Bayesian Inference of a Finite Population Mean Under Length-Biased Sampling

Jens Kinch og Morten Haubro
Niels Bohr Institute, University of Copenhagen
(Dated: March 6, 2019)

This write-up is a summary of an article by Xu, Nandram and Manandhar entitled "Bayesian Inference of a Finite Population Mean Under Length-Biased Sampling". It has to our knowledge not been published in any scientific journal

The purpose of the paper by Xu, Nandram and Manandhar is to develop a method to characterise a finite population of shrubs of unknown number and size with only a very small sampling. This is to better be able to estimate the regrowth in a quarry. The hope is that this can be achieved by transecting the quarry only a few times and assuming that the measurement of a shrub is biased towards the larger shrubs. A simplified schematic of the sampling technique can be seen in figure 2. Two independent replications are performed, each with three different transects.
It is assumed that the shrubs are distributed according to a generalised gamma function,

$$f(x|\alpha,\beta,\gamma) = \frac{\gamma x^{\gamma\alpha-1}}{\beta^{\gamma\alpha}\Gamma(\alpha)} \exp\left[-\left(\frac{x}{\beta}\right)^{\gamma}\right],\ x>0 \qquad (1)$$

This is what the paper calls the unweighted pdf. The paper assumes that the probability of one shrub being counted is proportional to it's width perpendicular to the transect direction (x. The binary value I = 0,1, denotes whether or not a particular shrub is counted. $P(I=1,x) = Cx$, $C = 1/w$ where w is the length of the base line (figure 2).
Using this as a prior it is possible to obtain the weighted pdf or the sample distribution $P(x|I=1) = g(x|\alpha_g,\beta_g,\gamma_g) = f(x|\alpha+\frac{1}{\gamma},\beta,\gamma)$.
A best estimate for the total number of samples in the finite population is obtained by using the Horvitz-Thompson unbiased estimator,

$$\hat{N}_i = \sum_i \frac{1}{Cx_i} = w\sum_i \frac{1}{x_i}. \qquad (2)$$

A Bayesian posterior is then created as,

$$\pi(N_i|n_i,\mu_0) = \frac{P(n_i|N_i,\mu_0)P(N_i)}{\int P(n_i|N_i,\mu_0)P(N_i)\mathrm{d}N_i}, \qquad (3)$$

where $P(n_i|N_i,\mu_0) \sim Binom(n_i|N_i,\mu_0)$. The second replicate is used to find an estimate of $\hat{N}_i$ and in turn estimate μ_0. The most likely value of $\pi(N_i|n_i,\mu_0)$ is $E(N_i|n_i,\mu_0) = \frac{n_i}{\mu_0} = \hat{N}_i$, which yields $\mu_0 = 0.0046$.

For a single transect the probability of being sampled is $p_\mathrm{s} = x_i/w$, then the probability for not being sampled is: $p_\mathrm{ns} = 1 - x_i/w$. we define I_j:

$$I_j = \begin{cases} 1 & \text{if } j \le n, \\ 0 & \text{if } j > n. \end{cases}$$

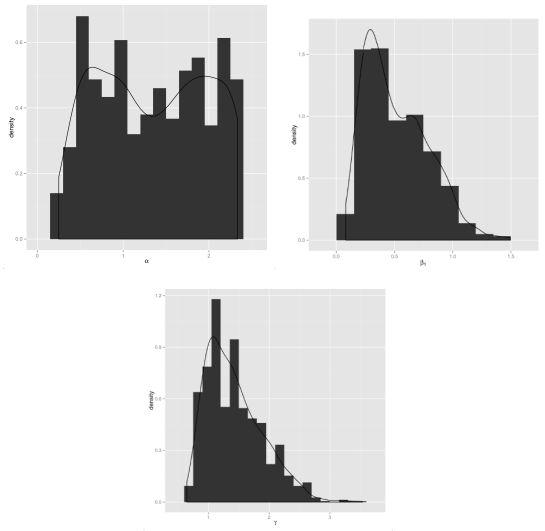

FIG. 1: Posterior distributions for α, β and γ.

Then:

$$\pi(I_j,x_j) \propto \left[\frac{x_j}{w}f(x_j)\right]^{I_j}\left[\left(1-\frac{x_j}{w}\right)f(x_j)\right]^{1-I_j}$$

so:

$$\pi(x_j|I_j=0) = \frac{\left(1-\frac{x_j}{w}\right)f(x_j)}{\int\left(1-\frac{x_j}{w}\right)f(x_j)\mathrm{d}x_j}$$

We are now looking for the posterior parameter distribution or,

$$\pi(\alpha,\beta,\gamma|x_s) = \frac{g(x_s|\alpha,\beta,\gamma)P(\alpha,\beta,\gamma)}{\int g(x_s|\alpha,\beta,\gamma)P(\alpha,\beta,\gamma)\mathrm{d}\alpha\mathrm{d}\beta\mathrm{d}\gamma}. \qquad (4)$$

$P(\alpha,\beta,\gamma) = \frac{1}{\beta}\frac{1}{(1+\alpha)^2}\frac{1}{(1+\gamma)^2}$, this is known as the shrinkage prior and essentially it reduces the effects of overfitting. This is to increase the prediction power of the model.
Since the posterior pdf for the parameters has been constructed, it is now possible to sample this using Markov-chain Monte Carlo and obtain posterior distributions for all the parameters, the distributions are shown in figure 1.

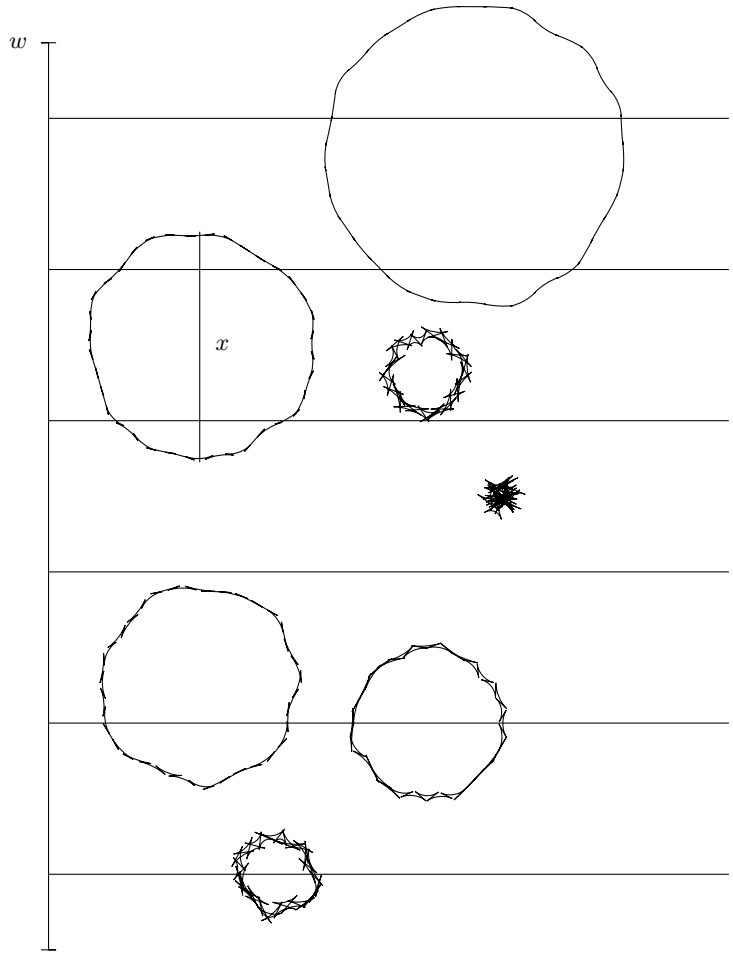

FIG. 2: Sketch of the method used to collect data; six transects perpendicular to a line of length w, data on the widths (x) of all shrubs hit by these lines is collected.