Sampling Methods for the Concentration Parameter of the Dirichlet Process

Lyric Yang Liu* Balgobin Nandram†

Abstract

There are many methods in current statistical literature for making inferences based on samples selected from a finite population. Parametric models may be problematic because statistical inference is sensitive to parametric assumptions. The Dirichlet process (DP) is very flexible and determines the complexity of the model. It is indexed by two hyper-parameters: the baseline distribution and concentration parameter. Current sampling methods for the concentration parameter only consider the continuous baseline distribution. We compare three different methods: Adaptive Reject Algorithm, Mixture of Gammas Method and Grid Method. We also propose a new method based on the ratio of uniforms. In practice, some survey responses are known to be discrete; if a continuous distribution is adopted as the baseline distribution, the model is misspecified and standard estimation/inference may be invalid. We propose a discrete baseline approach to the DP and conclude that the unobserved responses from the finite population can be sampled from a multinomial distribution if all possible outcomes are observed. We also applied our discrete baseline approach to a Phytophthora data set.

Key Words: Concentration Parameter, Discrete Baseline, Empirical Study, Grid Method, Nonparametric Bayesian Statistics

1. Introduction

We often know very little about the specific parametric forms of the distributions, and it is also difficult to validate the parametric assumptions. The parametric Bayesian models based on distributional assumptions may be problematic because inferences are sensitive to such assumptions. It may be more appealing to use a nonparametric Bayesian approach. The existence of the DP was established by Ferguson (1973). It is a distribution over distributions, that is, each draw from a DP itself is a distribution (i.e., we are working on functional spaces).

It is an open topic to sample the concentration parameter (α) of the DP. One can use Gilks' (1992) Adaptive Reject Sampling method which relies on the logconcavity of the logarithm transformation of α. Nandram and Yin (2016 a, b) used a grid method to sample α from the posterior density of $\rho = 1/(1+\alpha)$; they have used a noninformative prior for α, different from the proper (informative) prior suggested by Escobar and West (1995). Antonelli, Trippa and Haneuse (2016) reviewed several methods and suggested a more complex method. The problem of sampling the posterior density of α is a difficult one, and in this paper we will propose a new method called ratio of uniforms.

Another concern will be addressed is regarding the discreteness of the baseline distribution G_0. It is well-known that inference is sensitive to the specification of baseline measure (e.g., McAuliffe, Blei and Jordan 2006 and Nandram and Yin 2016 a). So it is more robust if we have an unspecified distribution G_0. However, the discreteness of G_0 means that the same value can come from either G_0 or from the balls already drawn in the Polya urn scheme. But it is mandatory to have G_0 discrete in this model if we have strong belief that

*Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609

†Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609

the observations are from a discrete family. In such case, the number of distinct values in the sample, k, is no long a sufficient statistics for α. This paper will correct this.

We will proceed this paper as follows. In Section 2, we briefly review the Dirichlet process (DP), different sampling algorithms for alpha, the concentration parameter; We also introduce our approach, the ratio of uniforms algorithm. In Section 3, we discuss one limitation that current literature has regarding the baseline distribution of the DP and how we resolve it, we also discuss the implementation of our method to the finite population mean. In Section 4, we run a small simulation study. In section 5 we discuss an illustrative example on Phytophthora data. We conclude this paper in section 6.

2. Dirichlet Process and Sampling the Concentration Parameter

2.1 Review of the Dirichlet Process

Let $(\Theta, \mathcal{B})$ be a measurable space, with G_0 a baseline measure (nonrandom) on the space, and let α be a positive real number. A Dirichlet process, DP(α, G_0), is defined as the distribution of a random probability measure G over $(\Theta, \mathcal{B})$ such that, for any finite measurable partition of the measurable space $(\Theta, \{A_i\}_{i=1}^n$,

$$\{G(A_1), \cdots, G(A_n)\} \sim \text{Dirichlet}\{\alpha G_0(A_1), \cdots, \alpha G_0(A_n)\}.$$

We write $G \sim \text{DP}(\alpha, G_0)$, if G is a random probability measure with a distribution given by the DP, where α is the concentration parameter. For any measurable set, A, we have $\text{E}[G(A)] = G_0(A)$, that is the mean of the DP is the baseline distribution G_0 and $\text{Var}[G(A)] = G_0(A)[1G_0(A)]/(\alpha + 1)$. The larger α is, the smaller the variance (i.e., the DP concentrates more of its mass around the baseline distribution). Here G_0 and α are both parameters and they play intuitive roles in the definition of the DP. Here G is constrained to be around G_0 and this is regulated by α.

Let $G \sim \text{DP}(\alpha, G_0)$ and $y_1, \cdots, y_n$ be a sequence of independent draws from G. The posterior distribution, $G|y_1, \cdots, y_n$ is

$$\text{DP}\left(\alpha + n, \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{y_i}\right),$$

where δ_{y_i} is the cdf of a point mass at y_i. This conjugate property of the DP was motivated by Ferguson (1973), desirable for easy algebra and computations.

For a one-sample problem, one might take

$$Y_1, \cdots, Y_n | G \sim G, G \sim DP(\alpha, G_0),$$

where G_0 is the baseline measure and α the concentration parameter. Assuming that there are k distinct values among $Y_1, \cdots, Y_n$, the baseline model is $Y_1^*, \cdots, Y_k^* | k \sim G_0$. Note that k is a random variable. The baseline measure G_0 is assumed continuous. Binder (1982) was the first to introduce this model to survey sampling; more recently, see Nandram and Yin (2016 a,b). Although G_0 can be discrete, it appears that this latter case was not discussed by Antoniak (1994).

Antoniak (1974) wrote down the distribution of k given α and he proved that k is a sufficient statistic for α. This is true when G_0 is continuous. It is easy to write down the posterior density with an appropriate prior. The sampling methods being discussed in this section are all based on continuous baseline.

However, if G_0 is discrete, k is no longer a sufficient statistic; this result appears to be not so well known. Therefore, if the result is used, this is a violation of the sufficiency principle; we will discuss this issue in Section 3.

2.2 Current Sampling Methods

We first review the Adaptive Reject Sampling method (Gilks 1992).
Theorem. Let $\phi = log(\alpha)$, where α is the concentration parameter. The posterior density $\pi(\phi|k)$ is logconcave, (i.e., strongly unimodal with a unique mode).
The proof can be found in the appendix. Knowing that $\pi(\phi|k)$ is logconcave, we can use the Adaptive Reject Sampling method (Gilks 1992) to draw ϕ. This sampling procedure was realized with the R package `ars`. Then we can compute α in the form $\alpha = e^{\phi}$. The algorithm is as follows:

1. Initialize n and S_n
2. Generate $X \sim g_n(x)$, $U \sim U(0,1)$
3. If $U \leq \frac{f(x)}{\bar{w}_n g_n(x)}$, accept x. Otherwise, update S_n to $S_{n+1} = S_n \cup \{x\}$. and repeat.

Nandram and Choi (2004) discussed the use of the gamma prior which was introduced earlier by Escobar and West (1995). One concern is that the mix of Gamma method gives bimodal sampling distribution whereas, we prefer the unimodal density of α.

Nandram and Yin (2016) transformed α according to $\rho = \frac{1}{1+\alpha}$, this is also the correlation in the DP. The posterior density of ρ is

$$\pi(\rho|k) \propto \frac{(1-\rho)^{k-1}\rho^{n-k}}{\prod_{j=1}^{n-1}(1-\rho+\rho j)}, \;\; 0 \leq \rho \leq 1.$$

We see that it is not in a simple form and a one-dimensional grid method was used to draw samples from it, thereby avoiding Markov chain Monte Carlo methods. The unit interval is simply divided into 100 sub-intervals of equal width, and the joint posterior density is approximated by a discrete distribution with probabilities proportional to the heights of the continuous distribution at the mid-points of these sub-intervals. Now, it is easy to draw a sample from this univariate discrete distribution of $\pi(\rho|k)$. The algorithm goes as follows:

1. Draw a number U between $[0,1]$ with probability proportional to the heights of the intervals.
2. Draw x uniformly from $[U - w/2, U + w/2]$, where w is the width of the interval

Nonetheless, there is drawback of this method, because it does not perform well near the tail.

2.3 Ratio of Uniforms Method

Original introduced by Kingderman and Monahan (1977), a point is generated uniformly over a certain region in the plane. To realize this, independent uniform random variables are simulated, U and V say, and those that fall outside some set are discarded. The ratio V/U is then calculated for those points inside the set. The ratio values obtained are used as observations from the required distribution. This method can proceed using the following algorithm: Suppose we want to draw samples from $h(x)$, an unknown distribution.

1. Generate u and v independently from $U(0,b)$ and $U(c,d)$.
2. Set $x = v/u$ if $u^2 \leq h(v/u)$ and return to (i) otherwise.

Here b, c and d are given by

$$b = \sup_x \sqrt{h(x)} \quad c = -\sup_x x\sqrt{h(x)} \quad d = \sup_x x\sqrt{h(x)}$$

Because α is positive, we can restricted $c = 0$. This algorithm is very easy to implement and very efficient to get samples.

3. Baseline Consideration

3.1 A Problem

Current literature has been using continuous baseline distributions, see Teh, Jordan, Beal and Blei(2006), Antonelli, Trippa and Haneuse(2016). Here we explored a possibility of using a discrete baseline. One problem is that the distinct values in the sample is no longer the true distinct ones because for discrete baseline, we allow observing a "new" value from the baseline distribution that is the same as one that is already in the sample. To solve this problem, we introduce a latent variable $Z_i \sim \text{Ber}(\frac{\alpha}{\alpha+i-1})$, with

$$Z_i = \begin{cases} 1, & \text{if a draw is made from the baseline,} \\ 0, & \text{if a draw is from the value that is already observed.} \end{cases}$$

The true number of distinct values k is the sum of z_i, $k = \sum_{i=1}^{n} z_i$.

3.2 Finite Population Prediction

Suppose we want to predict the finite population proportion for a given area based on a random sample from it. This could be applied to many areas of study, for example we want to predict the infectious rate of the given farmland for some disease and it is not feasible to observe all the plant on the farm, however, we could take a random sample and estimate the posterior mean using this sample. We have observed n of them and want to make predictions to the $N - n$ individuals. Consider following scenarios:

Scenario 1. We use the one-level DP model for the population values to make inference for a finite population mean. For this case, the baseline distribution is chosen to be normal. We assume that

$$y_1 \cdots, y_N | G \sim G$$
$$G \sim DP(\alpha, G_0)$$
$$G_0 \sim \ N(\mu, \sigma^2)$$

Scenario 2. We use the one-level DP model for the population values to make inference for a finite population mean. For the one-level DP model we assume that

$$y_1 \cdots, y_N | G \sim G$$
$$G \sim DP(\alpha, G_0)$$
$$G_0 \sim \ Bin(m, p), \ \text{ with } m \text{ known}$$

Scenario 3. Using the model in scenario 2, but suppose we have already observed $y_1^*, \cdots, y_d^*$, d distinct values ($1 \leq d \leq n$), with $n_1, \cdots, n_d$ being their corresponding counts. Now we want to predict $N_1 - n_1, \cdots, N_d - n_d$, for convenience, we write $N_1^*, \cdots, N_d^*$. Let $N^* = N - n$, Now

$$N_1^*, \cdots, N_d^* \sim \text{Multinomial}\Big\{N^*, (w_1, \cdots, w_d)\Big\},$$

where $w_1, \cdots, w_d$ are the weights in stick-breaking algorithm with $\sum_{s=1}^{\infty} w_s = 1$ and $w_1 = \nu_1, w_2 = \nu_2(1-\nu_1)\cdots, \nu_i \overset{iid}{\sim}$ Beta $(1, \alpha)$. Given α from the Gibbs sampler, we can draw ν_i and thus draw the predicted values from a Multinomial distribution. This is based on Sethuraman(1994).

3.3 The Sampling Procedure

We will describe the sampling procedure for the three scenarios mentioned in Section 3.2 respectively.

Scenario 1. Using a normal cdf as the baseline distribution, we can observe the number of distinct values k and then sample α as discussed in Section 2. For each sampled $alpha$ value, we predict the unobserved $Y_{n+1}, \cdots, Y_N$ using the Polya urn scheme,

$$Y_{n+i+1}|y_1, \cdots, y_n, y_{n+1}, \cdots, y_{n+i} \sim \frac{\alpha}{\alpha+n+i} G_0 + \frac{n+i}{\alpha+n+i} \sum_{j=1}^{n+i} \delta_{y_j},$$

for $i = 1, \cdots, N-n-1$. Now it is easy to draw the unsampled values one by one using the equation.

Scenario 2. We correct the true number of observations from the baseline distribution $k^{'} = \sum_{i=1}^{n} z_i$. And the Gibbs sampler is as follows:

1. $Z_i|\alpha, p \sim \text{Ber}(\frac{\alpha}{\alpha+i-1})$
2. $\pi(\alpha|\boldsymbol{z}, p) \propto \frac{\alpha^{k^{'}}}{\prod_{j=1}^{n-1}(j+\alpha)} \cdot \frac{1}{(1+\alpha)^2}$
3. $p|\boldsymbol{z}, \alpha = \frac{1}{k'} \sum_{\{i:z_i=1\}} y_i$

Scenario 3. From the sample, we have distinct $y_1^*, \cdots, y_d^*$, with corresponding weights $w_1, \cdots, w_d$, where
$w_1 = \nu_1, w_2 = \nu_2(1-\nu_1)\cdots, w_d = \prod_{s=1}^{d-1}(1-\nu_s), \nu_i \overset{iid}{\sim}$ Beta $(1, \alpha)$. So

$$\begin{aligned}\pi(\boldsymbol{\nu}|\alpha, d) \propto\ & \nu_1^{n_1}[\nu_2(1-\nu_1)]^{n_2} \cdots [\nu_{d-1}(1-\nu_1)\cdots(1-\nu_{d-2})]^{n_{d-1}} \\ & [(1-\nu_1)\cdots(1-\nu_{d-2})(1-\nu_{d-1})]^{n_d} \times \prod (1-\nu_i)^{\alpha-1} \\ \propto\ & \nu_1^{n_1}(1-\nu_1)^{n_2+\cdots+n_d+\alpha-1} \nu_2^{n_2}(1-\nu_2)^{n_3+\cdots+n_d+\alpha-1} \cdots \nu_{d-1}^{n_{d-1}}(1-\nu_{d-1})^{n_d+\alpha-1}.\end{aligned}$$

Using a Gibbs sampler,

1. Draw ν_i from $\pi(\nu_i|\alpha, d)$
2. $\pi(\alpha|d) \propto \frac{\alpha^d}{\prod_{j=1}^{n-1}(j+\alpha)} \cdot \frac{1}{(1+\alpha)^2}$

4. Simulation Study

It is convenient to compare different sampling methods using simulations because we can obtain the true distribution of α and compare the theoretical values with the sampled values. Firstly we find the theoretical percentiles of α using fine grids of width 0.0025. Then we perform the four sampling methods to get 10,000 sample points. We can find the sample percentiles by ordering the sample values and find corresponding quantiles as the theoretical values. Lastly, we compare the theoretical value vs. the sampled value using a

quantile-quantile plot. Result is shown in Figure 1. All four methods provide reasonable sampling distribution for α. However, as we mentioned in Section 2, the ARS and grid method have tail problems and mixture of gamma uses an informative prior which remains to be validated. Our method does not require informative gamma prior and is faster and easy to implement. So we recommend using ratio of uniforms to sample α

5. Real Data Analysis

The data we present here are about Phytophthora Epidemic in Bell Pepper Gumpertz (1997). The pathogen Phytophthora Capsici Leonian causes lesions on the crown, stem, and leaves of bell pepper, and rapidly causes the plant to die. For their analyses, they took one field which was a square lattice of 20 by 20 quadrats with 2 to 3 bell pepper plants per quadrat as an example. The response variable within each quadrat was presence or absence of disease in a quadrat. If any plant was wilted, dead, or had lesions on stem, crown, or leaves, disease was considered to be present in the quadrat. Disease presence or absence was recorded for each quadrat on nine dates throughout the growing season, from 6/16/92 to 8/5/92. Figure 2 shows the disease incidence on 6/25/92. We want to make this data set usable to mimic our discrete response scenario so we perform the following sampling procedure: we divide each row of the field by every five quadrats; and then we take one random sample within each row of the field. We assume that the sampled value follows a binomial distribution with total number of trials being 5. Now our goal is to predict the unobserved quadrats and estimate the infectious rate, which is the success probability for this binomial distribution. We performed the estimation using both discrete baseline and continuous baseline approach discussed in Section 3.2.

We report the posterior mean and the credible interval in Table 1. We found that the continuous baseline distribution provides a biased estimation to the infectious rate comparing to the discrete baseline approach we propose. Also the former is not precise as the latter with a wider credible interval. This result can be seen in Figure2, with red line indicating the true infectious rate.

6. Conclusions

We have proposed a new sampling method for the concentration parameter of the Dirichlet Process and compared it with other three exited methods. Our method performs as well as other methods and it is faster considering the computational time. In the mean time, we pointed out a problem that current researchers have ignored regarding the baseline distribution of the DP. We have corrected the true number of distinct values in the sample by introducing a latent variable which indicated which urn a new observation is from. By using this approach, we are able to give a more accurate estimation of the finite population mean when the observations are discrete. We used a Phytophthora example to illustrate our approach. And concluded the discrete baseline method is more reasonable.

There are two directions we could proceed to extend our current work. First, we could consider spatial model for the example provided in this paper. Also, we could easily extend the one-level DP model to a two or multilevel DP model. The hierarchical structure will make our inference more robust and accurate.

REFERENCES

Antoniak, C. E. (1974), "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2 (6), 1152-1174.

Blackwell, D. and MacQueen, J. B. (1973), "Ferguson distributions via Polya urn schemes," *The Annals of Statistics*, 1 (2), 353-355.

Antonelli, J., Trippa, L. and Haneuse, S. (2016), "Mitigating bias in generalized linear mixed models: The case for Bayesian nonparametrics," *Statistical Science*, 31 (1), 80-95.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90 (430), 577-588.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1 (2), 209-230.

Ishwaran, H. and James, L. F. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96 (453), 161-173.

Gumpertz, M. L. , Graham, J. M. and Ristaino,J. B. (1997), "Autologistic Model of Spatial Pattern of Phytophthora Epidemic in Bell Pepper: Effects of Soil Variables on Disease Presence," *Journal of Agricultural, Biological, and Environmental Statistics*, Vol. 2, No. 2, pp. 131-156

Kalli, M. and Griffin, J. E. and Walker, S. G. (2011), "Slice sampling mixture models," *Statistics and Computing*, 21 (1), 83-105.

Kinderman, A.J., Monahan J.F.(1977), "Computer Generation of Random Variables Using the Ratio of Uniform Deviates," *Association for Computing and Machinery, Inc.*

Nandram, B. and Choi, J. W. (2004), 'Nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse," *Journal of Nonparametric Statistics*, 16 (6), 821-839.

Nandram, B. and Yin, J. (2016a), "Bayesian predictive inference under a Dirichlet process with sensitivity to the normal baseline," *Statistical Methodology*, 28, 1-17

Nandram, B. and Yin, J. (2016b), "A nonparametric Bayesian prediction interval for a finite population mean," *Journal of Statistical Computation and Simulation*, 86 (16), 3141-3157.

Neal, R. M. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 9 (2), 249-265.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639-650.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006), "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, 101 (476), 1566-1581.

Appendix:

proof. Consider a "Cauchy" type prior for α, also called a shrinkage prior, of the form, $p(\alpha) = \frac{1}{(1+\alpha)^2}, \alpha > 0$.

$$\pi(\alpha|k) \propto \frac{\alpha^k}{\prod_{j=1}^{n-1}(j+\alpha)} \cdot \frac{1}{(1+\alpha)^2}$$

Letting $\alpha = e^{\phi}$,

$$\pi(\phi|k) \propto \frac{e^{k\phi}}{\prod_{j=1}^{n-1}(j+e^{\phi})} \cdot \frac{1}{(1+e^{\phi})^2}$$

Now we need to show that $\pi(\phi|k)$ is logconcave.

$$\log \pi(\phi|k) = k\phi - \sum_{j=1}^{n-1} \log(j+e^{\phi}) - 2\log(1+e^{\phi})$$

$$\frac{\partial}{\partial \phi} \log \pi(\phi|k) = k - \sum_{j=1}^{n-1} \frac{e^{\phi}}{j+e^{\phi}} - \frac{2 \cdot e^{\phi}}{1+e^{\phi}}$$

$$\frac{\partial^2}{\partial \phi^2} \log \pi(\phi|k) = -\sum_{j=1}^{n-1} \frac{je^{\phi}}{(j+e^{\phi})^2} < 0$$

And we conclude that the posterior density $\pi(\phi|k)$ is logconcave. □

Table 1: Estimation of the Infectious Rate (True Rate:0.1525)

Baseline Distribution	Posterior Mean	95% Credible Interval
Normal (μ, σ^2)	0.1691	(0.047,0.294)
Binomial (5,p)	0.1502	(0.114,0.174)

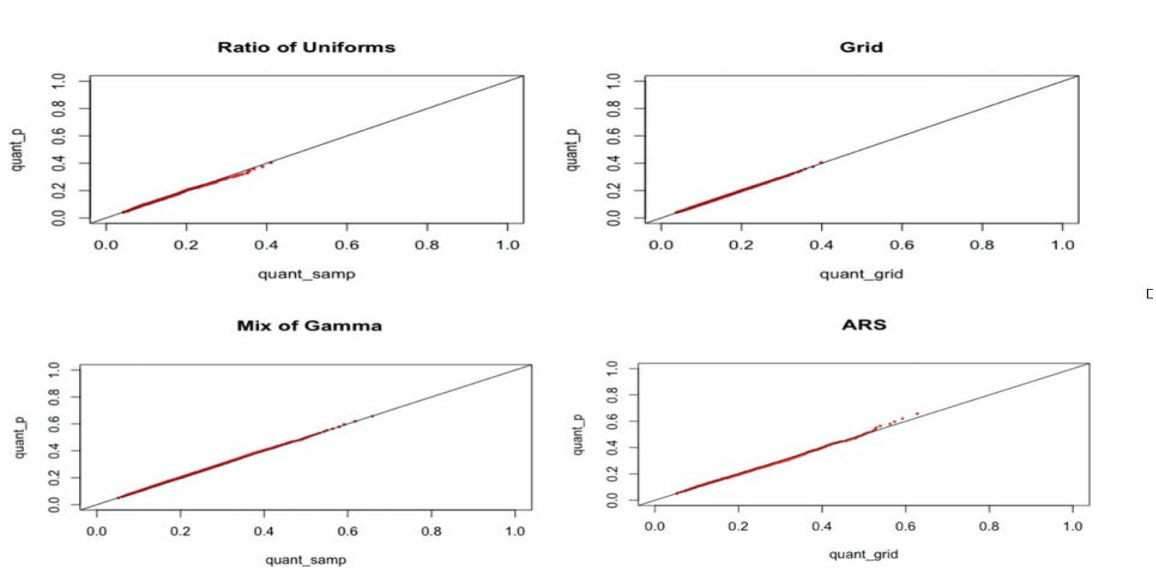

Figure 1: Quantile-Quantile Plot of Different Sampling Methods.

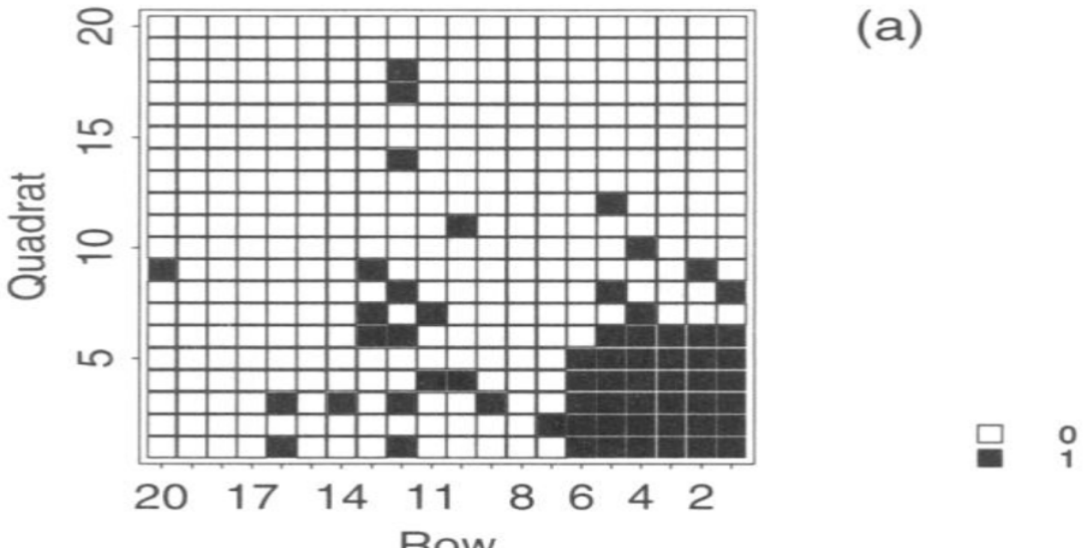

Figure 2: Map of Disease Incidence.

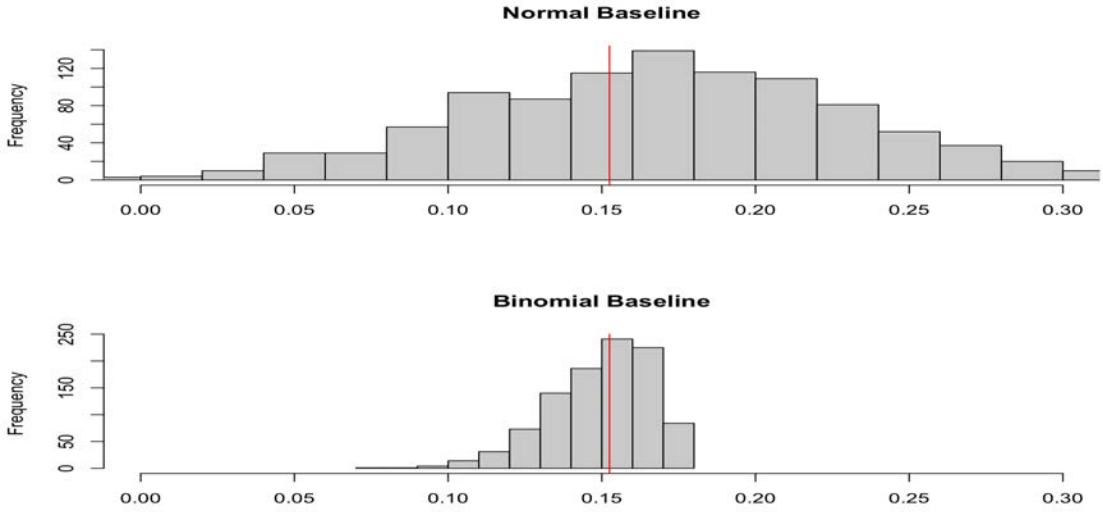

Figure 3: Posterior Distribution of Mean from the Two Baseline DP Models