COMPARING BINOMIAL BOOTSTRAP AND BAYESIAN ESTIMATION METHODS IN ASSESSING THE AGREEMENT BETWEEN CLASSIFIED IMAGES AND GROUND TRUTH DATA.

Bahman Shafii

William J. Price

COMPARING BINOMIAL BOOTSTRAP AND BAYESIAN ESTIMATION METHODS IN ASSESSING THE AGREEMENT BETWEEN CLASSIFIED IMAGES AND GROUND TRUTH DATA.

Bahman Shafii and William J. Price
Statistical Programs
P.O. Box 442337
University of Idaho
Moscow, Idaho 83844

ABSTRACT

The degree of agreement between classification and ground truth in remotely sensed data is often quantified with an error matrix and summarized using agreement measures such as Cohen's kappa. In the case of ground truth however, the kappa statistic can be shown to be a transformation of the marginal proportions commonly referred to as omissional and commissional error rates. A more meaningful statistical interpretation of remote sensing results and less ambiguous conclusions can be obtained via direct utilization of these measures. Several estimation techniques have been suggested for these marginal proportions. In this study, we will develop the exact binomial, bootstrap and Bayesian estimation methods for omissional and commissional errors. Emphasis will be placed on comparing the various estimation methods and their corresponding empirical distributions. Results are demonstrated with reference to a study designed to evaluate the detectability of yellow hawkweed and oxeye daisy using multispectral digital imagery in Northern Idaho.

Keywords: Agreement Measures, Omissional and Commissional Error rates, Remote Sensing.

I. INTRODUCTION

Accuracy in remote sensing is often assessed through the comparison of the classified points on an image (pixels) with their corresponding locations on the ground. The most common means of quantifying such comparisons is the error matrix (Card, 1982; Congalton, et al., 1983) which records the incidence of agreement and disagreement between classification and ground truth. The rows of the matrix denote the classified categories (i=1, 2, 3, ..., C) and the columns represent the reference or ground truth categories (j=1, 2, 3, ..., C):

Ground Truth

Classified	1	2	3	...	C	
1	x_{11}	x_{12}	x_{13}	...	x_{1c}	$N_{1.}$
2	x_{21}	x_{22}	x_{23}	...	x_{2c}	$N_{2.}$
3	x_{31}	x_{32}	x_{33}	...	x_{3c}	$N_{3.}$
...	...	...	...	...	...	...
C	x_{c1}	x_{c2}	x_{c3}	...	x_{cc}	$N_{C.}$
	$N_{.1}$	$N_{.2}$	$N_{.3}$	...	$N_{.C}$	N

where x_{ii} is the number of pixels correctly classified in category i, $N_{i.}$ and $N_{.i}$ are the corresponding marginal totals for classification and ground truth, respectively, and $N = \Sigma N_{i.} = \Sigma N_{.i}$.

Various measures have been suggested for assessing the degree of ground truth agreement

$$\hat{K}_i = (x_{ii}/N_{.i} - N_{i.}/N)/(1 - N_{i.}/N) \quad , \tag{1}$$

for each category. Some of the common measures include conditional kappa, a general index of agreement (Light, 1971) :
the omissional error rate, measuring the proportion of pixels incorrectly omitted from a classification:

$$\hat{O}_i = 1 - x_{ii}/N_{.i} \quad , \tag{2}$$

and the commissional error rate, measuring the proportion of pixels incorrectly committed to a classification (Aronoff, 1982):

$$\hat{C}_i = 1 - x_{ii}/N_{i.} \quad . \tag{3}$$

Kappa statistics have been suggested as a means of assessing the degree of agreement in remotely sensed data because they equally weight both omissional and commissional errors (Rosenfield and Fitzpatrick-Lins, 1986). However, remote sensing presents a unique situation for conditional kappa in which, for a given image classification, the marginal ground truth totals, $N_{.i}$, as well as classified totals, $N_{i.}$, are constant. Under these circumstances, (1) becomes a simple monotonic function of the omissional error rate (Wackerley, et al., 1978). Furthermore, although kappa treats misclassifications equally, in many cases it may be important to distinguish between the error types (Story and Congalton, 1986). Thus, it will be more advantageous to carry out

accuracy assessment based on the later two measures, namely omissional and commissional error rates. These measures also provide a better statistical interpretation of remote sensing results.

The objective of this study is to develop binomial, Bayesian, and bootstrap procedures for the estimation of omissional and commissional error rates. Comparison of various method via their corresponding empirical distributions are demonstrated with reference to a study concerned with detecting yellow hawkweed and oxeye daisy in Northern Idaho.

II. METHODS

Three estimation methods are considered, i) binomial, ii) Bayesian, and iii) bootstrap, which have certain components in common. These similarities along with a complete theoretical development for the aforementioned techniques are discussed below.

i) Binomial

If the area and location being imaged are held constant, the marginal totals for ground truth, $N_{.i}$, are fixed and the diagonal elements of the error matrix, x_{ii}, are distributed as binomial variates:

$$\Phi_i = \binom{N_{.i}}{x_{ii}} p_{ii}^{x_{ii}} (1-p_{ii})^{N_{.i}-x_{ii}} \tag{4}$$

where p_{ii} is the true proportion of correctly classified pixels. Hence, the omissional error rate (2) may be formed as a monotonic function of x_{ii} and the distribution of $\hat{O}_i$ maybe derived from Φ_i using the following transformation:

$$\Phi_i = p(x_{ii} = b) = p(1 - \frac{x_{ii}}{N_{.i}} = 1 - \frac{b}{N_{.i}}) = p(\hat{O}_i = b') \quad , \tag{5}$$

where b and b` are constant values. Further, if inferential results are limited to a specified classification, the $N_{i.}$ are also fixed and similar distributional developments can be made for the commissional error rate.

Statistical inference using the binomial distribution has typically been carried out using either a normal approximation or exact binomial confidence bounds (see, for example, Morisette and Khorram, 1998). An alternative approach is to numerically develop the full binomial distribution and hence, ascertain its associated probabilities, moments and percentiles. The later technique is adopted throughout this paper.

ii) Bayesian

The Bayesian perspective for $\hat{O}_i$ may be developed using (4) as a likelihood and assuming a prior distribution for p_{ii}. Using a constant non-informative prior (Shafii and Price, 1998), the posterior distribution for $p_{ii} \mid x_{ii}$ becomes:

$$\pi(p_{ii}|x_{ii}) \propto \pi(p_{ii})\cdot\Phi_i = A\cdot\Phi_i \quad . \tag{6}$$

Similar to the binomial case, the distribution for $\hat{O}_i$ can be generated from a transformation as shown in (5).

For the binomial method, inference on commissional error was dependent on limiting the scope of the problem to a specified classification algorithm. If this restriction is lifted, the commissional error rate becomes a function of binomial variates involving sums and ratios:

$$\hat{C}_i = 1 - \frac{x_{ii}}{N_{i.}} = 1 - \frac{x_{ii}}{x_{ii} + \sum x_{ij}} \quad . \tag{7}$$

where Σx_{ij} is the sum of the i^{th} row elements over j, $i \neq j$ and $x_{ii} > 0$. The x_{ij} are independent and distributed as:

$$\Phi_{ij} = \binom{N_j}{x_{ij}} p_{ij}^{x_{ij}} (1-p_{ij})^{N_j - x_{ij}} \tag{8}$$

Analytical derivation of a posterior distribution for (7) is troublesome. However, numerical derivation is possible using posterior distributions based on (4) and (8). To simplify the computations, an initial distribution for the inverse of $x_{ii} / N_{i.}$ is generated. The resulting values are subsequently re-inverted to obtain the final posterior distribution. This solution is restricted to $x_{ii} > 0$, encompassing all non-degenerate cases. Estimates, moments, and probability intervals are then derived using these posterior distributions.

iii) Bootstrap

Given the binomial distributions (4) and (8), a parametric bootstrap method (Efron and Tibshirani, 1993) may be used to generate the error matrix. Parameters p_{ii} and p_{ij} are replaced with their empirical estimates, $x_{ii} / N_{.i}$ and $x_{ij} / N_{.j}$, respectively. Bernoulli samples of size $N_{.i}$ or $N_{.j}$ are then drawn to fill the columns of the error matrix. Omissional and commissional error rates are then calculated from the bootstrapped error matrix. The resampling process is repeated a large number of times leading to empirical bootstrap distributions for each error rate. As with the other methods, the bootstrap distributions can be used to provide estimates, moments and confidence intervals.

All estimations were carried out using custom C codes and SAS (1991). Bootstrap computations were based on a large number of repetitions, i.e. 5000., as a means of accounting for the estimation of an entire error matrix at each iteration. Program codes are available from the authors at: http://www.uidaho.edu.ag/statprog/kansas01 .

III. DEMONSTRATION

The data used for the purpose of demonstration were reported in Lass and Callihan (1997), which consisted of remotely sensed images taken on June 10, June 21, and July 17, 1994 near Fernwood, Idaho. Originally, the data were classified into nine categories representing three levels of hawkweed, one level of oxeye daisy, three other vegetation types, and categories for soil and water. As for this demonstration, only the data obtained on the June 21 sampling date are considered. Also, the data are combined and reclassified into four categories: 1) hawkweed, 2) oxeye daisy, 3) other vegetation, and 4) non-vegetation categories, for which the presence or absence is recorded. Furthermore, to illustrate the techniques described above, the examples shown concentrate on hawkweed and oxeye daisy (weed) categories.

The error matrix representing the four categories as defined above is given in Table 1. The values within this matrix represent the number of correctly (on diagonal) or incorrectly (off diagonal) classified pixels. Column totals indicate the number of pixels sampled in a given category. For example, the hawkweed category shows 1254 correctly classified pixels out of 1511 possible pixels, indicating a producer's accuracy of 83% or an omissional error rate of 1 - 1254/1511 = 0.17. Oxeye daisy, on the other hand, has a user's accuracy of 47% or an omissional error rate of 1 - 69/147 = 0.53. This suggests a poor detection accuracy for oxeye daisy, in that more than half of the 'true' oxeye daisy pixels are omitted from the classified image. The category for other vegetation indicates a low omissional error rate, 12%, whereas the value of the omissional error rate for the non-vegetation category is zero.

The commissional error rate is often used as a post-classification measurement of accuracy. Based on the perspective of someone using the classified image, the error rates are computed relative to the row totals. In this example, hawkweed has a commissional error rate of 1 - 1254/1389 = 0.097 or equivalently, a user's accuracy of about 90%. However, the commissional error rate for oxeye daisy is 0.64 which represents a user's accuracy of about 36%. This relatively high value of commissional error indicates that more than 60% of pixels that had been classified as oxeye daisy were in fact committed to that class from other categories. Thus, the reliability of oxeye daisy classification from the user's perspective is low. The estimated values of the commissional error rates for the other vegetation and non-vegetation categories were15 and 8%, respectively.

The results of the three estimation methods are given in Table 2. Estimates of omissional and commissional error rates for the four categories along with the standard error and respective upper and lower confidence limits for each of the estimation method are provided. Estimated standard errors and corresponding confidence bounds on the omissional error rates are similar for all the estimation methods. This is to be expected as the bootstrap technique is based on a binomial model, and the binomial and Bayesian methods are identical given the constant prior on the later. However, the estimated standard errors and the corresponding bounds for commissional error rates show more disparity. The standard errors for the binomial method are relatively higher than those of Bayesian and bootstrap methods, resulting in wider confidence limits and a lower precision. For example, the estimated values of the commissional error rate for the hawkweed category is 0.0972. The binomial methods gives an estimated standard error of 0.00062, whereas the Bayesian and bootstrap methods give estimates of 0.00008 and 0.00017, respectively. This

represents a substantial reduction in variability for the Bayesian and bootstrap methods.

The above interpretations are evident from the distributions of omissional and commissional error rates generated from the three estimation methods (Figures 1 and 2). Similar to the results obtained from Table 2, the distributions for omissional error rates for both hawkweed and oxeye daisy are very similar across estimation methods (Figure 1). Each method gives a unimodal distribution with essentially the same shape, location and range having comparable expected values and variances. For the commissional error rate, the expected values of the three methods are basically the same, however, the Bayesian and bootstrap distributions indicate substantially smaller variances. Also, the distributions are multimodal with numerous spikes, whereas the binomial distribution is unimodal and smoother in appearance (Figure 2). The irregular nature of the Bayesian and bootstrap distributions is a consequence of computing commissional error as a function of sums and ratios. On the other hand, the binomial method gives more predictable results because it is derived from a smooth function which approximates the former methods. The binomial misses numerous potential values of commissional error that the Bayesian and bootstrap methods encompass. For both the bootstrap and Bayesian distributions, the uniformity (equal spacing) and symmetry of commissional error rates implied by the binomial model is not evident. In fact, the distributions are irregular and show patterns of sparse and concentrated values which closely reflect the underlying structure of commissional error rates.

In general, the bootstrap empirical distributions are similar to their Bayesian counterparts as demonstrated by the distributions of omissional and commissional error rates. This similarity becomes more apparent with increasing bootstrap sample sizes. As an example, Figure 3 illustrates changes in the standard deviation of the estimated commissional error rate for hawkweed with increasing bootstrap sample sizes. The estimated value of the standard deviation drops markedly from about 0.0033 to 0.0010 as the bootstrap sample size increases from 1000 to100,000, and it approaches the Bayesian estimate of 0.0079 as the bootstrap sample size is increased to1,000,000. This phenomenon is further emphasized in Figure 4, where the estimated bootstrap probability distributions of the commissional error rate for hawkweed are displayed at various bootstrap sample sizes (B). As bootstrap sample sizes increase, the probability distribution of commissional error rates show better definition and become smoother. At B = 1,000,000, the empirical bootstrap probability distribution resembles that of the Bayesian probability distribution. Thus for the bootstrap method, given a large number of samples, the empirical distribution will approach in limit to the Bayesian solution.

IV. CONCLUDING REMARKS

Three estimation methods were introduced to assess the agreement of classified images with ground truth thorough evaluation of omissional and commissional error rates. For the omissional error, the three estimation methods provide roughly the same estimates with equal precision for all classified categories. The binomial method is an attractive solution due to its computational simplicity. However, The Bayesian approach, by definition, can incorporate prior information into the estimation. This may not necessarily be the simple constant prior, and may involve a distribution derived from similar work or image. Since remote sensing applications often utilizes images taken sequentially in time, the Bayesian methodology may provide a means of

incorporating such previous information into the estimation procedure. This later aspect of analysis is currently under investigation. The commissional error rate presents a more complex situation. In the binomial case, a restrictive assumption on the domain of inference is necessary. Here, the Bayesian and bootstrap methods provide a better alternative by eliminating this restriction while utilizing the underlying distributions. The empirical bootstrap distribution approaches the Bayesian solution in the limit. Hence, the bootstrap method may provide a more attractive alternative, especially when prior information regarding the distribution of error rates is not available.

ACKNOWLEDGEMENTS

This work is published with the approval of the Idaho Agricultural Experiment Station as manuscript number 01-01.

REFERENCES

Aronoff, S. 1982. Classification accuracy: A user approach. Photogrammetric Engineering and Remote Sensing. 48: 1299-1307.

Card, D. H. 1982. Using map category marginal frequencies to improve estimates of thematic map accuracy. Photogrammetric Engineering and Remote Sensing. 49: 431-439.

Congalton, R. G., R. Oderwald, and R. A. Mead. 1983. Assessing landsat classification accuracy using discrete multivariate analysis statistical techniques. Photogrammetric Engineering and Remote Sensing. Vol. 49 No. 12, pp. 1671-1678.

Efron, B. and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman Hall, NY.

Lass, L. W., and R. H. Callihan. 1997. The effect of phenological stage on detectability of yellow hawkweed (Hieracium pratense) and oxeye daisy (Chrysthanemum leucanthemum) with remote multispectral digital imagery. Weed Tech. Vol. 11, pp. 248-256.

Light, R. J. 1971. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bull. 76:5 pp. 365-377.

Morisette, J. T. and Khorram, S. 1998. Exact binomial confidence interval for proportions. Photogrammetric Engineering and Remote Sensing. 64: 281-283.

Rosenfield, G. H. and K. Fitzpatric-Lins. 1986. A coefficient of agreement as a measure of thematic classification accuracy. Photogrammetric Engineering and Remote Sensing. Vol. 52 No. 2, pp. 223-227.

SAS Institute Inc. 1991. SAS Language: Reference, Version 6, First Edition. SAS Institute Inc., Cary, NC, 1042 pp.

Shafii, B, and W. J. Price. 1998. A Bayesian solution for assessing variability of agreement measures in remote sensing imagery. *In* proceedings of the American Statistical Association, Section on Bayesian Statistical Science. pp. 22-27.

Story, M. and Congalton, R. G. 1986. Accuracy assessment: A user's perspective. Photogrammetric Engineering and Remote Sensing. 52: 397-399.

Wackerly, D. D., J. T. McClave, and P. V. Rao. 1978. Measuring nominal scale agreement between a judge and a known standard. Pyschometrika. 43: 213-223.

Table 1. The error matrix representing the number of classified and ground truth pixels for each of the four categories (1 = Hawkweed, 2 = Oxeye daisy, 3 = Other Vegetation, and 4 = Non-vegetation).

		Ground Truth				
		1	*2*	*3*	*4*	Total
Classified	*1*	1254	39	96	0	*1389*
	2	84	69	38	0	*191*
	3	157	39	1113	0	*1309*
	4	16	0	20	398	*434*
	Total	*1511*	*147*	*1267*	*398*	*3323*

Table 2. Estimates of omissional ($\hat{O}_i$) and commissional ($\hat{C}_i$) error rates for the four categories along with the binomial, Bayesian, and bootstrap estimates of standard errors and the respective 95% upper and lower confidence bounds.

		Estimation Method									
			Binomial			Bayesian			Bootstrap		
Category		Estimate	Std Err	Lower	Upper	Std Err	Lower	Upper	Std Err	Lower	Upper
Hawkweed	$\hat{O}_i$	0.1701	0.00065	0.1522	0.1899	0.00065	0.1522	0.1899	0.00098	0.1542	0.1860
Oxeye		0.5306	0.00445	0.4490	0.6122	0.00445	0.4490	0.6122	0.00614	0.4625	0.5986
Other Veg.		0.1215	0.00069	0.1050	0.1405	0.00069	0.1050	0.1405	0.00104	0.1066	0.1365
Non-Veg.		0.0000	0.00044	0.0000	0.0075	0.00044	0.0000	0.0075	0.00016	0.0000	0.0000
Hawkweed	$\hat{C}_i$	0.0972	0.00062	0.0828	0.1138	0.00008	0.0848	0.1127	0.00017	0.0855	0.1092
Oxeye		0.6387	0.00356	0.5707	0.7016	0.00049	0.5862	0.6973	0.00088	0.5902	0.6875
Other Veg.		0.1497	0.00071	0.1314	0.1704	0.00010	0.1344	0.1679	0.00021	0.1359	0.1632
Non-Veg.		0.0829	0.00143	0.0599	0.1129	0.00056	0.0637	0.1154	0.00152	0.0829	0.1036

Figure 1. Omissional error rate probability distributions based on the binomial, Bayesian, and bootstrap estimation methods for hawkweed (a) and oxeye daisy (b).

a)

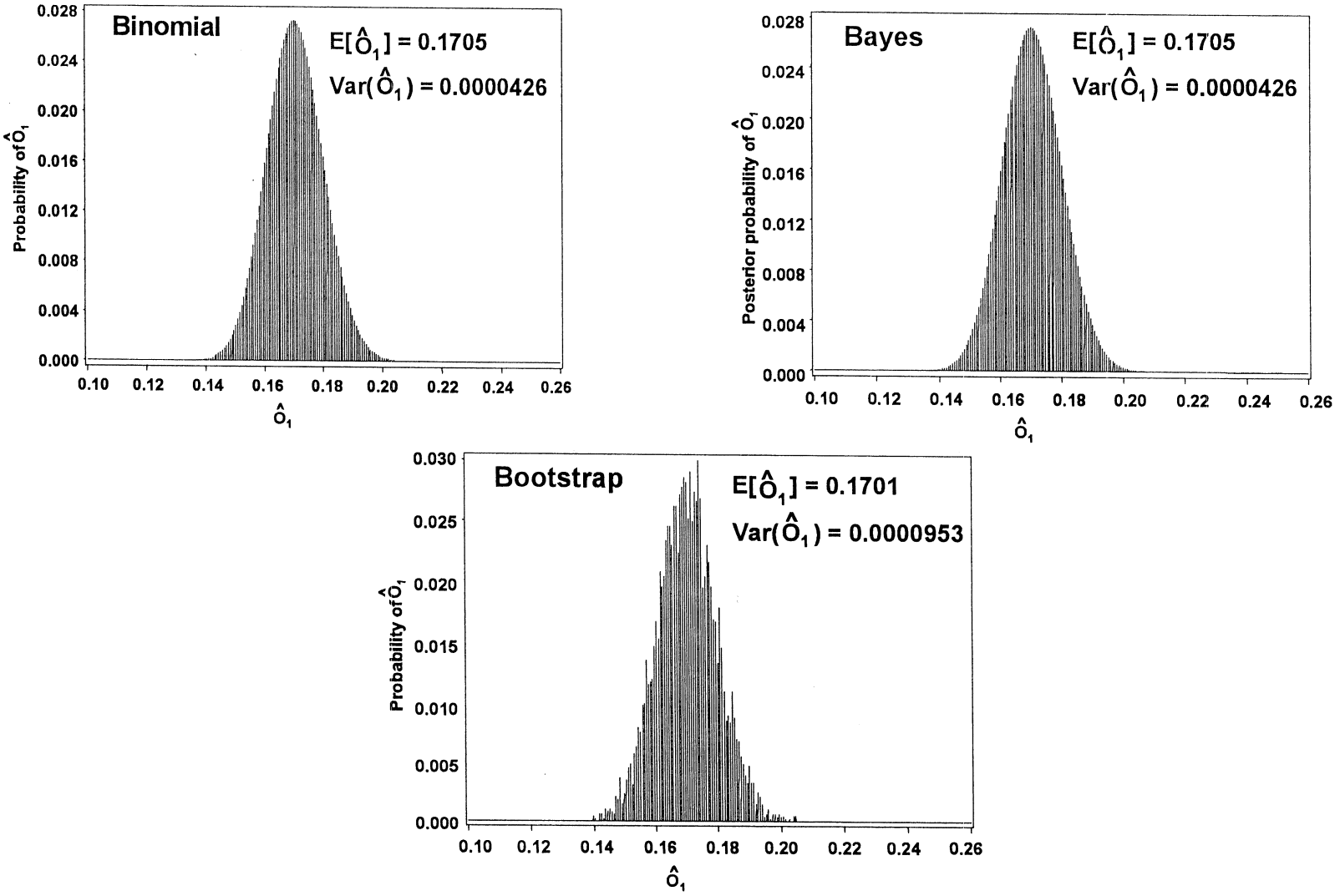
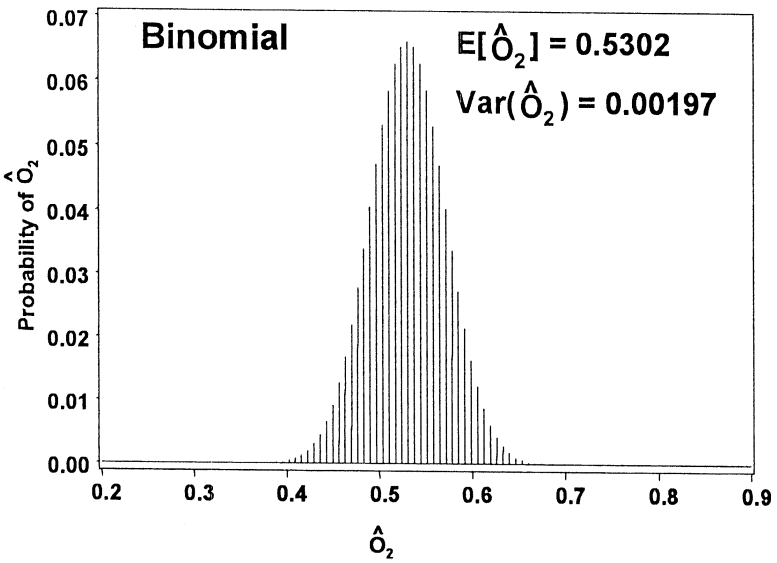
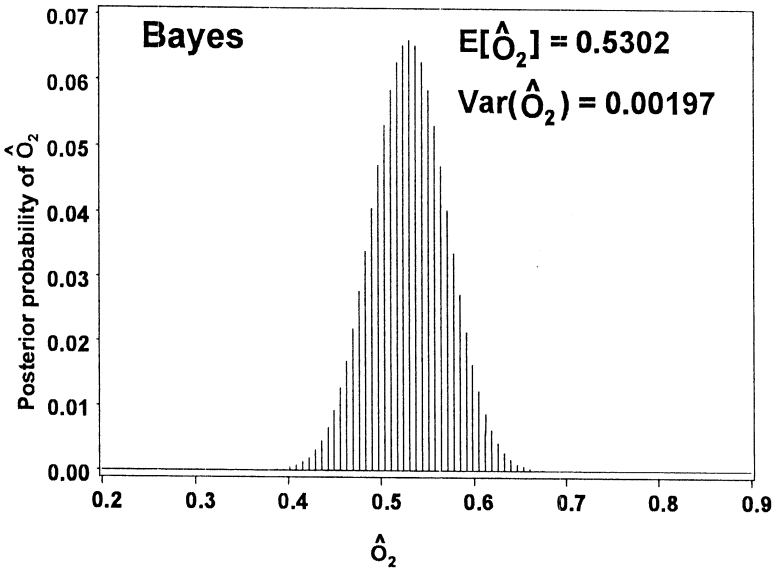


Figure1 (cont).

b)





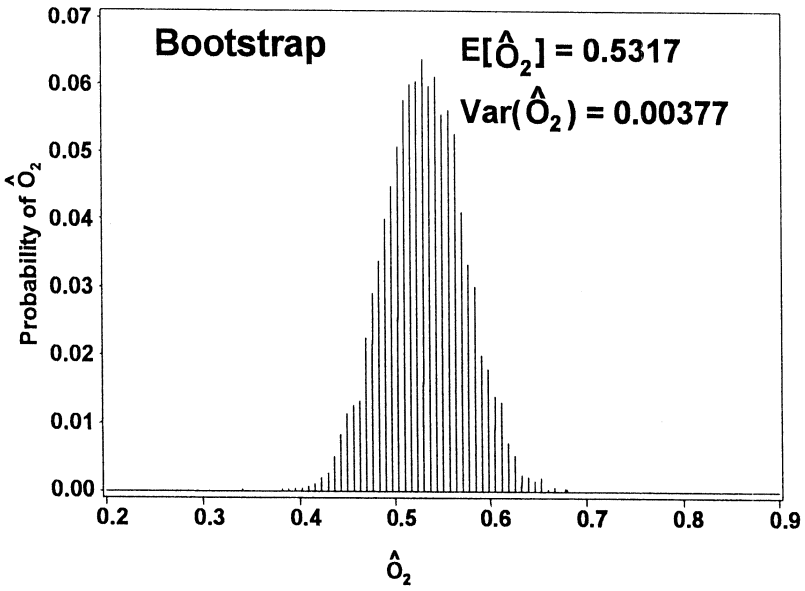


Figure 2. Commissional error rate probability distributions based on the binomial, Bayesian, and bootstrap estimation methods for hawkweed (a) and oxeye daisy (b).

a)

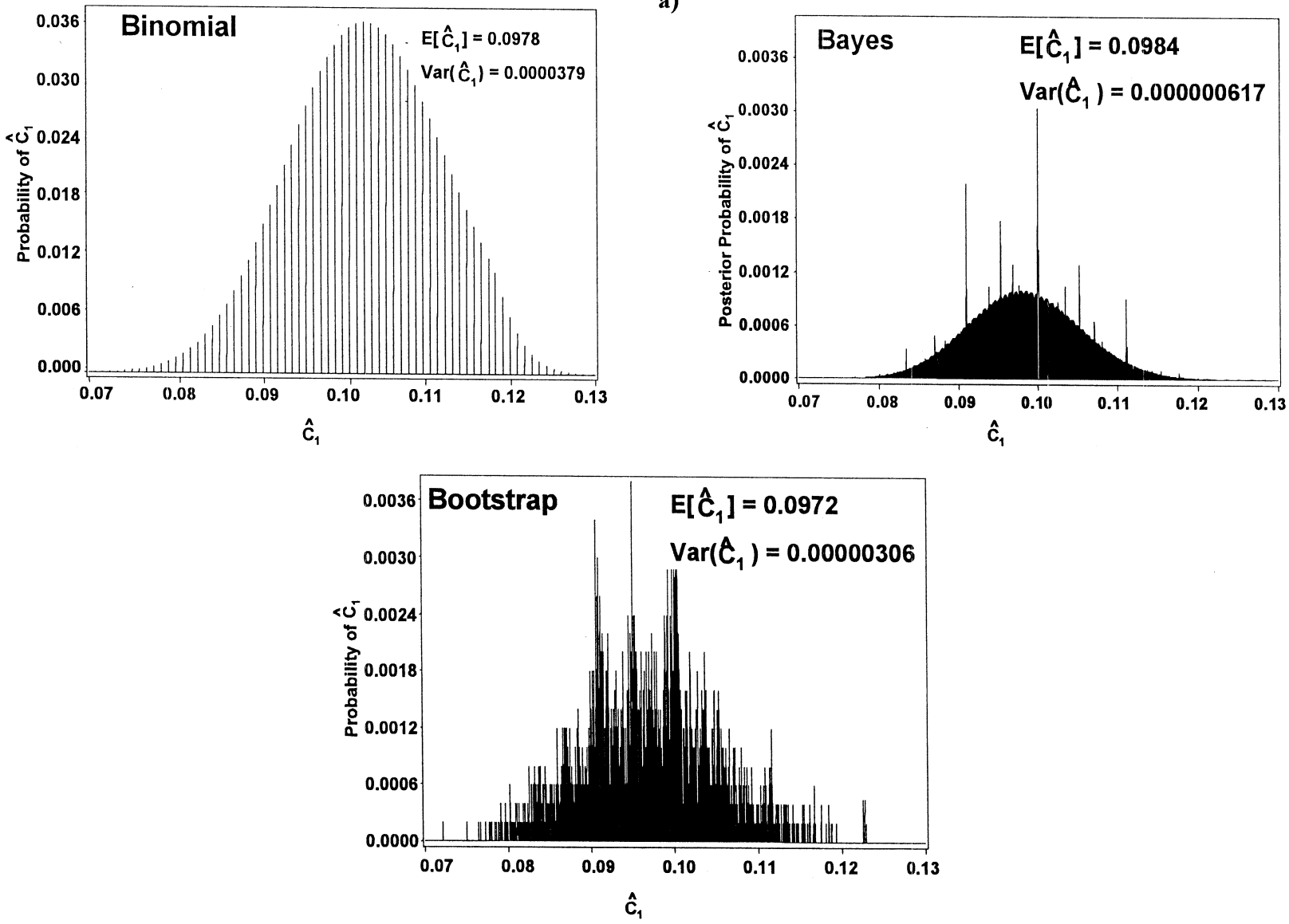
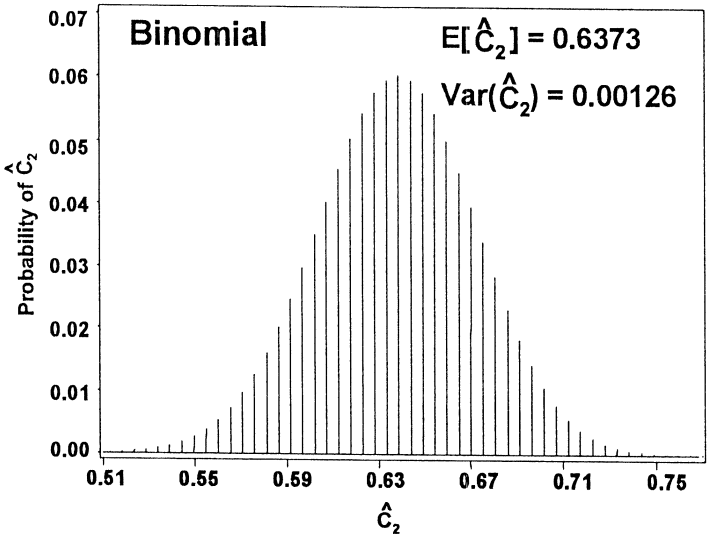
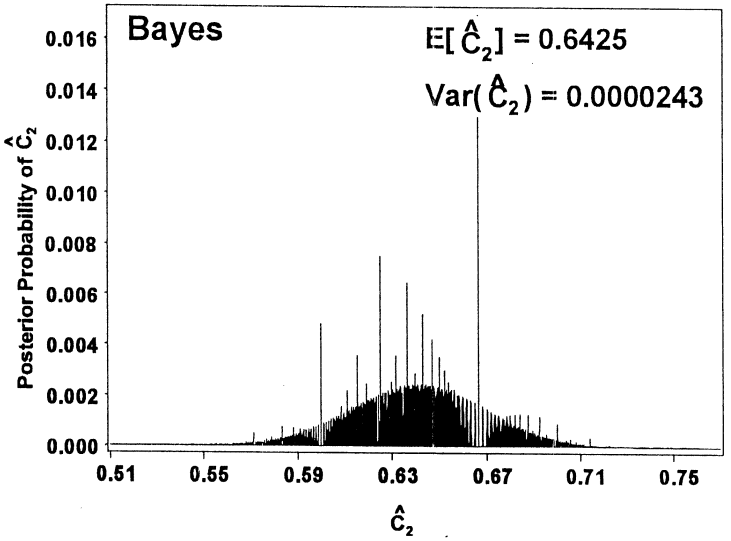


Figure 2 (cont).

b)





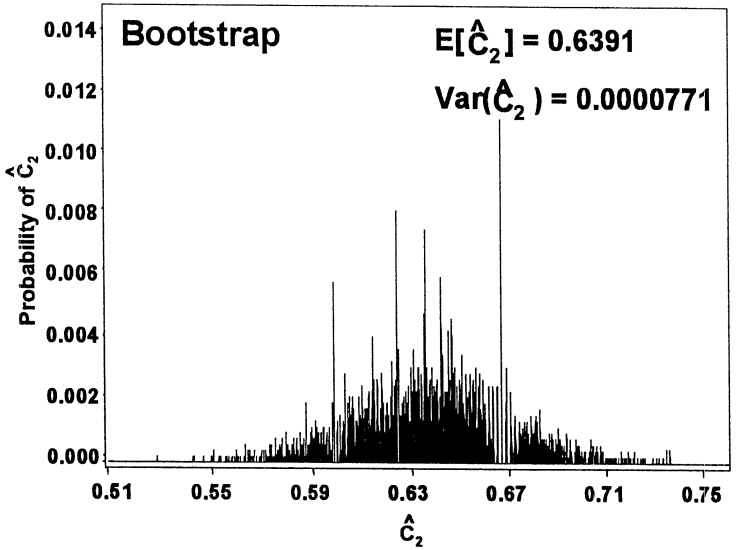


Figure 3. Standard deviation of the commissional error rate for hawkweed based on the bootstrap estimation method plotted against increasing sample sizes. The dashed line represents the corresponding standard deviation for the Bayesian estimate.

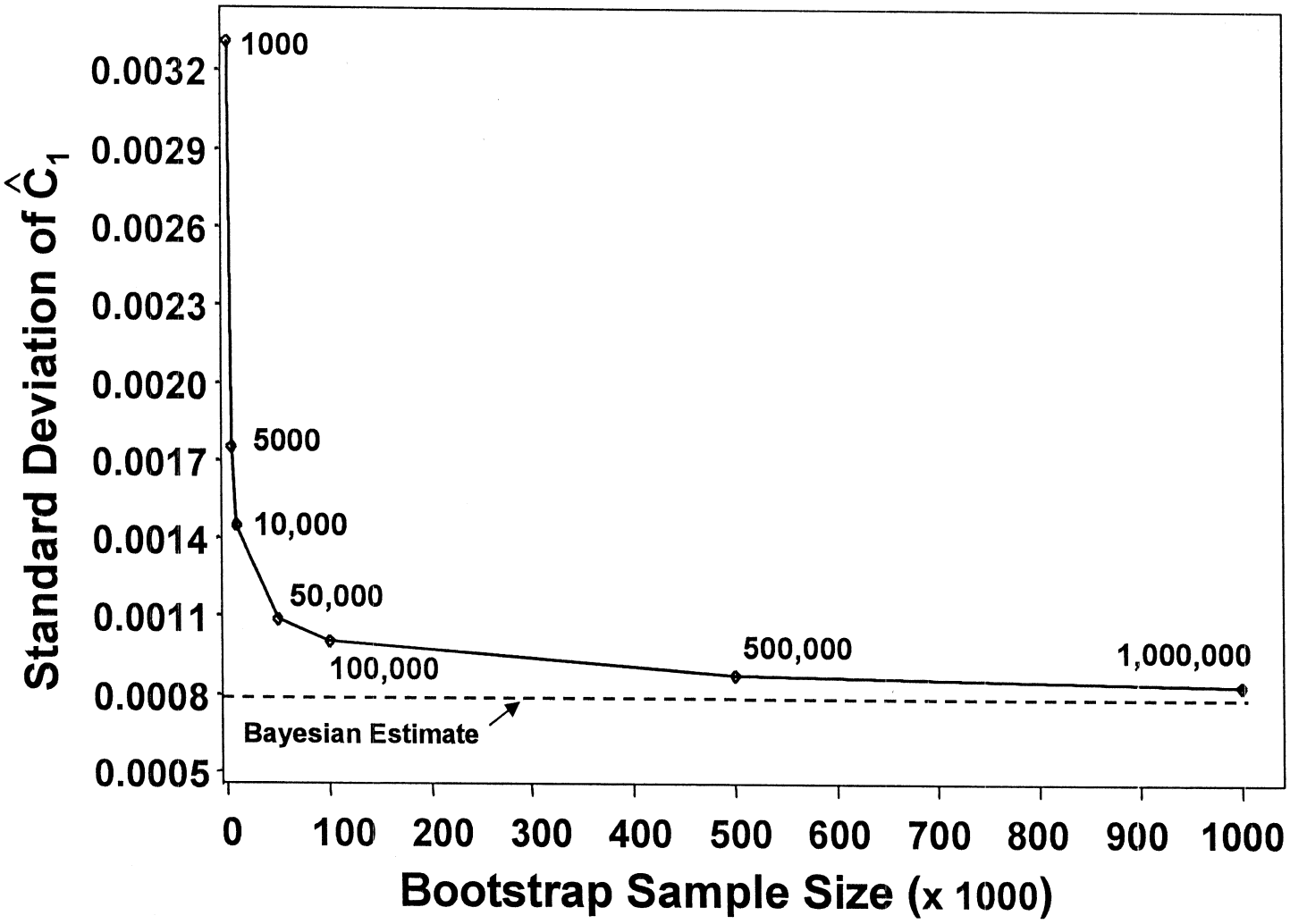


Figure 4. Estimated bootstrap probability distributions of the commissional error rate for hawkweed, p($\hat{C}_1$), at various bootstrap sample sizes (B) compared with the corresponding Bayesian probability distribution.