NONINFORMATIVE NONPARAMETRIC BAYESIAN ESTIMATION OF QUANTILES

Glen Meeden*
School of Statistics
University of Minnesota
Minneapolis, MN 55455

Appeared in *Statistics & Probability Letters*
Volume 16 (1993) 103-109

*Research supported in part by NSF grant DMS-8911548-01

ABSTRACT

In 1981 Rubin introduced the Bayesian bootstrap and argued that it was the natural Bayesian analogue to the usual bootstrap. We show here that when estimating a population quantile in a nonparametric problem it yields estimators that are often preferred to the natural naive estimators based on the order statistic.

AMS 1980 Subject Classification: 62G05, 62C15 and 62G30
Key Words: Bayesian Bootstrap, noninformative Bayes, nonparametric estimation, order statistic, and quantile.

1 Introduction

The Bayesian bootstrap was introduced in Rubin (1981) as a natural analogue of the bootstrap. Given the data it yields a 'posterior distribution' for the parameter of interest even though no prior distribution was actually specified. Because it is based on somewhat peculiar model assumptions Rubin suggested that in some situations it could lead to unreasonable inferences. In Meeden and Ghosh (1983) a new proof was given of the fact that the sample mean is an admissible estimator of the population mean in finite population sampling. This proof was based on the stepwise Bayes technique. Given the sampled units it yields a 'posterior distribution' for the unsampled members of the population. This 'posterior distribution' was called the polya posterior and is the Bayesian bootstrap adapted to the finite population setting. (See also Lo (1988).)

In Meeden and Vardeman (1991) a general approach to finite population sampling based on the 'polya posterior' was discussed. There it was observed that for estimating a population median an estimator based on the 'polya posterior' often was preferred to the sample median. This suggests that for nonparametric estimation of a quantile an estimator based on the Bayesian bootstrap could often be preferred to the usual naive estimator based on the order statistic. In section 2 we briefly review the Bayesian bootstrap and a related method, due to Banks (1988) called the smooth Bayesian bootstrap. A new modification of Rubin's Bayesian bootstrap is also introduced. In section 3 the results of various simulations are presented which show that all three techniques led to estimates which are preferred to the usual estimator.

2 Bayesian bootstraps

Let $X = (X_1, \ldots, X_n)$ where the X_i's are independent and identically distributed random variables with common distribution function F. We assume that F belongs to Θ, a large nonparametric family of possible distributions. Let q=q(F) denote the qth quantile of F. We consider the problem of using X to estimate q for some fixed q$\in (0,1)$. Let $b = (b_1, \ldots, b_r)$ be a vector of r distinct real numbers arranged in increasing order. Let $\Theta(b)$ denote all the distribution functions which concentrate all their mass on $b_1, \ldots, b_r$. We assume that for every r and for every b, $\Theta(b)$ is a subclass of Θ. If F is assumed to be-

long to just $\Theta(b)$ then X is a random sample from a multinomial$(1;p_1,\ldots,p_r)$ population where $p_i = P(X_j = b_i)\, for\, i = 1,\ldots,r\, and\, j = 1,\ldots,n$. Note that $\Theta(b)$ is equivalent to the (r-1)-dimensional simplex

$$\Lambda(r) = \{p = (p_1,\ldots,p_r) : p_i \geq 0\, for\, i = 1,\ldots,r\, and \sum_{i=1}^{n} p_i = 1\}$$

Finally, given $X = x$ is observed let $T(x) = t$ be the vector of distinct values appearing in x, arranged in increasing order and $V(x) = v$ be the count vector, i.e. v_i is the number of times t_i appears in the sample. If the members of x are all distinct then t is just the order statistic of x and v is a vector of 1's.

We now give a noninformative Bayesian method for estimating q=q(F) when F is assumed to belong to Θ. Assume X=x has been observed and T(x)=t and V(x)=v are known. Let r be the length of x. After X=x is in hand we will assume that the unknown F must belong to $\Theta(t)$. Furthermore we will take as our 'posterior distribution' over $\Theta(t)$, or equivalently over $\Lambda(r)$ the Dirichlet distribution with parameter vector v. This 'posterior' leads in turn to a 'posterior distribution' for q=q(F) which then can be used in estimating q=q(F). The above is just the Bayesian bootstrap of Rubin(1981) applied to the problem of estimating a quantile.

We believe that it cannot be a true posterior distribution for F, i.e. that it arises in the usual way from a prior distribution over Θ. Moreover it might seem silly to choose a distribution which is restricted to the subclass $\Theta(t)$ of Θ. It is easy to check that under this 'posterior' the predictive distribution of an additional observation from the population is just the empirical distribution function. This gives zero probability to the event that a future observation is unequal to the observed values in the sample. It was just this fact that lead Rubin to question the broad applicability of the Bayesian bootstrap. On the other hand, for a large sample size it does seem to give a possible noninformative Bayesian approach to this nonparametric problem. Moreover it does arise naturally from the stepwise Bayes point of view and leads to admissible estimators in nonparametric problems, see Meeden et al. (1985). Following their argument it yields admissible estimators of q=q(F) when the parameter space is $\Theta\, or\, \Theta(b)$ for every choice of b.

Suppose $X_1, X_2, \ldots$ are real valued random observations where ties have probability 0. Hill introduced the assumption A_n which asserts that conditional on the first n observations the next observation X_{n+1} is equally likely

to fall in any of the $n+1$ open intervals defined by the successive order statistics in the sample. Furthermore if the sequence is exchangeable and γ_i is the proportion of future observations falling in the ith interval then A_n implies that conditional on the first n observations the distribution for the vector of $\gamma_i's$ is Dirichlet with a parameter vector of all ones. This argument is made under the assumption that the probability measures involved are countably additive. But then Hill shows that there is no countably additive probability distribution on the space of observations such that the conditional distributions will agree with A_n.

From some points of view A_n is a more attractive formulation for a non-informative nonparametric Bayesian approach then the Bayesian bootstrap. For one thing it does not assume that any future observation must equal one of the past observations. On the other hand even if one is willing to assume that given the first n observations the conditional distribution of the $\gamma_i's$ is Dirichlet with a parameter vector of all ones it is not clear how the probability assigned to each interval should be spread out over the interval. One obvious possibility is to distribute the probability uniformly over each interval. Banks (1988) called this approach the smooth Bayesian bootstrap and compared it to the usual Bayesian bootstrap and several other bootstrap methods as well. Based on simulation studies he declared the smooth Bayesian bootstrap the winner. In part this was based on the fact that confidence intervals based on the smooth Bayesian bootstrap were slightly more accurate than those based on other methods.

As noted above one theoretical justification for the Bayesian bootstrap is that it is a stepwise Bayes procedure and hence will led to admissible procedures. Lane and Sudderth (1978) show that Hill's A_n is consistent with a finitely additive probability measures on the space of observations. This in turn shows that A_n is coherent. (Coherence is a slightly weaker property than admissibility since a procedure may be coherent but not admissible.) This however is not a justification for the smooth Bayesian bootstrap since their work does not directly compute the necessary conditional distributions. In fact the results of section 7 of their paper suggests that the smooth Bayesian bootstrap does not arise from some finitely additive prior distribution.

We now suggest another modification of the Bayesian bootstrap which is a smoother version of Rubin's Bayesian bootstrap and is also a stepwise Bayes procedure. We suppose that for every member of Θ its support is a subset of the finite interval $I = (c, d)$. Let $g = (g_1, \ldots, g_k)$ be a grid defined

on I, i.e. $c = g_1 < \cdots < g_k = d$. We assume that such a grid g is given and fixed. Let $\Theta(g)$ be the subset of Θ which contains all the density functions on I which are constant on each subinterval of I defined by the grid g. After X=x is observed let S(x)=s be the subintervals of the grid g which contain at least one observation. Let W(x)=w be the count vector, i.e w_i is just the number of observations that fell in subinterval s_i. Now given a sample X=x we will assume that our 'posterior' is concentrated on the subset of $\Theta(g)$ which consists of all those densities whose support is contained in S(x)=s. The 'posterior' distribution of the amount of probability assinged to each of these subintervals will just be Dirichlet with parameter vector W(x)=w. Then by assumption these 'posterior' probabilities are uniform over their respective subintervals. It is easy to check that this 'posterior' is in fact a stepwise Bayes procedure. The argument is exactly the same as the one used for the Bayesian bootstrap except it is now applied to subintervals defined by the grid rather than the actual values.

The above procedure can be thought of as a specialization of Rubin's Bayesian bootstrap to categorical data, where the categories are just the subintervals of the grid. It assumes a histogram model with the probabilities assigned to the subintervals as the parameters. So in this sense it is not really nonparametric since the grid is determined before the sample is chosen. The Banks procedure is similar in spirit except that he lets the subintervals of the grid be determined by the observed values in the sample. Although this is intuitively quite appealing, as we noted before there is no known Bayes or stepwise Bayes justification for a sample dependent grid.

In what follows we will be comparing these three methods for point and interval estimation of a population quantile. We will denote the Bayesian bootstrap by bbs, the smooth Bayesian bootstrap of Banks by sbbs, and the Bayesian bootstrap based on a grid by gbbs.

3 Estimating a quantile

In this section we will present some simulation results comparing various point and interval estimators of population quantiles. We begin by studying the usual sample quantile and an estimator based on Rubin's Bayesian bootstrap. Recall that given X=x we will take as a 'posterior distribution' for F a Dirichlet distribution with parameter vector v over the simplex $\Lambda(t)$

when T(x)=t and V(x)=x. This induces a 'posterior' for q=q(F). If the loss function is squared error the corresponding estimator is the mean of this distribution. This estimator can only be found approximately by simulating the 'posterior distribution' of q=q(F). For example we could take R independent observations from the appropriate Dirichlet distribution. For each observation, compute the qth quantile and then find the mean of these R computed quantiles. We will call this estimator the Bayesian bootstrap estimator of q=q(F) and denote it by bbsq. In what follows we will compare it to the usual naive estimator of q=q(F), i.e the qth quantile of the sample, denoted by smpq.

Since we are estimating population quantiles the most natural loss function is absolute error. For this loss function our estimator of q=q(F) is just the median of our 'posterior distribution'. When estimating the population median Ferguson(1973) shows that the median of our 'posterior distribution' is just the median of x, i.e. the sample median. For other quantiles this estimator cannot be found explicitly. In some cases it behaves quite similarly to the estimator smpq while in other instances it is out performed by both the bbsq and smpq estimators. For this reason we will only compare the estimators bbsq and smpq, but using absolute error as our loss function.

The two estimators were compared for six different populations. The populations were Beta(20,20), denoted by Beta_1 , Beta(.7,.8), denoted by Beta_2, Cauchy with location and scale parameters 0 and 1, the standard Exponential, the Gamma with shape parameter 20 and scale parameter 1, and the Lognormal with mean and standard deviation (of the log) 4.9 and .586. For each population we considered samples of size 11, 25 and 50. In each case we estimated the tenth, twenty-fifth, fiftieth, seventy-fifth and ninetieth quantiles. These quantiles are denoted by q10, q25, q50, q75 and q90. We also estimated the differences q75-q25 and q90-q10. In each case we observed 500 random samples from each population. Given a sample we computed bbsq by taking R=500 observations from the appropriate Dirichlet distribution. These simulations were done using S. When computing the quantile of a sample of size n, S linearly interpolates between order statistics of the sample, assuming the ith order statistic is the $(i-.5)/n$ quantile. The results of the simulations are given in the Table 1. Table 1 exhibits the ratio of the the average absolute error for the estimator smpq and the estimator bbsq and gives the percentage of times the estimator bbsq is closer to the true quantile than the estimator smpq.

place Table 1 about here

We see from the Table 1 that, except for the Cauchy distribution, the Bayesian bootstrap estimator performs better than the usual naive estimator in the vast majority of the cases. The usual estimator seems to do better for extreme quantiles in the heavy tail of the distribution especially for small sample sizes. The average amount of improvement for the bbsq over the smpq is somewhere between five and ten precent although it can reach as high as twenty precent. This suggests that for most nonparametric problems for which moments are assumed to exist one should use the Bayesian bootstrap rather the a sample quantile.

We will now present one possible explanation for this, perhaps, somewhat surprising fact. An holistic approach to nonparametric estimation suggests that one should first find a reasonable estimate of the entire population and then use this to find estimates of particular aspects of the population of interest. In some sense both the sample mean and sample median, as estimators of the population mean and median, seem to follow from this holistic approach. However the sample mean makes fuller use of the information in the sample than does the sample median. It is just this fact that makes the sample median more robust than the sample mean. It has long been recognized that when estimating a quantile the usual naive estimator can be improved upon if some additional information about the population is known. For example when estimating the median of a symmetric population a symmetric pair of order statistics can be preferred to the sample median. The difficulty for the general nonparametric problem when only vague prior information about the population is available is how to improve upon the usual naive estimator. What is surprising is that the 'posterior' leading to the Bayesian bootstrap seems to do just that in essentially an automatic manner.

Often one is interested in interval estimators as well as point estimators. In finite population sampling Meeden and Vardeman (1991) observed that the 'polya posterior', which is the Bayesian bootstrap adapted to finite population sampling, gave an interval estimator of the finite population median which was very close to a frequentist interval. For the nonparametric problem of estimating a median, Efron (1982) shows that asymptotically the Bayesian bootstrap intervals are equivalent to the usual nonparametric intervals based on the order statistic. Note that since this interval is based on two extreme order statistics it makes more use of the information in the sample than a

single sample quantile does.

We now wish to compare the three Bayesian bootstrap estimators bbs, sbbs and gbbs. We did this using the gamma distribution with shape parameter 20 and scale parameter 1. We also consider three different grids for gbbs. One grid just went from 0 to 60 with each step having length 1. For this distribution 8.96 is the .001 quantile while 36.7 is the .999 quantile. A second grid started with the values 0,2,4,5,6,7,8, and 8.5. It then divided (8.96,36.7) into 100 successive subintervals of equal probability. It then ended with the values 40, 45, 50, 55 and 60. The shortest subinterval in this grid was just slightly larger than .11. The third grid went from 0 to 8 in steps of 1, then from 8.5 to 30 in steps of .1 and from 30 to 60 in steps of 1. The simulation results for the three grids where very similar and hence we will present only the results for the third grid. So the choice of a grid does not seem to matter very much as long as there are no large subintervals which contain lots of probability.

To use the smooth Bayesian bootstrap one needs to assume that the support of the distribution is confined to some known interval. We took the lower bound to be 0. Although technically correct this is not a good idea because the subinterval defined by 0 and the minimum of the sample is too large and contains a region which is essentially assigned probability zero. This may not matter much when estimating the median but could have some affect when estimating smaller quantiles. For the upper end we assumed that the last interval was just (max,max+1) where max was the maximum value of the sample. Some comparisons are given in Table 2.

place Table 2 about here

Overall the preformance of the three 'Bayesian bootstrap' estimators are quite similar. For a small sample size the smooth Bayesian bootstrap seems to give more accurate coverage probabilities but as expected can be sensitive to the assumed bounds for the distribution. Now it is intuitively clear that Rubin's Bayesian bootstrap, for small sample sizes, will under estimate the amount of variation in the population. Hence Bank's smooth Bayesian bootstrap which spreads out the 'posterior probability' could be an improvement. However to call it a smooth Bayesian bootstrap seems to be a misnomer since there does not appear to be any Bayesian justification for the method. The modification of Rubin's Bayesian bootstrap when the sample space is

partitioned into a grid has the intuitive appealing property that the predictive distribution for a future observation is no longer concentrated on the observed values. However point and interval estimates based on it seem to differ little from those based on Rubin's Bayesian bootstrap. Moreover for moderate and larger sample sizes, based on the simulations given here, there seems to be little difference between the three of them.

Up until now most estimation studies have been concerned with interval estimation and obtaining estimates of variance. All though such problems are important in practice, the problem of point estimation is of interest as well. The main purpose of this note was to demonstrate that estimates of population quantiles based on Rubin's Bayesian bootstrap are preferred to the usual naive estimates. Moreover this method yields a sensible noninformative Bayesian approach to these problems. For a more informative Bayesian approach to such problems see Doss (1985). Yang (1985) presents another approach to the nonparametric estimation of quantiles. It is essentially a kernel type estimator. The optimal choice of the window width depends both on the sample size and the underlying distribution. He suggests a bootstrap technique to estimate the optimal width. Since this involves doing many different bootstraps for each sample it was not included in this study.

References

[1] Banks, D. L. (1988), Histospline smoothing the Bayesian bootstrap, *Biometrika*, 75, 673-684.

[2] Efron, B. (1982), *The Jackknife, the Bootstrap and other resampling plans*, Society for Industrial and Applied Mathematics, Philadelphia.

[3] Ferguson, T. S. (1973), A Bayesian analysis of some nonparametric problems, *Annals of Statistics* , 1, 209-230.

[4] Hill, B. M. (1968), Posterior distribution of precentiles: Bayes's theorem for sampling from a population, *Journal of the American Statistical Association* , 63, 677-691.

[5] Lane, D. A. and Sudderth W. D. (1978), Diffuse models for sampling and predictive inference, *Annals of Statistics* , 6, 1318-1336.

[6] Lo, A. Y. (1988), A Bayesian Bootstrap for a finite population, *Annals of Statistics* , 16, 1684-1695.

[7] Meeden, G. and Ghosh, M. (1983), Choosing between experiments: applications to fintie population sampling, *Annals of Statistics* , 11, 296-305.

[8] Meeden, G., Ghosh, M. and Vardeman, S. (1985), Some admissible nonparametric and related finite population sampling estimators, *Annals of Statistics* , 13, 811-817.

[9] Meeden, G. and Vardeman, S. (1991), A noninformative Bayesian approach to interval estimation in finite population sampling, to appear in *Journal of the American Statistical Association* .

[10] Rubin, D. B. (1981), The Bayesian bootstrap, *Annals of Statistics* 9, 130-134.

[11] Yang, S-S. (1985), A smooth nonparametric estimator of a quantile function, *Journal of the American Statistical Association* 80, 1004-1011.

Population	n	q10	q25	q50	q75	q90	q75 - q25	q90-q10
$Beta_1$	11	1.00	1.07	1.10	1.09	1.03	1.23	0.94
	25	1.18	1.10	1.08	1.12	1.13	1.10	0.94
	50	1.09	1.10	1.06	1.95	1.10	1.14	1.12
$Beta_2$	11	0.75	1.00	1.23	1.10	0.81	1.02	0.70
	25	0.94	1.03	1.16	1.12	0.93	1.05	0.84
	50	0.94	1.07	1.09	1.07	1.01	1.10	0.90
$Cauchy$	11	1.16	0.44	0.84	0.39	1.19	0.31	1.20
	25	0.40	0.66	1.01	0.82	0.41	0.62	0.34
	50	0.62	0.98	1.02	0.97	0.63	0.98	0.54
$Exponential$	11	0.80	0.95	1.06	1.09	1.04	1.15	1.04
	25	0.97	1.02	1.07	1.08	1.12	1.12	1.12
	50	1.01	1.04	1.07	1.06	1.08	1.07	1.08
$Gamma$	11	1.02	1.07	1.12	1.09	1.05	1.11	1.00
	25	1.08	1.08	1.11	1.07	1.14	1.14	1.14
	50	1.09	1.11	1.07	1.09	1.08	1.11	1.09
$Lognormal$	11	0.90	1.02	1.12	1.07	1.13	1.15	1.10
	25	1.09	1.06	1.06	1.09	1.13	1.14	1.14
	50	1.07	1.04	1.04	1.04	1.07	1.08	1.07

Table 1: The ratio of the average absolute error of the estimator smpq to the estimator bbsq for various sample sizes n.

n	Quantile	Estimator	Value	Absolute error	Length	Frequency of coverage
11	q25=16.83	bbs	17.21	1.14	6.31	0.916
		sbss	16.81	1.08	7.83	0.970
		gbss	17.21	1.14	6.26	0.916
25		bbs	16.96	0.77	4.30	0.944
		sbss	16.82	0.77	4.39	0.956
		gbss	16.96	0.77	4.31	0.948
11	q50=19.67	bbs	19.89	1.18	6.36	0.916
		sbss	19.89	1.18	6.68	0.936
		gbss	19.89	1.18	6.33	0.912
25		bbs	19.74	0.79	4.41	0.950
		sbss	19.75	0.80	4.39	0.960
		gbss	19.75	0.79	4.37	0.954
11	q75=22.81	bbs	22.68	1.37	7.92	0.904
		sbss	23.08	1.43	7.98	0.914
		gbss	22.68	1.36	7.81	0.900
25		bbs	22.83	0.95	5.46	0.966
		sbss	23.00	0.96	5.57	0.964
		gbss	22.83	0.95	5.41	0.962

Table 2: The average value and absolute error of the point estimator and average length and frequency of coverage for a 95% interval estimator for 500 random samples of size n from a Gamma distribution for three "Bayesian bootstrap" methods where in each case the estimates were computed from $R = 500$ simulated observations