Bayesian Conjugate Analysis of the Galton-Watson Process

M. Mendoza
Departamento de Estadística y Actuaría
ITAM, Río Hondo 1
01000 México D.F., Mexico

E. Gutiérrez-Peña
Departamento de Probabilidad y Estadística
IIMAS-UNAM, Apartado Postal 20-726
01000 México D.F., Mexico

Abstract

In this article we consider the Bayesian statistical analysis of a simple Galton-Watson process. Problems of interest include estimation of the offspring distribution, classification of the process, and prediction. We propose two simple analytic approximations to the posterior marginal distribution of the reproduction mean. This posterior distribution suffices to classify the process. In order to assess the accuracy of these approximations, a comparison is provided with a computationally more expensive approximation obtained via standard Monte Carlo techniques. Similarly, a fully analytic approximation to the predictive distribution of the future size of the population is discussed. Sampling-based and hybrid approximations to this distribution are also considered. Finally, we present some illustrative examples.

KEY WORDS: Branching Process; Dirichlet Distribution; Forecasting; Offspring Distribution; Reproduction Mean; Population Growth

1 INTRODUCTION

The Galton-Watson process is a simple probabilistic model often used to describe population growth. The basic model assumes that each individual in the population gives rise to a random number, X, of descendants. Specifically, individuals reproduce independently according to the *offspring distribution*

$$\Pr(X = j) = \pi_j, \qquad j = 0, 1, \ldots, k. \tag{1.1}$$

Thus, given Z_0 original ancestors the evolution of the population can be described by means of $\{Z_i : i = 0, 1, \ldots\}$, where Z_i denotes the total number of individuals in the i-th generation. More detailed information is obtained if we can observe $\{Z_{ij} : i = 0, 1, \ldots; j = 0, 1, \ldots, k\}$, where Z_{ij} is the number of individuals in the i-th generation giving rise to exactly j descendants. In any event, the basic model is a discrete distribution determined by the vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)'$, where $0 < \pi_j$ $(j = 1, \ldots, k)$ and $\sum_{j=1}^{k} \pi_j < 1$, with $\pi_0 = 1 - \sum_{j=1}^{k} \pi_j$. If no further constraints are imposed on $\boldsymbol{\pi}$, this structure is referred to as *nonparametric*, as opposed to the *parametric* case where the offspring distribution is assumed to belong to a specific family (e.g. the binomial).

The main probabilistic properties of the simple Galton-Watson process have been considered by a number of authors (see, for example, Harris, 1963; Jagers, 1975; Karlin and Taylor, 1975; Guttorp, 1991), and will only be briefly reviewed here.

The sequence $\{Z_i : i = 0, 1, \ldots\}$ is a Markov process with transition probabilities

$$\Pr(Z_i = s | Z_{i-1} = t) = p(s, t|\boldsymbol{\pi}),$$

where $p(s, t|\boldsymbol{\pi})$ is the t-fold convolution of the offspring distribution (1.1). Clearly, if $Z_i = 0$ for some i then $Z_r = 0$ for all $r \geq i$, and so the population becomes extinct. Thus interest is often focused on the *extinction probability*

$$q = \Pr(Z_i = 0 \text{ for some } i = 1, 2, \ldots |\boldsymbol{\pi}).$$

Specifically, it is of interest to determine whether $q = 1$, in which case extinction is certain.

It is well-known that q is the smallest root of the equation $f(s) = s$, where $f(\cdot)$ is the probability generating function of the offspring distribution (1.1). Note that q is one of the roots of a polynomial of order k, so in general it cannot be determined analytically. Hence the significance of the following result (Harris, 1963).

Result. Let $m = m(\boldsymbol{\pi}) = \sum_{j=0}^{k} j\,\pi_j$ be the *reproduction mean* of the process. Then, provided $\pi_1 \neq 1$, $q < 1$ if and only if $m > 1$.

In other words, we can establish whether or not extinction is certain simply by looking at the reproduction mean m. This parameter is easier to compute and gives rise to the following classification:

$$\text{the process is} \begin{cases} \text{subcritical} & \text{if } m < 1 \\ \text{critical} & \text{if } m = 1 \\ \text{supercritical} & \text{if } m > 1 \end{cases}$$

Thus, as far as the classification of the process is concerned, we can regard m as the parameter of interest. Note, however, that a full characterization of the process is only possible through the vector parameter $\boldsymbol{\pi}$. Nevertheless, it is possible to analyze the process by considering only a few functionals of the offspring distribution, such as m and the reproduction variance $\sigma^2 = \sigma^2(\boldsymbol{\pi}) = \sum_{j=0}^{k} \{j - m(\boldsymbol{\pi})\}^2 \pi_j$.

Despite the simple structure of the Galton-Watson process, its statistical analysis raises some challenging problems. Inference for this model, as well as for more general branching processes, has been considered by Dion (1972, 1974), Dion and Keiding (1977), Basawa and Prakasa-Rao (1980), Maki and McDunnough (1989), Guttorp (1991), and Prakasa-Rao (1992), among others. Most of these authors are mainly concerned with maximum-likelihood estimation and related classical methods of inference.

This article presents a Bayesian analysis of the simple Galton-Watson process. In the next section we briefly review the maximum-likelihood-based inference for this model. Section 3 is devoted to the corresponding Bayesian conjugate analysis. An elementary treatment of this structure, as described in Guttorp (1991, Section 4.4), is reviewed in Section 3.1 and then a more general approach is proposed. Specifically, inferences about m are discussed in Section 3.3. The problem of predicting the size of a future generation is also considered in some detail in Section 3.4. In addition, Section 3.2 deals with the elicitation of a simple vague prior for this model. Some illustrative examples are presented in Section 4. Finally, Section 5 contains some concluding remarks.

2 CLASSICAL INFERENCE FOR THE GALTON-WATSON PROCESS

If the evolution of a given population follows a simple Galton-Watson process but the offspring distribution is not known, what can be said about the process given data from the first n generations? In this section we shall review some basic results concerning maximum-likelihood estimation and inference for the Galton-Watson process (see, for example, Jagers, 1975; and Guttorp, 1991).

Suppose we get detailed information up to the n-th generation, i.e. we observe $\mathcal{Z}_n = \{Z_{ij} : i = 0, \ldots, n; j = 0, \ldots, k\}$. For notational convenience we shall sometimes write $\mathcal{Z}_n = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n\}$, with $\boldsymbol{Z}_i = (Z_{i0}, \ldots, Z_{ik})'$. It is worth noting that $Z_i = \sum_{j=0}^k Z_{ij}$ and $Z_{i+1} = \sum_{j=0}^k j\, Z_{ij}$. It is then easy to see that

$$p(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n | \boldsymbol{\pi}) = \prod_{i=0}^n p(\boldsymbol{Z}_i | Z_i, \boldsymbol{\pi}),$$

where $p(\boldsymbol{Z}_i | Z_i, \boldsymbol{\pi})$ is a multinomial distribution with index Z_i and parameter $\boldsymbol{\pi}$. Hence the likelihood function based on observations from the first n generations is given by

$$L(\boldsymbol{\pi} | \mathcal{Z}_n) \propto \prod_{i=0}^n \prod_{j=0}^k \pi_j^{Z_{ij}} \propto \prod_{j=0}^k \pi_j^{Y_{nj}} \tag{2.1}$$

for all vectors $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)'$ such that $0 < \pi_j \;\; (j = 1, \ldots, k)$ and $\sum_{j=1}^k \pi_j < 1$, with $\pi_0 = 1 - \sum_{j=1}^k \pi_j$.

The maximum-likelihood estimate for the offspring distribution is given by

$$\hat{\pi}_j = \frac{Y_{nj}}{Y_n}, \qquad j = 0, \ldots, k, \tag{2.2}$$

where $Y_{nj} = \sum_{i=0}^n Z_{ij}$ is the number of individuals giving rise to exactly j descendants, whereas $Y_n = \sum_{j=0}^k Y_{nj} = \sum_{i=0}^n Z_i$ denotes the total number of individuals up to the n-th generation. It follows from (2.2) that the maximum-likelihood estimate of the reproduction mean is simply

$$\hat{m} = \sum_{j=0}^k j\, \hat{\pi}_j = \frac{Y_{n+1} - Z_0}{Y_n}.$$

We note that $\boldsymbol{Z}_n$ determines the value of Z_{n+1} (the size of the population at the $(n+1)$-th generation), which in turn determines Y_{n+1}. The estimates $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_k)$ and $\hat{m}$ are not unbiased. Also, $\hat{m}$ is consistent only conditional on the nonextinction of the process (Guttorp, 1991, Chapter 2). Note that $\hat{m}$ only depends on the data through the statistics (Y_n, Y_{n+1}). Unfortunately, the same is not true for the sampling distribution of $\hat{m}$. It is possible, however, to establish the asymptotic normality of $\hat{m}$, again, only provided the process does not become extinct. Therefore, it is not possible to classify the process using asymptotic arguments. In addition, asymptotic probabilistic statements about $\hat{m}$ require the value of σ^2 to be known, or at least estimated. On the other hand, the nonasymptotic distribution of $\hat{m}$ does not have a standard form. It should then be clear that, from a frequentist perspective, it is not easy to test the hypothesis of extinction, $H_0 : q = 1$ (equivalently, $H_0 : m \leq 1$).

In the next section we shall present a Bayesian analysis on the Galton-Watson process that does not rely on asymptotic arguments and for which the condition of nonextinction is not required.

3 BAYESIAN CONJUGATE ANALYSIS

3.1 *Basic Structure*

A Bayesian analysis of model (2.1) requires the specification of a prior distribution on $\boldsymbol{\pi}$. A convenient class of prior distributions is the natural conjugate family (see, for example, DeGroot, 1970, Chapter 9), given in this case by densities of the form

$$p(\boldsymbol{\pi}) \propto \prod_{j=0}^{k} \pi_j^{\alpha_j - 1} \tag{3.1}$$

with $\alpha_j > 0$, i.e. the Dirichlet family. It then follows that the posterior distribution is also Dirichlet,

$$p(\boldsymbol{\pi} | \mathcal{Z}_n) \propto \prod_{j=0}^{k} \pi_j^{\beta_j - 1}, \tag{3.2}$$

where $\beta_j = \alpha_j + Y_{nj}$.

Careful choice of the parameter $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_k)'$ allows us to represent various forms of prior beliefs. Moreover, a more flexible class of priors is available if we consider finite mixtures of Dirichlet distributions (see Diaconis and Ylvisaker, 1985, for a thorough discussion on the use of mixtures of conjugate distributions in Bayesian inference). In this article we shall only be concerned with the class of Dirichlet priors. It must be pointed out, however, that some of the results presented here hold more generally for mixtures thereof.

The properties of the Dirichlet distribution (3.2) are well-known. Let $\beta = \sum_{j=0}^{k} \beta_j$. Then, for all $j = 0, 1, \ldots, k$,

$$\begin{aligned} E(\pi_j | \mathcal{Z}_n) = \frac{\beta_j}{\beta}, \qquad & \mathrm{Var}(\pi_j | \mathcal{Z}_n) = \frac{\beta_j\,(\beta - \beta_j)}{\beta^2\,(\beta + 1)}, \\ \mathrm{Cov}(\pi_j, \pi_l | \mathcal{Z}_n) \;=\; & -\frac{\beta_j\,\beta_l}{\beta^2\,(\beta+1)}, \qquad j \neq l. \end{aligned} \tag{3.3}$$

In addition, the posterior mode is given by the vector $\boldsymbol{\pi}^M = (\pi_1^M, \ldots, \pi_k^M)'$ with entries

$$\pi_j^M = \frac{\beta_j - 1}{\beta - (k+1)}, \qquad j = 0, 1, \ldots, k.$$

It is worth noting that, given a proper conjugate prior, the posterior distribution of $\boldsymbol{\pi}$ is also proper for any sample size. Moreover, such a posterior is well-defined even if the process is subcritical. Thus, Bayesian inference for $\boldsymbol{\pi}$ is rather straightforward. Unfortunately, this is not the case for m, which in the present context is often regarded as the parameter of interest since its value determines the nature of the process. Bayesian analyses of the Galton-Watson process have been previously considered by Dion (1972), Heyde (1979), and Johnson, Susarla and van Ryzin (1979). Guttorp (1991, Chapter 4) provides an excellent review of both parametric and nonparametric Bayesian methods. In particular, he discusses the difficulties involved in the calculation of the posterior distribution for the reproduction mean m in the nonparametric case.

It should be pointed out that, conditional on the nonextinction of the process, the posterior for $\boldsymbol{\pi}$ is asymptotically normal as the number of generations increases. To see this, note that the likelihood (2.1) is proportional to the one that would have been obtained had $(Y_{n0}, \ldots, Y_{nk})$ been an observation from a multinomial distribution with index Y_n and parameter $\boldsymbol{\pi}$. Under these circumstances, $\boldsymbol{\pi}$ would be just a one-to-one transformation of the canonical parameter of a multinomial family. The result then follows from the asymptotic normality of the posterior distribution for the canonical parameter of an exponential family (Bernardo and Smith, 1994, Section 5.3), together with the preservation of the asymptotic normality under transformations (Mendoza, 1994). This argument requires that Y_n (and not just n) tend to infinity. Of course this is only possible essentially for supercritical processes. An alternative proof of this result is given by Guttorp (1991, Section 4.3)

Note that m is simply a linear combination of the entries of the vector $\boldsymbol{\pi}$, so the first two posterior moments of m can be readily obtained from those of $\boldsymbol{\pi}$. These are given in (3.3). Let $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ and $\Sigma = \Sigma(\boldsymbol{\beta})$ denote the posterior mean vector and variance-covariance matrix of $\boldsymbol{\pi}$, where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)'$. Then

$$E(m|\mathcal{Z}_n) = \boldsymbol{h}'\boldsymbol{\mu} = m(\boldsymbol{\mu})$$

and

$$\mathrm{Var}(m|\mathcal{Z}_n) = \boldsymbol{h}'\Sigma\boldsymbol{h} = \sigma^2(\boldsymbol{\mu})/(\beta+1), \tag{3.4}$$

with $\boldsymbol{h} = (1, \ldots, k)'$. Unfortunately, the full posterior marginal distribution of m is not available in closed form, even though m is just a linear function of a Dirichlet variate. However, conditional on the nonextinction of the process, the asymptotic normal approximation for $p(\boldsymbol{\pi}|\mathcal{Z}_n)$ induces a normal approximation to $p(m|\mathcal{Z}_n)$, with mean $m(\boldsymbol{\pi}^M)$ and variance $\sigma^2(\boldsymbol{\pi}^M)/\{\beta-(k+1)\}$. Other approximations will be considered in Section 3.3.

We close this section with a remark concerning the choice of the prior distribution for $\boldsymbol{\pi}$. Such a distribution should of course represent our beliefs about $\boldsymbol{\pi}$. In particular it must reflect our prior beliefs about the parameter of interest. Here we are mainly concerned with the classification of the process, so, even though the model is originally parameterized in terms of $\boldsymbol{\pi}$, the parameter of interest is just

$$\phi = \phi(\boldsymbol{\pi}) = \begin{cases} 1 & \text{if } m(\boldsymbol{\pi}) > 1 \\ 0 & \text{if } m(\boldsymbol{\pi}) \leq 1. \end{cases}$$

The prior for $\boldsymbol{\pi}$ must then satisfy the conditions $\Pr(\phi = 1) = p$ and $\Pr(\phi = 0) = 1 - p$, where the value of p is assigned *a priori* and represents our prior beliefs concerning the nature of the process. In applications, however, it is sometimes not feasible, or desirable, to assign an informative prior on $\boldsymbol{\pi}$. In such cases a *noninformative prior*, reflecting 'prior ignorance' in some sense, is often required. This prior must of course take account of the fact that ϕ is the parameter of interest. Specifically, it must be such that $p = 1/2$.

A number of methods for deriving noninformative priors have been proposed. These include Jeffreys' rule (Jeffreys, 1946; Box and Tiao, 1973), reference analysis (Bernardo, 1979; Berger and Bernardo, 1992), and limits of conjugate distributions (DeGroot, 1970, Chapter 10). Jeffreys' rule yields a (joint) noninformative prior for the entries of the parameter vector $\boldsymbol{\pi}$, without regard to the parameter of interest. Thus it is not appropriate for the classification of the process. Indeed, the Jeffreys prior for $\boldsymbol{\pi}$ under the model (2.1) is given by

$$p_J(\boldsymbol{\pi}) \propto \left\{ \sum_{i=0}^{n} m(\boldsymbol{\pi})^i \right\}^{1/2} \prod_{j=0}^{k} \pi_j^{-1/2},$$

and does not satisfy the condition $p = 1/2$. The reference analysis for this problem does distinguish between parameters of interest and nuisance parameters, but the calculations are much more involved.

A simple alternative is to use a vague prior on $\boldsymbol{\pi}$ corresponding to some limit of the conjugate density (3.1). If, for instance, we set $\alpha_j = \alpha_*$ $(j = 0, 1, \ldots, k)$, with $\alpha_* \approx 0$, the resulting prior is typically such that $\Pr(\phi = 1) \neq 1/2$. In fact, in this case $\Pr(\phi = 1)$ increases with k. It is therefore of interest to determine whether a conjugate prior can be found that represents vague prior knowledge about ϕ.

3.2 *A Vague Prior for* ϕ

Consider a reparameterization of the form $(m(\boldsymbol{\pi}), \boldsymbol{\theta}(\boldsymbol{\pi})')$, with $m(\cdot)$ as defined in Section 1, and where in general $\boldsymbol{\theta} : \mathbb{R}^k \to \mathbb{R}^{k-1}$. Denote by $p(m, \boldsymbol{\theta})$ the distribution induced on $(m, \boldsymbol{\theta})$ by the Dirichlet prior (3.1). Then we can in principle find the marginal distribution of m via

$$p(m) = \int p(m, \boldsymbol{\theta})\, d\boldsymbol{\theta}, \qquad 0 < m < k.$$

As pointed out in the previous section, however, this integral in not available in closed form. Nevertheless, if we write

$$p(m) = \int p(m|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

then a simple modal approximation yields

$$\hat{p}(m) = p(m|\hat{\boldsymbol{\theta}}),$$

where $\hat{\boldsymbol{\theta}}$ is the mode of $p(\boldsymbol{\theta})$. Such an approximation corresponds to a naive standard form of the Laplace approximation to $p(\boldsymbol{\theta})$; to some extent, the accuracy of $\hat{p}(m)$ depends on the form of $p(\boldsymbol{\theta})$ (see, for example, Kass, Tierney and Kadane, 1991). Loosely speaking, the more 'normal' $p(\boldsymbol{\theta})$ is, the more accurate $\hat{p}(m)$ can be expected to be. Note that $(\pi_1, \ldots, \pi_{k-1})$ follows a $(k-1)$-dimensional Dirichlet distribution with parameter $(\alpha_0 + \alpha_k, \alpha_1, \ldots, \alpha_{k-1})$. Thus a natural choice for $\boldsymbol{\theta}$ is $\boldsymbol{\theta} = \boldsymbol{\theta}(\pi_1, \ldots, \pi_{k-1})$, with

$$\theta_j = \log(\pi_j) - \log\left(1 - \sum_{l=1}^{k-1} \pi_l\right), \qquad j = 1, \ldots, k-1,$$

since in this case $p(\boldsymbol{\theta})$ becomes a generalized logistic distribution. This distribution is log-concave (hence unimodal) for all values of the parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)'$. Recalling that the likelihood (2.1) is proportional to a multinomial likelihood, we note that $(\pi_1, \ldots, \pi_{k-1})$ can then be identified with the mean-value parameter of a $(k-1)$-dimensional multinomial family. If the latter were to be written in exponential form, then $\boldsymbol{\theta}$ would correspond to the canonical parameter of the family. Now it follows from Section 3 of Gutiérrez-Peña (1992) that

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(E(\pi_1), \ldots, E(\pi_{k-1})).$$

Indeed $E(\pi_j) = \alpha_j/\alpha$ $(j = 0, 1, \ldots, k)$, with $\alpha = \sum_{j=0}^{k} \alpha_j$, so the mode of $p(\boldsymbol{\theta})$ is readily found to be

$$\hat{\theta}_j = \log(\alpha_j) - \log\left(\alpha - \sum_{l=1}^{k-1} \alpha_l\right), \qquad j = 1, \ldots, k-1.$$

Thus the Laplace approximation yields

$$\hat{p}(m) \propto \left\{k - \sum_{j=1}^{k-1} (k-j)\,(\alpha_j/\alpha) - m\right\}^{\alpha_0 - 1} \left\{m - \sum_{j=1}^{k-1} j\,(\alpha_j/\alpha)\right\}^{\alpha_k - 1}, \tag{3.5}$$

corresponding to the kernel of a nonstandard beta density with parameters α_0 and α_k, and supported on the interval

$$\left(\sum_{j=1}^{k-1} j\,(\alpha_j/\alpha), k - \sum_{j=1}^{k-1} (k-j)\,(\alpha_j/\alpha)\right). \tag{3.6}$$

The approximation $\hat{p}(m)$ suggests a vague prior on $\boldsymbol{\pi}$ corresponding to a noninformative prior for ϕ. First, note that it is natural to choose $\boldsymbol{\alpha}$ such that $\alpha_j \approx 0$ $(j = 1, \ldots, k-1)$, since in this case the support of $\hat{p}(m)$ corresponds approximately to the support of the true density $p(m)$, namely the interval $(0, k)$. Thus, essentially no possible value of m is excluded *a priori*. If, for simplicity, we now set $\alpha_0 = 1$, then

$\hat{p}(m)$ only depends on α_k. It is then easy to see that the choice $\alpha_k = \log(2)/\log(k)$ is such that $\widehat{\Pr}(m > 1) \approx 1/2$. Note that $\alpha_j \leq 1$ for all $j = 0, 1, \ldots, k$, which is conventionally regarded as consistent with a vague prior specification on $\boldsymbol{\pi}$.

It must be stressed that $\hat{p}(m)$ is just an asymptotic approximation to $p(m)$ as $\alpha \to \infty$, and so it cannot be expected to be accurate as an overall approximation to $p(m)$ for the small values of $\{\alpha_j\}$ selected above. Nevertheless, for these values of $\{\alpha_j\}$ the Dirichlet prior (3.1) satisfies $\Pr(m > 1) \approx 1/2$ regardless of the value of k, and the approximation is surprisingly accurate. Hence $\hat{p}(m)$ does provide a vague conjugate prior on $\boldsymbol{\pi}$ that is approximately noninformative for ϕ.

3.3 *Inferences about m*

Inferences about the nature of the process are based on the posterior marginal distribution of m. Specifically, the classification of the process can be carried out by calculating $\Pr(\phi = 1|\mathcal{Z}_n) = \Pr(m > 1|\mathcal{Z}_n)$. As pointed out in Section 3.1, however, even though the true posterior mean and variance of m can be found analytically, the full posterior distribution of m is not available in closed form. We now discuss some approaches to approximating $p(m|\mathcal{Z}_n)$.

Monte Carlo. Consider the parameterization discussed at the beginning of Section 3.2, and write

$$p(m|\mathcal{Z}_n) = \int p(m|\boldsymbol{\theta}, \mathcal{Z}_n)\, p(\boldsymbol{\theta}|\mathcal{Z}_n)\, d\boldsymbol{\theta}.$$

Then, given a sample $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}\}$ from $p(\boldsymbol{\theta}|\mathcal{Z}_n)$, the posterior marginal distribution of m can be approximated via the 'Rao-Blackwellized' Monte Carlo estimate

$$p(m|\mathcal{Z}_n) \approx \frac{1}{N} \sum_{t=1}^{N} p(m|\boldsymbol{\theta}^{(t)}, \mathcal{Z}_n) \tag{3.7}$$

(see, for example, Bernardo and Smith, 1994, Section 5.5.5). This approximation is generally highly accurate provided that N is sufficiently large. The parameter $\boldsymbol{\theta}$ should be chosen in such a way as to make sampling from $p(\boldsymbol{\theta}|\mathcal{Z}_n)$ easy. A natural choice here is $\boldsymbol{\theta} = \boldsymbol{\theta}(\pi_1, \ldots, \pi_{k-1}) = (\pi_1 \ldots, \pi_{k-1})'$, since in this case $p(\boldsymbol{\theta}|\mathcal{Z}_n)$ is just a $(k-1)$-dimensional Dirichlet distribution with parameters $(\beta_0 + \beta_k, \beta_1, \ldots, \beta_{k-1})$. Note that, if we sampled $\{\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(N)}\}$ from the posterior distribution of the full parameter vector $\boldsymbol{\pi}$, then not only could we compute (3.7) but also we could approximate essentially any other characteristic of the process. In particular, samples from the posterior distribution of m are readily obtained, from which $\Pr(m > 1|\mathcal{Z}_n)$ can be calculated. Inferences for other functionals of $\boldsymbol{\pi}$, such as σ^2, are also straightforward using this approach.

Monte Carlo approximations may be computationally expensive. However, for all practical purposes they can be regarded as exact and will be used here to assess the accuracy of the following analytic approximations.

Normal approximations. An alternative approach is to use (3.4) and a simple parametric model in order to approximate $p(m|\mathcal{Z}_n)$. One possibility, suggested by the form of the asymptotic posterior distribution of m, is the normal model. Using the minimum logarithmic divergence criterion (Bernardo and Smith, 1994, Section 3.4.3), it can be shown that the best normal approximation to $p(m|\mathcal{Z}_n)$ has mean and variance given by (3.4), namely the true posterior mean and variance of m. This approximation is well defined even if the process is subcritical, unlike the asymptotic normal approximation mentioned in Section 3.1.

Beta approximations. Another possibility is to use the modal approximation discussed in Section 3.2, but now applied to the posterior distribution $p(m, \boldsymbol{\theta}|\mathcal{Z}_n)$. This is obtained simply by substituting $\boldsymbol{\beta}$ for $\boldsymbol{\alpha}$ in (3.5) and (3.6). The approximation thus obtained is based on asymptotic arguments, and so may not be adequate unless the process does not become extinct and Y_n is sufficiently large. In particular, it fails to take into account the uncertainty about $\boldsymbol{\theta}$. Nonetheless (3.5) does suggest the use of an alternative beta model, now supported on the interval $(0, k)$. As in the normal case, the minimum logarithmic divergence criterion can be used to justify a nonstandard beta approximation with the same mean and variance as $p(m|\mathcal{Z}_n)$, and this nonasymptotic approximation is well-defined regardless of the nature of the process.

The three approximations discussed above will be illustrated and compared in Section 4.

3.4 *Predictive Inference*

Let Z_r, for $r > n+1$, denote the size of a future generation. Despite the fact that inferences about Z_r are of interest in a number of applications, only a few related results appear in the literature. In particular, Jagers (1975, Section 2.2) provides the following expressions for the conditional mean and variance of Z_r given the data $\mathcal{Z}_n$ and the value of $\boldsymbol{\pi}$.

$$E(Z_r|\boldsymbol{\pi}, \mathcal{Z}_n) = E(Z_r|m, Z_{n+1}) = Z_{n+1}\, m^{r-(n+1)}$$

$$\mathrm{Var}(Z_r|\boldsymbol{\pi}, \mathcal{Z}_n) = \mathrm{Var}(Z_r|m, \sigma^2, Z_{n+1}) = \begin{cases} Z_{n+1}\, \sigma^2 \left[\frac{m^{r-(n+2)}\,(m^{r-(n+1)}-1)}{m-1}\right] & \text{if } m \neq 1 \\ Z_{n+1}\, \sigma^2 \left[r-(n+1)\right] & \text{if } m = 1. \end{cases} \tag{3.8}$$

The standard frequentist approach involves replacing m and σ^2 by their respective estimates in (3.8). This procedure, however, tends to overestimate the precision of the prediction.

From a Bayesian point of view, inferences about Z_r should be based on the corresponding predictive distribution, $p(Z_r|\mathcal{Z}_n)$, for which no closed-form expression is available. A first, 'brute force' approach to calculating $p(Z_r|\mathcal{Z}_n)$ is to simulate, for each vector $\boldsymbol{\pi}^{(t)}$ in a sample $\{\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(N)}\}$ from $p(\boldsymbol{\pi}|\mathcal{Z}_n)$, $r-(n+1)$ generations of the process starting from Z_{n+1}. Recall from Section 3.1 that $p(\boldsymbol{Z}_i|Z_i, \boldsymbol{\pi})$ is a multinomial distribution with index Z_i and parameter $\boldsymbol{\pi}$. Let $Z_{n+1}^{(t)} = Z_{n+1}$ for all $t = 1, \ldots, N$. Now, for each $t = 1, \ldots, N$ and for $l = n+2, \ldots, r$, we can:

(a) sample $\boldsymbol{Z}_{l-1}^{(t)} = (Z_{l-1,0}^{(t)}, \ldots, Z_{l-1,k}^{(t)})'$ from $p(\boldsymbol{Z}_{l-1}|Z_{l-1}^{(t)}, \boldsymbol{\pi}^{(t)})$;
(b) set $Z_l^{(t)} = \sum_{j=1}^k j\, Z_{l-1,j}^{(t)}$.

In this way we obtain, for each $l = n+2, \ldots, r$, a sample $\{Z_l^{(1)}, \ldots, Z_l^{(N)}\}$ from $p(Z_l|\mathcal{Z}_n)$. These samples can then be used to approximate essentially any characteristic of the corresponding predictive distributions.

An alternative approach, pretty much in the spirit of the normal and beta approximations to $p(m|\mathcal{Z}_n)$ discussed in the previous section, is to use $E(Z_r|\mathcal{Z}_n)$ and $\mathrm{Var}(Z_r|\mathcal{Z}_n)$ and a simple parametric model to approximate $p(Z_r|\mathcal{Z}_n)$. Unfortunately, in this case the mean and variance of the predictive distribution cannot be obtained analytically. We note however that (3.8) suggest the following two approximations.

Monte Carlo. Given a sample $\{\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(N)}\}$ from $p(\boldsymbol{\pi}|\mathcal{Z}_n)$, the first two predictive moments of Z_r can be approximated via the Monte Carlo estimates

$$\begin{aligned}
E(Z_r|\mathcal{Z}_n) &\approx \frac{1}{N}\sum_{t=1}^N E(Z_r|m^{(t)}, Z_{n+1}) \\
\mathrm{Var}(Z_r|\mathcal{Z}_n) &\approx \frac{1}{N}\sum_{t=1}^N \mathrm{Var}(Z_r|m^{(t)}, \sigma^{2,(t)}, Z_{n+1}) \\
&+ \frac{1}{N}\sum_{t=1}^N E(Z_r|m^{(t)}, Z_{n+1})^2 - \left\{\frac{1}{N}\sum_{t=1}^N E(Z_r|m^{(t)}, Z_{n+1})\right\}^2,
\end{aligned}$$

where $\{m^{(1)}, \ldots, m^{(N)}\}$ and $\{\sigma^{2,(1)}, \ldots, \sigma^{2,(N)}\}$ are samples from the posterior distributions of m and σ^2, respectively, and are easily obtained from $\{\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(N)}\}$.

Analytic approximation. Note that $E(Z_r|\mathcal{Z}_n) = E[\,E(Z_r|m, Z_{n+1})\,]$, this latter expectation being with respect to $p(m|\mathcal{Z}_n)$. Expanding $E(Z_r|m, Z_{n+1})$ in a Taylor series about $\tilde{m} = E(m|\mathcal{Z}_n)$, and ignoring the cubic and higher order terms in the expansion before taking expectations, we get the following second order approximation to the predictive mean of Z_r

$$E(Z_r|\mathcal{Z}_n) \approx Z_{n+1}\left\{\tilde{m}^{r-(n+1)} + \frac{1}{2}[r-(n+1)][r-(n+2)]\,\tilde{m}^{r-(n+3)}\,\mathrm{Var}(m|\mathcal{Z}_n)\right\}.$$

A similar argument then yields the following first order approximation to the predictive variance

$$\mathrm{Var}(Z_r|\mathcal{Z}_n) \approx Z_{n+1}\,\mathrm{Var}(m|\mathcal{Z}_n)\,\tilde{m}^{r-(n+2)}$$

$$\times \quad \left\{ (\beta+1) \left(\frac{\tilde{m}^{r-(n+1)} - 1}{\tilde{m} - 1} \right) + [r-(n+1)]^2 \, Z_{n+1} \, \tilde{m}^{r-(n+2)} \right\}.$$

Either of the above approximations to the mean and variance can now be used in order to fit a simple parametric model to the predictive distribution $p(Z_r|\mathcal{Z}_n)$. We advocate the use of a gamma family, for which we have found strong empirical support based on a number of simulated examples under a wide range of initial conditions. Once more, the minimum logarithmic divergence criterion serves to justify the use of a gamma model with mean $E(Z_r|\mathcal{Z}_n)$ and variance $\mathrm{Var}(Z_r|\mathcal{Z}_n)$. The adequacy of this approximation is illustrated in Section 4.2.

4 EXAMPLES

In order to illustrate the methods discussed in this paper, both for the classification and the prediction problems, we now analyze two simulated examples. The first one concerns a subcritical process where the standard asymptotic theory does not apply. The second example deals with a supercritical model and compares the approximations proposed in Section 3.3. We also present a real data set previously discussed by Guttorp (1991). In all cases we use the vague conjugate prior introduced in Section 3.2.

4.1 *Classification*

Example 1. A subcritical process with $k = 3$, $Z_0 = 1$, and vector parameter $\boldsymbol{\pi} = (0.3, 0.1, 0.1)'$ was simulated. The true values of m and σ^2 are 0.8 and 0.96, respectively. The population extinguished after the second generation, and the relevant data are shown in Table 1.

Table 1: Data set for Example 1

Y_{20}	Y_{21}	Y_{22}	Y_{23}
1	1	0	0

Up to the second generation, $Y_2 = 2$ individuals were observed, and $Y_3 = 2$. The maximum-likelihood estimate for m is given by $\hat{m} = 0.5$, whereas the Bayesian analysis leads to $E(m|\mathcal{Z}_n) = 0.797$ and $\mathrm{Var}(m|\mathcal{Z}_n) = 0.260$. Figure 1 shows the posterior marginal density of m obtained via Monte Carlo using $50,000$ sample values. The normal and beta nonasymptotic approximations are also displayed, the latter clearly being better. In fact, the true probability $\Pr(m < 1|\mathcal{Z}_n)$ can be shown to be 0.71, whereas the corresponding beta approximation leads to the value 0.66. It should be stressed that this is an extreme case where the population not only becomes extinct but also the number of descendants is quite small.

Figure 1 about here

Fig 1: Posterior marginal density of m for Example 1

Example 2. In this case 10 generations of a process with $k = 5$, $Z_0 = 1$, and vector parameter $\boldsymbol{\pi} = (0.50, 0.05, 0.05, 0.05, 0.05)'$ were simulated. The true values of m and σ^2 are 1.2 and 1.76 respectively. Thus the process is supercritical. The relevant data are shown in Table 2.

Table 2: Data set for Example 2

$Y_{10,0}$	$Y_{10,1}$	$Y_{10,2}$	$Y_{10,3}$	$Y_{10,4}$	$Y_{10,5}$
137	238	29	29	24	24

Up to the $10th$ generation, $Y_{10} = 481$ individuals were observed, and $Y_{11} = 600$. The maximum-likelihood estimate for m is given by $\hat{m} = 1.245$. If the Bayesian approach is adopted we get $E(m|\mathcal{Z}_n) = 1.246$ and $\text{Var}(m|\mathcal{Z}_n) = 0.00369$. Figure 2 shows the Monte Carlo posterior marginal density of m, based on $20,000$ samples, together with the normal and beta approximations. For all practical purposes these three densities are indistinguishable from each other. This suggests that, even for moderate sample sizes, the simple analytical approximations discussed in Section 3.3 offer a reasonable alternative to the (computationally more expensive) Monte Carlo approach.

Note that $\Pr(m < 1|\mathcal{Z}_n)$ is essentially zero, so the process is classified correctly even though we only have a moderate amount of data in this case.

Figure 2 about here

Fig 2: Posterior marginal density of m for Example 2

Example 3. We now consider the Barnett data set analyzed by Guttorp (1991, Chapter 4), and described in Gani and Saunders (1977). These data are presented in Table 3 and relate to the number of cells with $j = 0, 1, \ldots, 8$ offspring obtained experimentally from a variety of yeast (*Saccharomyces cerevisiae*) in exponential growth phase.

Table 3: Barnett's data set

Y_{n0}	Y_{n1}	Y_{n2}	Y_{n3}	Y_{n4}	Y_{n5}	Y_{n6}	Y_{n7}	Y_{n8}
287	149	98	43	16	5	1	3	2

We note that no information regarding either the number of original ancestors, Z_0, or the number of generations, n, is available. At any rate, we have $Y_n =$

604. The maximum-likelihood estimate of m is $\hat{m} = 1.0033$, whereas the Bayesian analysis yields $E(m|\mathcal{Z}_n) = 1.0055$, $\mathrm{Var}(m|\mathcal{Z}_n) = 0.0028$ and $\Pr(m < 1|\mathcal{Z}_n) = 0.463$. Recalling that $\Pr(m < 1) = 0.5$, we see that the data provides only slight evidence in favour of the hypothesis of supercriticality. Figure 3 shows the posterior marginal density of m obtained via Monte Carlo using $15,000$ sample values. The normal and beta nonasymptotic approximations are also displayed. As in Example 1, all these densities lead essentially to the same results.

Figure 3 about here

Fig. 3: Posterior marginal density of m for Example 3

4.2 *Prediction*

Example 2. (cont.) In this case $Z_{11} = 119$. Using the available information from the 10 observed generations, the predictive distributions of Z_{12}, Z_{14} and Z_{16} were obtained via simulation. For each of these distributions, $20,000$ samples were simulated according to the procedure described in Section 3.4. The resulting histograms are shown in Figure 4, together with the corresponding gamma approximations. We used the Monte Carlo estimates of the means and variances to fit these densities. The gamma model provides remarkably good approximations, as can be verified from the respective $qq-$plots. Table 4 presents Monte Carlo estimates of $E(Z_r|\mathcal{Z}_{10})$ along with 95% prediction intervals, both for the sampling-based and gamma approximations to the predictive densities. As expected, these intervals are quite similar.

Table 4: Predictive summaries for Example 2

	Expected	Intervals	
	Value	Sampling-based	Gamma Model
Z_{12}	148	(118,181)	(118,181)
Z_{14}	232	(146,340)	(145,338)
Z_{16}	366	(187,626)	(179,616)

Figure 4 about here

Fig. 4: Predictive distributions and qq-plots for Example 2

Example 3. (cont.) Here, for illustrative purposes we assume that $Z_0 = 100$, so that $Y_{n+1} = 706$ and $Z_{n+1} = 102$. We simulated 15,000 samples from the predictive distributions of Z_{n+2}, Z_{n+4} and Z_{n+6}. Figure 5 shows the corresponding histograms, as well as the gamma approximations. Again, the gamma model provides an extremely good fit, as can be seen from the $qq-$plots. Predictive summaries for these data are presented in Table 5. As in Example 2, both the sampling-based approach and the gamma model lead essentially to the same 95% prediction intervals.

Table 5: Predictive summaries for Example 3

	Expected Value	Intervals	
		Sampling-based	Gamma Model
Z_{n+2}	103	(76,132)	(77,132)
Z_{n+4}	105	(57,167)	(57,166)
Z_{n+6}	108	(43,205)	(42,204)

Figure 5 about here

Fig. 5: Predictive distributions and qq-plots for Example 3

5 CONCLUDING REMARKS

One of the main inferential problems arising in the analysis of a Galton-Watson process concerns its classification into either subcritical or supercritical. The nature of the process can be determined simply by looking at the reproduction mean, m. From a frequentist viewpoint, however, making inferences about m is a rather difficult problem. Even though the maximum-likelihood estimator of m is readily available, its asymptotic distribution can only be obtained conditional on the nonextinction of the process, and depends on the unknown value of the reproduction variance. Therefore it is not possible to test the hypothesis of extinction using asymptotic arguments. Another relevant problem, which, in our opinion, has not been satisfactorily dealt with in the literature, is that of predicting the size of a future generation of the process.

On the other hand, from a Bayesian perspective it is possible to make inferences about essentially any functional of the offspring distribution, regardless of the nature of the process. In particular, the posterior distribution of m can be used to solve the classification problem. This distribution is not available in closed form, but can be accurately approximated using the simple beta density discussed in this paper. Inferences about the size of a future generation can be carried out in a similar fashion. Again, a closed form expression for the corresponding predictive distribution is not available but a gamma model provides a reasonable approximation. The accuracy of these approximations was assessed by comparison with the true distribution obtained via standard Monte Carlo methods.

In this article we have been concerned with the conjugate class of Dirichlet priors. In situations of vague prior information, the most common methods proposed in the literature to obtain noninformative priors produce distributions which do not belong to this class. As an alternative, we propose a pseudo-noninformative Dirichlet prior suggested by a Laplace approximation to the marginal density of m.

ACKNOWLEDGEMENTS

This work was partially supported by the Sistema Nacional de Investigadores, Mexico. The authors would like to thank an anonymous referee for several valuable comments which improved the presentation of the paper.

REFERENCES

Basawa, I.V. and Prakasa-Rao, B.L.S. (1980). *Statistical Inference for Stochastic Processes.* New York: Academic Press.

Bernardo, J.M. (1979). Reference Posterior Distributions for Bayesian Inference (with discussion). *Journal of the Royal Statistical Society B* **41**, 113-147.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.* Chichester: Wiley.

Berger, J.O. and Bernardo, J.M. (1992). On the Development of Reference Priors (with discussion). In *Bayesian Statistics 4,* eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith. Oxford: Clarendon Press, pp. 35-60.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis.* Reading: Addison-Wesley.

DeGroot, M.H. (1970). *Optimal Statistical Decisions.* New York: McGraw-Hill.

Diaconis, P. and Ylvisaker, D. (1985). Quantifying Prior Opinion (with discussion). In *Bayesian Statistics 2,* eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith. Amsterdam: North-Holland, pp. 133-156.

Dion, J.P. (1972). Estimation des Probabilités Initiales et de la Moyenne d'un Processus de Galton-Watson. *Doctoral Thesis.* University of Montreal, Montreal.

Dion, J.P. (1974). Estimation of the Mean and the Initial Probabilities of the Branching Processes. *Journal of Applied Probability* **11**, 687-694.

Dion, J.P. and Keiding, N. (1978). Statistical Inference in Branching Processes. In *Branching Processes,* eds. A. Joffe, and P. Ney. New York: Marcel Dekker, pp. 105-140.

Gani, J. and Saunders, I.W. (1977). Fitting a Model to the Growth of Yeast Colonies. *Biometrics* **33**, 113-120.

Gutiérrez-Peña, E. (1992). Expected Logarithmic Divergence for Exponential Families. In *Bayesian Statistics 4,* eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith. Oxford: Clarendon Press, pp. 669-674.

Guttorp, P. (1991). *Statistical Inference for Branching Processes.* New York: Wiley.

Harris, T.E. (1963). *The Theory of Branching Processes.* Berlin: Springer Verlag.

Heyde, C.C. (1979). On Assessing the Potential Severity of an Outbreak of a Rare Infectious Disease: A Bayesian Approach. *Australian Journal of Statistics* **21**, 282-292.

Jagers, P. (1975). *Branching Processes with Biological Applications.* New York: Wiley.

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society A* **186**, 453-461.

Johnson, R.A., Susarla, V. and van Ryzin, J. (1979). Bayesian Non-Parametric Estimation for Age-Dependent Branching Processes. *Stochastic Processes and their Applications,* **9**, 307-318.

Karlin, S. and Taylor, J.M. (1975). *A First Course in Stochastic Processes.* New York: Academic Press.

Kass, R.E., Tierney, L. and Kadane, J.B. (1991). Laplace's Method in Bayesian Analysis. In *Statistical Multiple Integration,* eds. N. Fluornoy and R.K. Tsutakawa. Providence: ASA, pp. 89-99.

Maki, E. and McDunnough, P. (1989). What Can or Can't Be Estimated in Branching and Related Processes? *Stochastic Processes and their Applications,* **31**, 307-314.

Mendoza, M. (1994). Asymptotic Normality under Transformations: a Result with Bayesian Applications. *Test* **3**, 173-180.

Prakasa-Rao, B.L.S. (1992). Nonparametric Estimation for Galton-Watson type Processes. *Statistics and Probability Letters* **13**, 289-293.