2=0}^{n_2} (-1)^{t_0} \binom{n_0}{t_0} \binom{n_2}{t_2} \\ &\quad \times B(n_1 + 2n_2 + \gamma_1 + t_0 - t_2, n_0 + n_1 + \gamma_2 + t_2) \\ &\quad \times B(\delta_1 + t_2, n_1 + \delta_2 + t_0). \end{aligned}$$

After integrating out ρ, the marginal posterior of μ is

$$\begin{aligned} \Pr_1(\mu \mid r = 2, \{n_k\}) &= \frac{C 2^{n_1}}{B(\gamma_1, \gamma_2) B(\delta_1, \delta_2)} \\ &\quad \times \sum_{t_0=0}^{n_0} \sum_{t_2=0}^{n_2} (-1)^{t_0} \binom{n_0}{t_0} \binom{n_2}{t_2} \\ &\quad \times B(\delta_1 + t_2, n_1 + \delta_2 + t_0) \mu^{n_1 + 2n_2 + \gamma_1 + t_0 - t_2 - 1} \\ &\quad \times (1 - \mu)^{n_0 + n_1 + \gamma_2 + t_2 - 1}. \end{aligned} \quad (10)$$

Similarly, the marginal posterior of ρ is

$$\Pr_2(\rho \mid r = 2, \{n_k\}) = \frac{C2^{n_1}}{B(\gamma_1, \gamma_2)B(\delta_1, \delta_2)} \times \sum_{t_0=0}^{n_0} \sum_{t_2=0}^{n_2} (-1)^{t_0} \binom{n_0}{t_0} \binom{n_2}{t_2} \times B(n_1 + 2n_2 + \gamma_1 + t_0 - t_2, n_0 + n_1 + \gamma_2 + t_2) \times \rho^{\delta_1+t_2-1}(1 - \rho)^{n_1+\delta_2+t_0-1}. \tag{11}$$

Because of the appearance of the prior kernel $[\mu^a (1 - \mu)^b \rho^c (1 - \rho)^d]$ in the joint and marginal posteriors, we might say that the beta prior has a conjugate-type relation to the beta-binomial model. Because the posterior is expressed in terms of weighted sums of these kernels, however, it is not strictly a conjugate relation. A similar result, in which the posterior is a mixture of the conjugate priors, was encountered in a Bayesian analysis of Warner's randomized response model by Winkler and Franklin (1979).

The sth moments of the marginal posteriors can also be obtained in closed form as

$$M_{s,\mu} = E(\mu^s \mid r = 2, \{n_k\}) = \frac{C2^{n_1}}{B(\gamma_1, \gamma_2)B(\delta_1, \delta_2)} \sum_{t_0=0}^{n_0} \sum_{t_2=0}^{n_2} (-1)^{t_0} \binom{n_0}{t_0} \binom{n_2}{t_2} \times B(n_1 + 2n_2 + \gamma_1 + t_0 - t_2 + s, n_0 + n_1 + \gamma_2 + t_2) \times B(\delta_1 + t_2, n_1 + \delta_2 + t_0) \tag{12}$$

and

$$M_{s,\rho} = E(\rho^s \mid r = 2, \{n_k\}) = \frac{C2^{n_1}}{B(\gamma_1, \gamma_2)B(\delta_1, \delta_2)} \sum_{t_0=0}^{n_0} \sum_{t_2=0}^{n_2} (-1)^{t_0} \binom{n_0}{t_0} \binom{n_2}{t_2} \times B(n_1 + 2n_2 + \gamma_1 + t_0 - t_2, n_0 + n_1 + \gamma_2 + t_2) \times B(\delta_1 + t_2 + s, n_1 + \delta_2 + t_0). \tag{13}$$

One possible estimator for μ would be $M_{1,\mu}$ and correspondingly, $M_{1,\rho}$ for ρ. The mode of the marginal posterior may also be used. Calculation of the mode, however, requires a numerical search for the values of μ and ρ that maximize the corresponding marginal posteriors.

The Bayesian prediction of k' for the units observed to have k successes in r opportunities in the past is based on

$$\Pr(k' \mid r', k, r, \{n_k\}) = \iint \Pr(k' \mid r', k, r, \alpha, \beta)\Pr(\mu, \rho \mid r, \{n_k\})\, d\mu\, d\rho.$$

Using (5) and the general version of (9), this integral may be evaluated. For the special case of $r = 2$ and $r' = 1$—that is, predicting the next outcome based on the previous two outcomes—we have

$$\Pr(k' = 1 \mid r' = 1, k, r = 2, \{n_k\}) = \frac{C2^{n_1}}{B(\gamma_1, \gamma_2)B(\delta_1, \delta_2)} \sum_{t_0=0}^{n_0} \sum_{t_2=0}^{n_2} (-1)^{t_0} \binom{n_0}{t_0} \binom{n_2}{t_2} \times \{B(n_1 + 2n_2 + \gamma_1 + t_0 - t_2 + 1, n_0 + n_1 + \gamma_2 + t_2) \times \sum_{j=0}^{\infty} (-1)^j B(\delta_1 + t_2 + j, n_1 + \delta_2 + t_0 + 1) + k \cdot B(n_1 + 2n_2 + \gamma_1 + t_0 - t_2, n_0 + n_1 + \gamma_2 + t_2) \times \sum_{j=0}^{\infty} (-1)^j B(\delta_1 + t_2 + j + 1, n_1 + \delta_2 + t_0)\}. \tag{14}$$

For the second prediction problem,

$$\Pr(k' \mid r', r, \{n_k\}) = \iint \Pr(k' \mid r', \alpha, \beta)\Pr(\mu, \rho \mid r, \{n_k\})\, d\mu\, d\rho.$$

Using (6) and the general version of (9), this integral may be evaluated. For the special case of $r = 2$ and $r' = 1$,

$$\Pr(k' = 1 \mid r' = 1, r = 2, \{n_k\}) = \frac{C2^{n_1}}{B(\gamma_1, \gamma_2)B(\delta_1, \delta_2)} \sum_{t_0=0}^{n_0} \sum_{t_2=0}^{n_2} (-1)^{t_0} \binom{n_0}{t_0} \binom{n_2}{t_2} \times B(n_1 + 2n_2 + \gamma_1 + t_0 - t_2 + 1, n_0 + n_1 + \gamma_2 + t_2) \times B(\delta_1 + t_2, n_1 + \delta_2 + t_0). \tag{15}$$

4. THE IMPORTANCE OF THE $r = 2$ CASE

The beta-binomial model has been so widely applied that it is difficult to be exhaustive in discovering applications in which $r = 2$ is potentially an important case. Rather, in this section, we will describe three such situations. The first is in epidemiology, the second is in marketing, and the third is in psychometrics, although the connection to marketing is obvious. It should also be noted that the $r = 2$ case is often chosen as a basis of discussion for clarity and simplicity (see, e.g., Kupper and Haseman 1978).

Example 1. Griffiths (1973) applied the beta-binomial model to describe the household distribution of the incidence of diseases. Consider the population of households with two adults. Then $r = 2$, since each household may have 0, 1, or 2 infected members. Griffiths discussed the appropriateness of the beta-binomial model with respect to the nature of the disease and assumptions about infectiousness.

Example 2. In marketing, the beta-binomial and other stochastic models have been applied to purchasing behavior, often focusing on brand-switching, which involves two choice occasions (see Massy et al. 1970). These data are often obtained from telephone surveys or personal interviews by asking respondents what brand was bought last time and the time before. Because of problems in accuracy of recall, it is inadvisable to ask for more than two previous purchases. For any given brand, the respondents will have made 0, 1, or 2 purchases in the previous $r = 2$ choice occasions. Kalwani and Morrison (1977) showed how this approach may be used to infer the competitive structure of the market.

Example 3. Morrison (1981) made a strong case for testing the discriminability of subjects in taste-testing situations, where preferences are of ultimate concern. The double triangle discrimination test is frequently used to see if subjects can discriminate between two stimuli, A and B. Subjects are asked to identify A and B respectively in each of two triads, (A, B, B) and (A, A, B). Therefore, subjects will be correct 0, 1, or 2 times. Morrison applied the beta-binomial model with $r = 2$, but modified to allow for guessing by nondiscriminating subjects, as described by Morrison and Brockway (1979). These ideas are fundamental for correctly interpreting the taste (preference) tests that are increasingly being used for product design and comparative advertising. Buchanan and Morrison (1985) developed an approach for designing such tests and for evaluating the results.

5. APPROXIMATING THE MARGINAL POSTERIOR DISTRIBUTION

Because of the computational difficulties in evaluating integrals of the posterior distribution for calculating interval estimates or the median, it would be worth approximating the posterior by a known distribution for which methods are more developed. We propose a Pearson Type I distribution based on the first four moments of the marginal posterior. This distribution has four parameters and is essentially a beta distribution defined over an interval other than (0, 1), the limits of which are two of the parameters. Such an approach has great value in Bayesian analysis, in which the posterior distribution is difficult to work with but the moments are available. Approximation using the Pearson system has been used elsewhere—for example, by Krishnaiah, Lee, and Chang (1976) for likelihood ratio criteria in multivariate analysis and Solomon and Stephens (1978) for nonparametric tests. Another approach to posterior computation was suggested by Naylor and Smith (1982).

Since the procedure is not widely known, the essential details are available from us on request (see also Elderton and Johnson 1969; Johnson and Kotz 1970; Johnson, Nixon, Amos, and Pearson 1963). If a Pearson Type I distribution is appropriate, let the original random variable (actually, μ or ρ) be transformed to x so that $-a_1 < x < a_2$ and $x = 0$ at the mode. Then, the approximation used is

$$[(x + a_1)/(a_1 + a_2)] \sim \beta(a, b),$$

where a, b, a_1, and a_2 are calculated from the first four moments.

The mode and mean are easily calculated. The median and other percentiles require calculation of the incomplete beta function, which does not present a problem. From these percentiles, the symmetric or the hpd interval estimate of μ or ρ may be obtained.

6. COMPARATIVE PERFORMANCE OF ESTIMATORS USING SIMULATED DATA

The analytical results for the posterior distributions of μ and ρ are sufficiently complex, even for $r = 2$, to warrant demonstration of the feasibility of the numerical computations. With this in mind, data were simulated from four true models that represent quite distinct beta-binomial models. The four true models are the following.

Model 1: $\mu = .4, \rho = .50$.
Model 2: $\mu = .4, \rho = .25$.
Model 3: $\mu = .5, \rho = .33$.
Model 4: $\mu = .8, \rho = .29$.

The underlying beta distributions are shown in Figure 1.

The other parameters of the simulation were the sample size ($n = 20$ and $n = 30$), the number of replications (100), and the priors (uniform and non-uniform). The procedure that was followed was to generate n values of k, the number of successes in $r = 2$ opportunities, for a given μ and ρ. The resulting frequency distribution, $\{n_0, n_1, n_2\}$, was used to compute Bayesian (B), approximate Bayesian (AB), and ML estimates of μ and ρ. This step was repeated 100 times, and summary measures of the discrepancy between the estimates and the true values were computed. For point estimators of μ and ρ, which were chosen to be the posterior means for the B method, the summary measures were mean absolute deviation (MAD), root mean squared error (RMSE), and the proportion closer to the true value relative to the ML estimators (CLOS). For interval estimates (at a given probability level), the average width of the interval (AW) was used, with a check on the proportion including the true value (INCL).

The first set of results is based on uniform priors—that is, with $\gamma_1 = \gamma_2 = \delta_1 = \delta_2 = 1$. Table 1 summarizes results for the point estimators. The B estimators systematically outperform the ML estimators on MAD and RMSE. Although the difference for μ is small and in

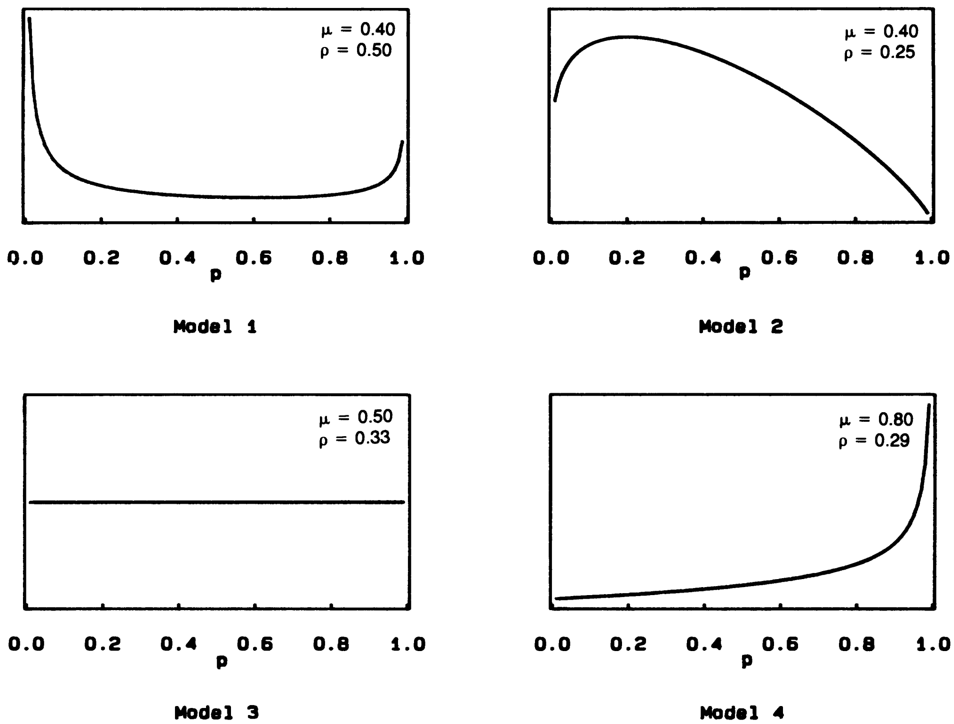

Figure 1. True Models for Simulation.

some cases negligible, the difference for ρ is considerable in all cases. For example, for Model 2, $n = 20$, for μ the MAD for B is .059 versus .061 for ML, and for ρ, the MAD for B is .081 versus .113 for ML. In all but two cases (for μ, Model 4), the B estimators are closer to the true value than the ML estimators in much more than 50% of the replications. On this measure as well, the impact on ρ is greater.

It should be noted that in several cases ML estimates cannot be obtained. Although we have not established a general test for the regularity conditions, it is clear that for $r = 2$ for several realizations the likelihood equations cannot be solved. This problem has been largely ignored in the theoretical and applied literature on the beta-binomial model. Notable exceptions were Kleinman (1973), who suggested that regularity conditions are met based on some numerical examples only, and Wilcox (1979), who recognized the possibility of negative-valued (and therefore meaningless) solutions to the ML equations. The magnitude of the problem is far from negligible. In our simulation, for $n = 20$, the numbers of replications (out of the 100 in each case) in which ML estimators could not be obtained are 4, 17, 9, and 17 for Models 1–4. The corresponding values for $n = 30$ are 2, 12, 4, and 12. An intuitive explanation for the undefined ML estimators is that the particular (n_0, n_1, n_2) distribution is basically inconsistent with respect to a beta-binomial model, because the observed

Table 1. Comparison of Point Estimators Using Simulated Data and Uniform Priors

	$n = 20$						$n = 30$					
	μ			ρ			μ			ρ		
Estimator	*ML*	*B*	*AB*	*ML*	*B*	*AB*	*ML*	*B*	*AB*	*ML*	*B*	*AB*
MAD												
Model 1	.079	.074	.074	.160	.136	.136	.059	.057	.057	.141	.126	.126
Model 2	.061	.059	.059	.113	.081	.081	.058	.058	.058	.133	.101	.101
Model 3	.065	.063	.063	.155	.114	.114	.057	.054	.054	.127	.103	.103
Model 4	.053	.058	.058	.167	.120	.120	.043	.043	.043	.148	.111	.111
RMSE												
Model 1	.095	.091	.091	.191	.160	.160	.073	.071	.071	.173	.154	.154
Model 2	.077	.073	.073	.138	.103	.103	.073	.073	.073	.156	.124	.124
Model 3	.085	.084	.084	.187	.141	.141	.071	.068	.068	.157	.128	.128
Model 4	.074	.076	.076	.206	.156	.156	.058	.059	.059	.178	.143	.143
CLOS												
Model 1	—	.71	.74	—	.73	.73	—	.53	.53	—	.68	.68
Model 2	—	.61	.61	—	.69	.69	—	.54	.54	—	.82	.82
Model 3	—	.86	.86	—	.99	.99	—	.89	.89	—	.90	.90
Model 4	—	.29	.29	—	.69	.69	—	.42	.42	—	.77	.77

Table 2. Comparison of Interval Estimators Using Simulated Data and Uniform Priors

	n = 20						n = 30					
	μ			ρ			μ			ρ		
Estimator	ML	B	AB	ML	B	AB	ML	B	AB	ML	B	AB
AW												
Model 1	.303	.287	.287	.610	.546	.544	.253	.242	.242	.505	.471	.470
Model 2	.283	.271	.271	.592	.496	.497	.233	.225	.226	.504	.443	.443
Model 3	.298	.282	.282	.597	.520	.520	.244	.235	.235	.518	.467	.466
Model 4	.246	.236	.236	.675	.574	.574	.196	.181	.193	.580	.510	.510
INCL												
Model 1	.85	.90	.90	.90	.93	.93	.90	.92	.92	.83	.90	.90
Model 2	.94	.96	.96	.99	.99	.99	.86	.89	.89	.93	.94	.94
Model 3	.90	.90	.90	.92	.93	.93	.91	.91	.91	.88	.92	.92
Model 4	.87	.92	.92	.90	.96	.96	.92	.90	.93	.93	.93	.93

NOTE: The probability for interval estimates is .9.

distribution is either too concentrated (where ρ tends to 0 and α, β tend to ∞) or too disperse or polarized (where ρ tends to 1 and α, β tend to 0). The B estimators presented no problems at all.

Consider now the interval estimation results of Table 2. The pattern is the same. The B intervals are consistently narrower than the ML intervals. For example, for Model 2, $n = 20$, for ρ the AW is .592 for ML and .496 for B. Moreover, in only one of the cases shown does the B interval fall below the expected 90% inclusion, whereas this occurs five times for the ML intervals. Another major problem with ML intervals is that, by the asymptotic normal property of ML estimators, the interval may fall out of the valid range (from 0 to 1 for μ and ρ, in our case). In spite of truncating the interval, the results of Table 2 show that the B intervals are still narrower. Although none of the ML estimates for μ were out of bounds, it was a severe problem for ρ. The magnitude of the problem may be gauged by the number of replications, assuming that ML estimators could be obtained, in which the ML interval estimate for ρ fell below 0 or above 1. For $n = 20$, for Models 1–4, the frequencies of such occurrences are 39, 63, 46, and 64! The corresponding numbers for $n = 30$ are 15, 45, 34, and 51. Therefore, for Model 4, $n = 20$, in 17 out of 100 replications ML estimates could not be obtained and of the remaining 83, 64 interval estimates of ρ fell partly outside the (0, 1) interval in either direction. Naturally, besides truncating the interval, there is no way to adjust the estimate.

Table 3. Comparison of Bayesian Symmetric and Asymmetric Interval Estimators of ρ using Simulated Data and Uniform Priors

	AW		INCL		Proportion of asym. better
	Sym.	Asym.	Sym.	Asym.	
$n = 20$					
Model 1	.546	.539	.93	.87	28
Model 2	.496	.472	.99	.97	73
Model 3	.520	.505	.93	.89	52
Model 4	.574	.551	.96	.94	59
$n = 30$					
Model 1	.471	.468	.90	.87	11
Model 2	.443	.427	.94	.92	50
Model 3	.467	.459	.92	.86	33
Model 4	.510	.491	.93	.92	55

NOTE: The probability for interval estimates is .9.

The ML interval estimates are symmetric in the tail probabilities (except when they are truncated). The B intervals may be chosen to be asymmetric, however. For our simulated data with uniform priors, the leftmost (with 0 as one endpoint) and rightmost (with 1 as one endpoint) interval estimates were also constructed. For μ, these asymmetric intervals were never narrower than the symmetric. For ρ, however, the asymmetric intervals could do better. The results are presented in Table 3. For example, for Model 2, $n = 20$, the symmetric AW was .496, but the AW of the best of the three intervals was .472. Of the 100 replications, the asymmetric interval was chosen 73 times. Therefore, it appears that the marginal posterior of ρ is quite skewed. For computational economy, the interval chosen was the shortest of only three alternatives—one with a lower limit of 0, one with an upper limit of 1, and the symmetric interval. With further computational effort, it is possible to obtain a more precise hpd interval.

The approximate distributions could always be obtained and were remarkably close to the B estimates. This fact is reflected in Tables 1 and 2 in the statistics for the AB and B methods.

To examine the impact of non-uniform priors, Model 1 was simulated ($n = 20$) and analyzed with three sets of priors shown in Figure 2. In each case—that is, for μ and ρ—prior A is expected to be "best," prior C is the uniform, and prior B is a non-uniform prior, which can be construed to be "misinformed" but not necessarily worse than prior C. Prior A is defined by $\gamma_1 = 1.6$, $\gamma_2 = 2.4$, $\delta_1 = 2.0$, and $\delta_2 = 2.0$. Prior B is defined by $\gamma_1 = 2.0$, $\gamma_2 = 2.0$, $\delta_1 = 3.6$, and $\delta_2 = 1.2$.

The results are summarized in Table 4 and indicate that non-uniform (informed) priors can do much better. Again, the effect is greater for ρ. For example, for ρ, the MAD of the ML estimators is .160 as compared with .136 for the uniform prior B estimates and .104 for the B estimates with prior A! The corresponding quantities for the AW of the .9 probability interval are

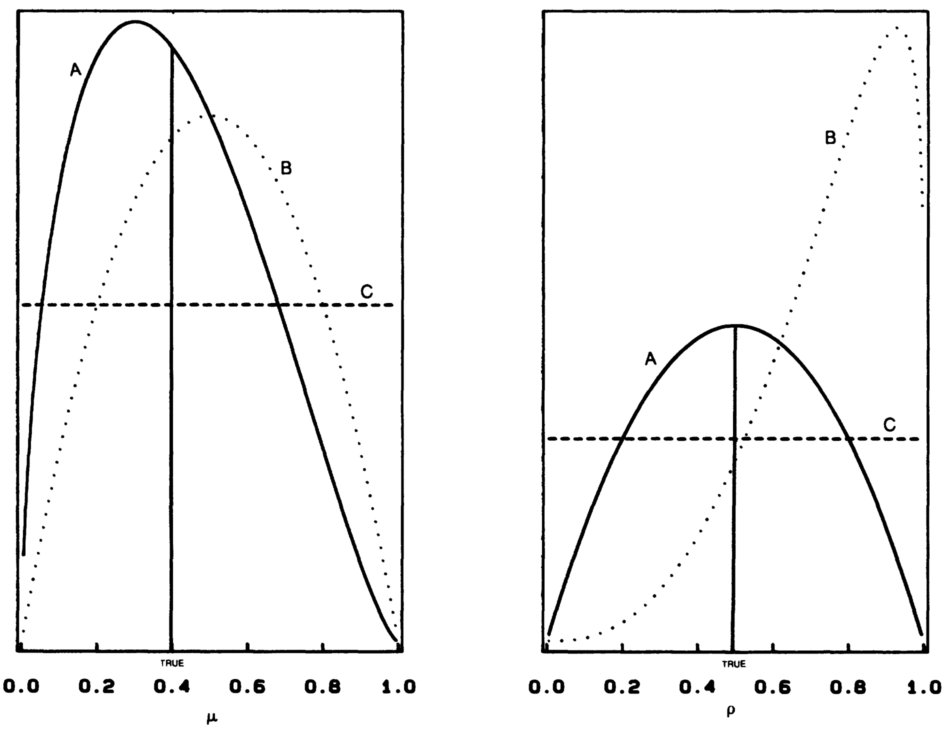

Figure 2. Comparison of Uniform and Non-Uniform Priors, True Model: $\mu = .4$, $\rho = .5$.

.610, .546, and .494. Prior A is better than prior C on all criteria; prior B is more erratic, however, since it appears to perform better than prior C on MAD, RMSE, and AW, but it is poorer on CLOS and INCL. The results for non-uniform priors are encouraging in that they demonstrate the potential value of including good prior information.

7. AN ILLUSTRATION USING TELEVISION-VIEWING DATA

The beta-binomial model is used for describing and predicting audience accumulation in several commercially available media models. Each member of some potential audience is assumed to have a fixed probability of being exposed to a media vehicle (e.g., viewing a particular television show or reading a particular magazine), and this probability varies across audience members according to a beta (μ, ρ) distribution. Empirical evidence indicates that for most established media the beta distributions are reverse-J-shaped or U-shaped.

Table 4. Comparison of Estimators Using Simulated Data and Priors A, B, and C for Model 1 and n = 20

Estimator	ML	Bayesian Prior A	Bayesian Prior B	Bayesian Prior C
Estimation of μ				
MAD	.079	.070	.066	.074
RMSE	.095	.085	.084	.091
CLOS	—	.95	.70	.71
AW	.303	.279	.286	.287
INCL	.85	.91	.90	.90
Estimation of ρ				
MAD	.160	.104	.115	.136
RMSE	.191	.124	.143	.160
CLOS	—	.84	.48	.73
AW	.610	.494	.435	.546
INCL	.90	.97	.89	.93

In media exposure applications, the parameters μ and ρ have specific interpretations. μ, the average probability of exposure, is an important measure of the expected audience size. For television, μ may be interpreted as the expected value of the rating, which is the proportion of television households tuned to a station. Notice that the alternative to viewing the station includes both viewing another station and not viewing. ρ is a measure of heterogeneity of the potential audience and was proposed as a measure of strength of preference for television shows by Sabavala and Morrison (1977). Their empirical results, based on 149 programs at two points in time, showed that ρ provided valuable information on program performance that would be of interest to advertisers and to program-scheduling decision makers. In that analysis, the correlation across programs between μ and ρ was low, suggesting that ρ adds information not contained in μ.

In addition to estimating μ and ρ, media producers and advertisers are also interested in estimating the frequency distribution of the number of exposures. For example, an advertiser placing an advertisement in r telecasts of a program would like to know how many viewers had k ($k = 0, 1, 2, \ldots, r$) opportunities to see the advertisement. Based on the advertiser's assumptions about the value of repeated exposures, this distribution provides the information needed to compare alternative programs. In some cases, it is also of

Table 5. Analysis of Television-Viewing Data: Prediction of Day 5 Using Informed Priors

	No. times viewed on days 1–4				
Program/slot	*0*	*1*	*2*	*3*	*4*
1. Number of homes	30	12	3	3	12
Actual number viewing on day 5	1	1	1	3	11
Bayesian day 5 prediction	2.2	2.8	1.0	1.3	9.8
ML day 5 prediction	1.3	3.1	1.4	2.1	10.9
2. Number of homes	47	6	1	1	5
Actual number viewing on day 5	1	0	0	1	4
Bayesian day 5 prediction	1.1	.5	.4	.4	4.0
ML day 5 prediction	.7	1.4	.5	.7	4.5
3. Number of homes	44	8	1	3	4
Actual number viewing on day 5	0	0	0	2	3
Bayesian day 5 prediction	1.9	1.8	.4	1.2	3.0
ML day 5 prediction	1.1	1.9	.4	2.0	3.5
4. Number of homes	32	7	5	9	7
Actual number viewing on day 5	2	2	1	6	6
Bayesian day 5 prediction	.4	1.6	.9	2.4	3.1
ML day 5 prediction	1.6	1.8	2.3	6.0	6.2

interest to predict the number of exposures, k', out of r' future telecasts of a program, given that we have observed the number of exposures k in r past telecasts. For example, in the third of a televised series of three political debates, it may be strategically worthwhile for the participants to understand the likely composition of the audience based on observation of the past two telecasts for a sample of viewers. This application corresponds to the special case of $r = 2$ and $r' = 1$ mentioned previously. [It is natural to ask whether the assumption that each viewer has a fixed exposure probability is tenable, since the outcomes on the first two occasions may alter that probability. This is a separate issue that has been well studied in many fields (e.g., see Sabavala and Morrison 1981) and is beyond the scope of this discussion.]

The data that we analyze were obtained from a sample of 60 homes located in a particular television market and collected for a study by a television network on the strength of loyalty for news programs. An assigned member of each household maintained a five-day diary of prime-time viewing. We will focus on four separate program/time slots, an example of which might be *The CBS Evening News*/6:00–6:30 p.m. For each we would like to predict day 5 viewing for the 60 homes based on observations of days 1–4.

Two different approaches will be used. In the first, we would like to illustrate the use of informed priors in prediction by doing the Bayesian estimation in two stages: (a) analyzing days 1–2 with a uniform prior and approximating the Pearson approximation to the marginal posteriors by beta distributions of the form (2), and (b) using these, assuming independence, as priors for analyzing days 3–4. We will predict day 5 based on the updated posterior using (14). These predictions will be compared with ML estimation with $r = 4$ and subsequent use of (5) by plugging in $\hat{\mu}_{m1}$ and $\hat{\rho}_{m1}$. Although the B and ML results are not substantially different, the low sample frequencies caution against making any comparative generalizations from this particular data set. Our objective is merely to illustrate the applicability of the Bayesian approach for making predictions of viewing behavior.

The results are shown in Table 5. The interpretation is as follows: 12 of the 60 homes viewed program/slot 1 for 4 out of 4 telecasts during days 1–4, and 11 of these 12 also viewed on day 5. The B prediction is 9.8 viewing on day 5, but the ML approach predicts 10.9. For the four program/slot examples shown in Table 5, the ML predictions appear to do about the same as the B predictions, which do not exhibit any unusual behavior.

The second approach consists of making predictions based on the previous two days' behavior using uniform priors. For our data we can predict, separately, days 3, 4, and 5 based on days 1–2, 2–3, and 3–4, respectively. The results are shown in Table 6, in which the three

Table 6. Analysis of Television-Viewing Data: Prediction Based on the Past Two Days Using Uniform Priors

	No. times viewed in past two days		
Program/slot	*0*	*1*	*2*
1. Number of homes	110	29	41
Actual number viewing on next day	6	9	36
Bayesian prediction	6.0	12.2	32.9
ML prediction	7.9	13.1	34.4
2. Number of homes	148	16	16
Actual number viewing on next day	2	3	14
Bayesian prediction	5.4	6.4	12.3
ML prediction	4.7	6.5	12.7
3. Number of homes	144	21	15
Actual number viewing on next day	4	6	11
Bayesian prediction	7.5	7.7	10.5
ML prediction	6.6	7.8	11.1
4. Number of homes	107	37	36
Actual number viewing on next day	10	17	26
Bayesian prediction	15.2	16.5	29.3
ML prediction	10.8	15.5	28.3

days' predictions are added together; that is, each home accounts for three observations. We can extend our findings from the simulation by observing that in prediction as well, the Bayesian method using uniform prior does as well as ML.

In both approaches, it is surprising that the beta-binomial predictors were as good as they are, since there is some evidence of differences in viewing levels across the five days. To account for these day-of-the-week effects, it would probably be more appropriate to predict Wednesday (say) viewing based on the two previous Wednesdays rather than on the two previous days.

8. CONCLUSIONS

This article has studied the beta-binomial model from a Bayesian viewpoint. The simulation results clearly showed that B point and interval estimators were consistently better than ML estimators. The improvement was greatest for ρ, especially for smaller samples sizes. In fact, it can be shown that B estimators asymptotically approach ML estimators with only one slight difference in the variance–covariance matrix (see Cox and Hinkley 1974, p. 400). The posterior distribution of ρ was sufficiently skewed in our simulated models to allow use of asymmetric interval estimates. The analysis of the television viewing data provided a clear illustration of the prediction results.

Several problems with the ML estimation were encountered and discussed. On the other hand, the B estimators presented no problems. The proposed approximation of the posterior is promising. Failing such approximation, the numerical calculations are not only tedious but may present major problems in precision.

In addition to posterior approximation, there are several opportunities for future research. These include sample size determination and using the predictive distribution for model discrimination. It is also possible to explore the nature of the dependence between μ and ρ that is reflected in the posterior through the conditional distributions $\Pr(\mu \mid \rho, r, \{n_k\})$ and $\Pr(\rho \mid \mu, r, \{n_k\})$.

The results suggest the importance of exploring Bayesian analysis for other stochastic mixture models. Some preliminary work on the negative binomial model is encouraging. The negative binomial model, which can result from a gamma mixture of Poisson individual models, is

$$\Pr_{NB}(x; v, a) = \binom{x + v - 1}{v - 1} \left(\frac{a}{a + 1}\right)^{v} \left(\frac{1}{a + 1}\right)^{x},$$
$$x = 0, 1, 2, \ldots, v, \quad a > 0.$$

Closed-form posteriors and the conjugate-type property can be obtained assuming prior independence, a gamma marginal prior for v, and a beta marginal prior for $[a/(1 + a)]$.

It is worth pointing out that when Bayesian methods are developed for a mixture model, effectively, the groundwork has been laid for a two-stage hierarchical Bayesian solution for the underlying individual unit model. Therefore, our results are likely to be of some interest in the development of hierarchical Bayesian methods for the Bernoulli or binomial model.

In conclusion, in applications in which the beta-binomial model is established, the Bayesian methods provide a reasonable approach to using the accumulated experience of the researcher or decision maker. For example, in applications to buying behavior, a manager's prior information on the market share (μ) and polarization index (ρ) can be formally incorporated in the analysis. In this article, Bayesian methods have been developed for the beta-binomial model, closed-form results have been obtained, and a conjugate-type property has been established. The computations are not trivial, and estimation for $r \geq 3$ involves infinite sums. In many applications, however, $r = 2$ and the superior small-sample performance with respect to ML estimates have been demonstrated using simulated data. The Bayesian prediction results have been illustrated using some real data on television viewing.

ACKNOWLEDGMENTS

We are grateful to Donald G. Morrison and Seymour Geisser for their valuable comments on a draft of this article. We also acknowledge the careful and thorough computer programming assistance of Mark Schroeder.

[Received May 1985. Revised September 1986.]

REFERENCES

Buchanan, B. S., and Morrison, D. G. (1985), "Measuring Simple Preferences: An Approach to Blind, Forced-Choice Product Testing," *Marketing Science*, 4, 93–109.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, New York: Halsted Press.

Elderton, W. P., and Johnson, N. L. (1969), *Systems of Frequency Curves*, New York: Cambridge University Press.

Fienberg, S., and Holland, P. (1973), "Simultaneous Estimation of Multinomial Cell Probabilities," *Journal of the American Statistical Association*, 68, 683–691.

Greene, J. D. (1970), "Personal Media Probabilities," *Journal of Advertising Research*, 10, 12–18.

Griffiths, D. A. (1973), "Maximum Likelihood Estimation for the Beta Binomial Distribution and an Application to the Household Distribution of the Total Number of Cases of a Disease," *Biometrics*, 29, 637–648.

Huynh, H. (1979), "Statistical Inference for Two Reliability Indices in Mastery Testing Based on the Beta Binomial Model," *Journal of Educational Statistics*, 4, 231–246.

Johnson, N. L., and Kotz, S. (1969), *Distributions in Statistics: Discrete Distributions*, New York: John Wiley.

——— (1970), *Distributions in Statistics: Continuous Distributions—I*, New York: John Wiley.

Johnson, N. L., Nixon, E., Amos, D. E., and Pearson, E. S. (1963), "Table of Percentage Points of Pearson Curves, for Given $\sqrt{\beta_1}$ and β_2 Expressed in Standard Measure," *Biometrika*, 50, 459–498.

Kalwani, M. U. (1980), "Maximum Likelihood Estimation of Zero-Order Models Given Variable Numbers of Purchases per Household," *Journal of Marketing Research*, 17, 547–551.

Kalwani, M. U., and Morrison, D. G. (1977), "A Parsimonious Description of the Hendry System," *Management Science*, 23, 467–477.

Kleinman, J. C. (1973), "Proportions With Extraneous Variance: Single and Independent Samples," *Journal of the American Statistical Association,* 68, 46–54.

Krishnaiah, P. R., Lee, J. C., and Chang, T. C. (1976), "The Distribution of the Likelihood Ratio Statistics for Tests of Certain Covariance Structures of Complex Multivariate Normal Populations," *Biometrika,* 63, 543–549.

Kupper, L. L., and Haseman, J. K. (1978), "The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments," *Biometrics,* 34, 69–76.

Leonard, T. (1977), "A Bayesian Approach to Some Multinomial Estimation and Pretesting Problems," *Journal of the American Statistical Association,* 72, 869–874.

Lord, F. M. (1965), "A Strong True-Score Theory, With Applications," *Psychometrika,* 30, 234–270.

Massy, W. F., Montgomery, D. B., and Morrison, D. G. (1970), *Stochastic Models of Buying Behavior,* Cambridge, MA: MIT Press.

Morrison, D. G. (1966), "Testing Brand-Switching Models," *Journal of Marketing Research,* 3, 401–409.

——— (1981), "Triangle Taste Tests: Are the Subjects Who Respond Correctly Lucky or Good?" *Journal of Marketing,* 45, 111–119.

Morrison, D. G., and Brockway, G. (1979), "A Modified Beta Binomial Model With Applications to Multiple Choice and Taste Tests," *Psychometrika,* 44, 427–441.

Naylor, J. C., and Smith, A. F. M. (1982), "Applications of a Method for the Efficient Computation of Posterior Distribution," *Applied Statistics,* 31, 214–225.

Novick, M. R., Lewis, C., and Jackson, P. H. (1973), "Estimation of Proportions in *M* Groups," *Psychometrika,* 38, 19–46.

Pearson, E. S. (1925), "Bayes' Theorem in the Light of Experimental Sampling," *Biometrika,* 17, 388–442.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), New York: John Wiley.

Sabavala, D. J., and Morrison, D. G. (1977), "A Model of TV Show Loyalty," *Journal of Advertising Research,* 17, 35–43.

——— (1981), "A Non-stationary Model of Binary Choice Applied to Media Exposure," *Management Science,* 27, 637–657.

Skellam, J. G. (1948), "A Probability Distribution Derived From the Binomial Distribution by Regarding the Probability of Success as Variable Between the Sets of Trials," *Journal of the Royal Statistical Society,* Ser. B, 10, 257–261.

Solomon, H., and Stephens, M. A. (1978), "Approximations to Density Functions Using Pearson Curves," *Journal of the American Statistical Association,* 73, 153–160.

Wilcox, R. R. (1979), "Estimating the Parameters of the Beta Binomial Distribution," *Educational and Psychological Measurement,* 39, 527–535.

——— (1981), "A Review of the Beta Binomial Model and Its Extensions," *Journal of Educational Statistics,* 6, 3–32.

Williams, D. A. (1975), "The Analysis of Binary Responses From Toxicological Experiments Involving Reproduction and Teratogenicity," *Biometrics,* 31, 949–952.

Winkler, R. L., and Franklin, L. A. (1979), "Warner's Randomized Response Model: A Bayesian Approach," *Journal of the American Statistical Association,* 74, 207–214.