IMPROVED ESTIMATES FROM SAMPLE SURVEYS WITH EMPIRICAL BAYES METHODS

Douglas H. Johnson
U. S. Fish and Wildlife Service
Northern Prairie Wildlife Research Center

ABSTRACT

Surveys are often conducted at regular intervals, and information from previous surveys can be used to improve current estimates by redesigning the survey. Empirical Bayes (EB) methodology offers another approach to using prior information profitably, by modifying the analysis.

This paper illustrates how EB methods can improve estimates of wild waterfowl populations, which are counted in sample surveys each year. Standard EB estimators, based on the average count in previous years, and EB estimators based on regression-like priors are developed and compared with usual estimators.

INTRODUCTION

Sample surveys are often conducted at regular intervals to monitor various features of a population. Information from one survey can be used effectively to improve the next, usually by modifying the design. Examples include revising the stratification and optimal allocation of the sample units.

Such prior information can be employed in the analysis of results from a survey, as well as in its design. It is particularly useful when variation from one survey period to the next is modest, sample estimates are imprecise, and cost or other considerations prohibit ready increases in sample size. This paper will illustrate by example how information from previous surveys can be incorporated into estimators for the current survey. The methodology will be empirical Bayes, and the application will be to counts of breeding waterfowl, which are conducted annually. Although EB methods have been used in a variety of applications, some of which were reviewed by Morris (1983), they have seen little use in sample surveys. Fay and Herriot (1979) employed the related James-Stein estimation procedures for determining income for small places, and Marks and Woodruff (1979) used EB techniques to determine optimal sample allocation.

THE SURVEY

The U. S. Fish and Wildlife Service (FWS), in cooperation with the Canadian Wildlife Service and several state and provincial wildlife agencies, annually collects certain kinds of information about the status of important species of waterfowl. This information is gathered in time for analysis and promulgation of regulations for each fall hunting season. An important component of this process is the May population survey, in which breeding waterfowl are counted in sample transects throughout the major breeding areas of North America (Fig. 1). In addition to waterfowl, the number of wet ponds in the transect is tallied.

The survey involves both aerial counts and on-ground calibrations. Details are not included here; see Bowden (1973) or Martin et al. (1979) for further information. The important considerations for the purpose of this paper are that the survey is conducted regularly (each May), it is a fairly expensive procedure, and estimators of waterfowl numbers frequently have rather large variances. Because of the costs, improved accuracy through larger samples is not a ready solution. The purpose of the study, which is described in part here, is to develop improved estimators without increased sample sizes.

For this report, only four of the 49 strata (Fig. 1) are examined, and further these are collapsed. The strata are 28 and 29 (southeastern Alberta) and 30 and 31 (central Saskatchewan). These are combined to facilitate the subsampling experiment to be described shortly. This report considers only two of the waterfowl species, the mallard and the canvasback ducks. Both species are highly regarded by waterfowl hunters and have recently declined in number, so they command especial interest.

METHODS

Five estimators are evaluated. Two are usual sample estimators, one based on the average density of ducks in each transect, the other a ratio estimator involving the area of the transect. Two empirical Bayes (EB) estimators are constructed, one based on a prior involving the mean density of ducks in the stratum during previous years, the other based on a regression estimate involving the number of wet ponds. In addition, a ratio estimator involving ponds is developed, under the assumption that the number of wet ponds in the entire stratum could be known exactly (this is not currently feasible, but may become so with remote sensing information).

Data for the 1967-76 period are used to develop the necessary prior knowledge. Estimators are then applied to the data for each year of the 1977-81 evaluation period.

The estimators are evaluated by a subsampling experiment. Stratum 28/29 contains seven transects; stratum 30/31 contains nine. The average density of a species from all transects in a stratum is treated as the true density, against which the estimators are compared. Estimators are based on a random sample of three transects drawn without replacement each year from the seven or nine available. New samples are drawn each year.

The main criterion for evaluation is the root mean square error, obtained by squaring the difference between an estimate and the true value, averaging across all years in the evaluation period, then taking its square root. Secondary criteria are the average standard error and the coverage (the percentage of confidence intervals that contain the true value).

THE ESTIMATORS

For a particular species and stratum, let

N = the number of transects in the stratum (7 in stratum 28/29, 9 in stratum 30/31)
X_{ij} = the count of birds in transect i in year j
A_i = the area of transect i
$Y_{ij} = X_{ij}/A_i$ = density of birds in transect i in year j
W_{ij} = the number of wet ponds in transect i in year j.

Define the sums

$$X._j = \Sigma_i X_{ij}$$
$$A. = \Sigma_i A_i$$
$$Y._j = \Sigma_i Y_{ij}$$
$$W._j = \Sigma_i W_{ij}.$$

Then the average density, $\theta_j = X._j/A.$, is to be estimated. For the n = 3 transects drawn randomly in year j, let x_{ij}, a_{ij}, y_{ij}, w_{ij}, $x._j$, $a._j$, $y._j$, and $w._j$ be defined analogously. Let $\overline{x}._j = x._j/n$, etc. be the corresponding means; $s_{xj}^2 = \Sigma_i(x_{ij} - \overline{x}._j)^2/(n - 1)$, etc. be the variances; and s_{xaj} and s_{xwj} be the covariances. These values are observed.

The average density estimator.--This estimates the average density for the stratum by the average of the densities in the sample transects:

$$\hat{\theta}_j^{(1)} = \overline{y}._j.$$

The estimated variance of this estimator can be obtained from

$$V_j^{(1)} = (1 - n/N)s_{yj}^2/n,$$

where (1 - n/N) is the finite population correction.

The ratio estimator.--This estimator uses the fact that the transects are ordinarily of different areas. It is

$$\hat{\theta}_j^{(2)} = x._j/a._j.$$

Its variance can be estimated in the usual way (Cochran 1977):

$$V_j^{(2)} = (1-n/N)\{s_{xj}^2+(\hat{\theta}_j^{(2)})^2 s_{aj}^2 - 2\hat{\theta}_j^{(2)} s_{xaj}\}/(na._j^2).$$

Empirical Bayes estimator based on previous counts.--This is the customary EB estimator, using counts from the previous K = 10 years. Let z_j be the average estimated density of the species in that stratum during the previous 10 years:

$$z_j = \Sigma_{k=j-K}^{j-1} \overline{y}._k/K.$$

The EB estimator is formed by taking a weighted average of this prior value, z_j, and the sample estimate, $\overline{y}._j$:

$$\hat{\theta}_j^{(3)} = \{B_j z_j + (1-B_j)\overline{y}._j\}.$$

The weights B_j and $(1 - B_j)$ involve the relative variances of z_j and $\overline{y}._j$. Morris (1983) proposed calculating B_j as

$$B_j = (K-r-2)\overline{V}_j^{(1)}/\{(K-r)(\overline{V}_j^{(1)}+D_j)\},$$

where r is the number of parameters estimated (r = 1 here),

$$D_j = S_j/(K - 1) - \overline{V}_j^{(1)}$$

is an estimate of the among-years variance,

$$\overline{V}_j^{(1)} = \Sigma_{k=j-K}^{j-1} V_k^{(1)}/K$$

is a pooled estimate of the within-years variance, and

$$S_j = \Sigma_{k=j-K}^{j-1} (\overline{y}._k - z_j)^2$$

is the sum of squares about the mean of the previous K observations. If D_j is negative, it is replaced by zero. Morris (1983) also suggested an estimated standard error, which is used but not described here.

Empirical Bayes estimator based on pond regression.--This estimator is of the same form as the previous one, except that the prior value is obtained from a regression equation relating bird densities to the density of ponds with water $(w._j/a._j)$. Assume that a linear model is appropriate:

$$\overline{y}._j = \alpha + \beta(w._j/a._j) + e_j,$$

where e_j is an error term. Least squares estimates of the coefficients α and β based on the previous K = 10 years of data will be used. Then the EB estimator is

$$\hat{\theta}_j^{(4)} = \{C_j(\hat{\alpha} + \hat{\beta}w._j/a._j) + (1-C_j)\bar{y}._j\}.$$

Again the weights C_j and $(1 - C_j)$ depend on the relative variances of the two components. The same procedure as above is followed, except that

$$C_j = (K-r-2)\bar{V}_j^{(1)}/\{(K-r)(\bar{V}_j^{(1)}+D_j)\},$$

$$S_j = \Sigma_{k=j-K}^{j-1} (\bar{y}._k - \hat{\alpha} - \hat{\beta}w._j/a._j)^2,$$

and

$$D_j = S_j/(K - 1) - \bar{V}_j^{(1)}.$$

Ratio estimator based on ponds.--If we assume that the count of wet ponds for the entire stratum is known, and equal to say $W._j$, then a ratio estimator of the form

$$\hat{\theta}_j^{(5)} = (x._j/w._j)(W._j/A.)$$

is appropriate. Its estimated variance is

$$V_j^{(5)} = (1-n/N)\{s_{xj}^2 + (x._j/w._j)^2 s_{wj}^2 - 2(x._j/w._j)s_{xwj}\}(W._j/A.)^2.$$

RESULTS

Table 1 displays the data values obtained by randomly subsampling in each year and stratum. Sample averages are compared with true values based on the entire set of transects in Table 2. Averages shown for mallards, canvasbacks, and ponds are ratios of total numbers to total area.

Calculation of the various estimators will be illustrated with the mallard data for stratum 28/29 in 1977. The average density estimate, $\hat{\theta}^{(1)}$, is simply the average of the densities along each transect:

$$(187.3/45 + 196.2/45 + 160.5/45)/3$$

$$= 4.03 \text{ (Tables 1 and 2).}$$

Its variance is the usual variance of a mean:

$$(1-3/7)\{(4.36-4.03)^2 + (4.16-4.03)^2 + (3.57-4.03)^2\}/\{(3-1)3\}$$

$$= (0.18)^2.$$

The ratio estimate, $\hat{\theta}^{(2)}$, is

$$(187.3+196.2+160.5)/(45+45+45) = 4.03.$$

Its variance is

$$(1-3/7)\{345.3233 + 4.03^2(0) - 2\ (4.03)(0)\}/\{(3)(45^2)\}$$

$$= (0.18)^2.$$

These values coincide with those for the average density estimator in this situation because all transects had the same area.

The EB estimate based on the previous mean, $\hat{\theta}^{(3)}$, uses the fact that for the K = 10 years prior to 1977 the average estimated density of mallards in stratum 28/29 was 13.72, with a standard deviation of 10.60. (Notice that the mallard density was appreciably lower in 1977-79 than during earlier years.) The EB estimate is a weighted average of the prior mean, 13.72, weighted by 0.163, and the average sample density in 1977, 4.03, weighted by 0.837:

$$13.72(0.163) + 4.03(0.837) = 5.61.$$

The standard error, the calculation of which is not shown here (see Morris 1983), is $(1.31)^2$. This value, which is based on the assumption that the within-year variance is constant, is noticeably larger than the others obtained above. This disparity is because the observed variance in 1977 was smaller than in most years.

The EB estimator based on the pond regression, $\hat{\theta}^{(4)}$, uses the relation developed for 1967-76 data:

$$y = 9.74 + 0.637(\text{pond density}).$$

With an estimated pond density in 1977 of 2.53, this equation gives a prior estimate of mallard density of 11.35. Its estimated standard error is 9.66. The EB estimator becomes a weighted average of this prior estimate and the average density:

$$11.35(0.168) + 4.03(0.832) = 5.29,$$

with a variance of $(1.29)^2$.

The ratio estimate based on pond numbers, $\hat{\theta}^{(5)}$, involves the ratio of mallards to wet ponds in the sample transects (544/342 = 1.59) and the assumed known ratio of wet ponds to area in all transects (860/306 = 2.81). Thus the estimate is (1.59)(2.81) = 4.47. Its estimated variance is

$$(1 - 3/7)\{345.3233 + (1.59)^2(36) - 2(1.59)(26.7)\}(2.81)^2$$

$$= (0.20)^2.$$

Resulting estimates and standard errors for both species, both strata, and all years are given in Table 3. The average density and ratio estimates, $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$, are similar, a result due largely to the modest variation in the areas of the transects. Estimates of standard errors are closer than might be expected from Cochran's (1977) rule of thumb that n should exceed 30 for the formula for the standard error of a ratio estimator to apply.

Compared with $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$, the EB estimates, $\hat{\theta}^{(3)}$ and $\hat{\theta}^{(4)}$, tend to be shifted toward the prior average density. The shift was greater for stratum 30/31 than for stratum 28/29, reflecting greater weights given to prior estimates; average values of B_i and C_i were respectively 0.094 and 0.086 in stratum 28/29 and 0.226 and 0.228 in stratum 30/31. Standard errors tended to be larger for the EB estimates than for the usual ones. This results in part from the EB assumption that within-years variance is constant, and is estimated from the long-term average, which was larger than the values for most of the individual years of the evaluation period.

Table 4 provides the square root of the mean squared error (RMSE), average standard error, and coverage probabilities for the five estimators. RMSE is a succinct summary of the accuracy of an estimator. On this basis, both EB estimators and the pond ratio estimator outperform $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ on the mallard data in stratum 28/29 and especially in stratum 30/31. For canvasbacks, the pond ratio estimator was somewhat better than the other four in stratum 28/29, but the EB estimators were best in stratum 30/31.

Average standard errors, together with coverage probabilities, indicate the adequacy of estimates of precision for the estimators. For the mallard, average standard errors tended to be smallest for $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$, largest for the EB estimators, and intermediate for $\hat{\theta}^{(5)}$ (Table 4). Among canvasbacks, average standard errors were similar for all estimators.

Table 4 suggests that the estimated standard errors for the non-EB estimators were optimistically small in this modest evaluation. Coverage probabilities for the 20 species-stratum-year combinations averaged 65 percent for $\hat{\theta}^{(1)}$, 60 percent for $\hat{\theta}^{(2)}$, and 70 percent for $\hat{\theta}^{(5)}$. Each EB estimator produced confidence intervals covering true values in 90 percent of the cases.

DISCUSSION

This report is intended to demonstrate that prior information can be usefully employed in sample surveys, to illustrate two methods for incorporating it (the empirical Bayes and ratio estimation with known pond counts), and to exemplify such applications. A more comprehensive assessment of the EB approach is in progress, involving ten species of ducks and all 49 strata. Most published evaluations of EB procedures have been based on theoretical considerations or on contrived data sets. There is a need to assess the performance of EB estimators on real data sets, such as those discussed here.

From the foregoing analysis, results of which are typical of those obtained to date, several conclusions seem warranted. First, the EB estimators have promise. Although they sacrifice unbiasedness, they gain a reduction in mean square error that is often substantial; RMSE averaged 30 percent less for $\hat{\theta}^{(3)}$ than for $\hat{\theta}^{(2)}$, for example. Their estimated standard errors, which may be either larger or smaller than those of usual estimators, seem conservative and provide better coverage probabilities.

The ratio estimator involving the total wetland count also appears to be worthy of further consideration, particularly if those total counts become readily available. Estimators of this form could probably also be improved through EB methodology.

ACKNOWLEDGMENTS

This report benefitted from review by S. G. Machado and T. L. Shaffer.

REFERENCES

BOWDEN, D.C. (1973), "Review and Evaluation of May Waterfowl Breeding Ground Survey," Unpublished report to U.S. Fish and Wildlife Service.

COCHRAN, W.G. (1977), Sampling Techniques (3rd ed.), New York, New York: John Wiley & Sons.

FAY, R.E., III and HERRIOT, R.A. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," Journal of the American Statistical Association, 74, 269-277.

MARTIN, F.W., POSPAHALA, R.S., and NICHOLS, J.D. (1979), "Assessment and Population Management of North American Migratory Birds," Pages 187-239 in CAIRNS, J., JR., PATIL, G. P., and WATERS, W.E. (eds.), Environmental Biomonitoring, Assessment, Prediction, and Management - Certain Case Studies and Related Quantitative Issues, Statistical Ecology vol. 11, Fairland, Maryland: International Co-operative Publishing House.

MARKS, H.M. and WOODRUFF, S.M. (1979), "Allocation Using Empirical Bayes Estimators," American Statistical Association Proceedings of Section on Survey Research Methods, 309-313.

MORRIS, C.N. (1983), "Parametric Empirical Bayes Inference: Theory and Applications," Journal of the American Statistical Association, 78, 47-55.

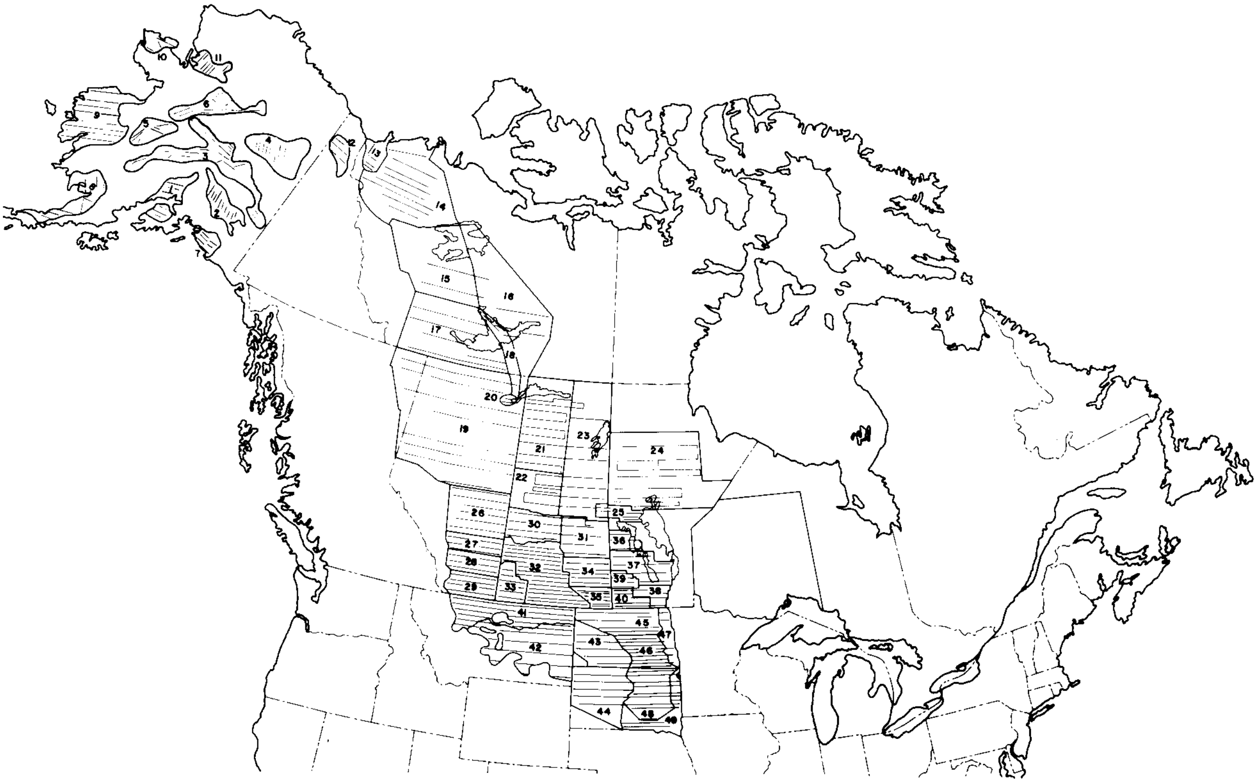

Figure 1. Strata and transects used in waterfowl surveys.

Table 1. Values obtained from random subsampling of n=3 transects in each stratum each year; Area is in square miles, Mal.=Mallards, Can.=Canvasbacks.[a]

	Sample	Stratum 28/29				Stratum 30/31			
Year	unit	Area	Mal.	Can.	Ponds	Area	Mal.	Can.	Ponds
1977	1	45.0	187.3	5.2	108	40.5	1851.5	255.3	426
	2	45.0	196.2	25.8	120	27.0	1446.0	234.6	488
	3	45.0	160.5	20.7	114	22.5	928.2	46.0	310
1978	1	45.0	424.4	16.6	278	40.5	578.1	41.0	302
	2	45.0	300.9	16.6	394	22.5	414.5	20.5	536
	3	40.5	189.0	16.6	172	27.0	343.6	0.0	184
1979	1	45.0	281.5	0.0	234	40.5	682.0	70.5	1120
	2	45.0	308.6	24.2	186	36.0	868.9	119.0	1114
	3	40.5	434.6	9.7	320	22.5	548.5	57.3	956
1980	1	45.0	490.3	17.0	276	40.5	976.6	288.3	342
	2	45.0	584.2	42.5	396	31.5	532.4	95.3	226
	3	45.0	314.4	8.5	178	22.5	636.8	30.1	328
1981	1	45.0	414.4	12.1	254	40.5	423.3	27.0	274
	2	45.0	430.9	0.0	454	36.0	748.4	54.0	232
	3	40.5	497.2	3.0	208	22.5	468.7	22.5	214

---

[a] Counts of ducks may be non-integer because a visibility adjustment has been applied.

Table 2. True values (from all transects) and sample values (based on random subsample of n=3 transects) of selected parameters; values are densities per square mile.

	Stratum 28/29						Stratum 30/31					
	Mallards		Canvasbacks		Ponds		Mallards		Canvasbacks		Ponds	
Year	True	Sample	True	Sample	True	Sample	True	Sample	True	Sample	True	Sample
1977	5.18	4.03	0.53	0.38	2.81	2.53	35.42	46.95	3.76	5.95	9.27	13.60
1978	6.83	7.01	0.40	0.38	7.36	6.47	16.40	14.85	1.60	0.68	13.08	11.36
1979	9.10	7.85	0.52	0.26	5.90	5.67	20.50	21.21	3.43	2.49	26.56	32.22
1980	8.98	10.29	0.24	0.50	5.59	6.30	22.25	22.71	3.20	4.38	10.27	9.48
1981	8.56	10.29	0.10	0.12	5.22	7.02	18.88	16.57	1.76	1.05	7.97	7.27

Table 3. Comparison of true values and various estimates of duck density, by species, stratum, and year; estimated standard errors are in parentheses.

Year (j)	θ_j	$\hat{\theta}_j^{(1)}$	$\hat{\theta}_j^{(2)}$	$\hat{\theta}_j^{(3)}$	$\hat{\theta}_j^{(4)}$	$\hat{\theta}_j^{(5)}$
			Mallard, Stratum 28/29			
1977	5.18	4.03 (0.18)	4.03 (0.18)	5.61 (1.72)	5.29 (1.67)	4.47 (0.20)
1978	6.83	6.93 (1.04)	7.01 (1.03)	7.58 (1.45)	7.56 (1.48)	7.97 (1.42)
1979	9.10	7.95 (1.06)	7.85 (1.02)	8.24 (1.26)	8.20 (1.27)	8.16 (0.49)
1980	8.98	10.29 (1.33)	10.29 (1.33)	10.35 (1.28)	10.34 (1.29)	9.14 (0.47)
1981	8.56	10.35 (0.73)	10.29 (0.71)	10.38 (1.09)	10.38 (1.09)	7.64 (1.61)
			Mallard, Stratum 30/31			
1977	35.42	46.84 (2.94)	46.95 (2.50)	36.28 (6.62)	38.03 (6.32)	32.00 (3.57)
1978	16.40	15.14 (1.39)	14.85 (1.13)	18.41 (4.45)	19.22 (4.76)	17.10 (4.55)
1979	20.50	21.78 (2.02)	21.21 (2.20)	22.60 (3.64)	22.05 (3.64)	17.48 (1.39)
1980	22.25	23.11 (2.72)	22.71 (2.42)	23.71 (3.62)	24.69 (3.72)	24.59 (2.48)
1981	18.88	17.36 (2.82)	16.57 (3.09)	18.51 (3.45)	19.27 (3.58)	18.15 (3.30)
			Canvasback, Stratum 28/29			
1977	0.53	0.38 (0.10)	0.38 (0.10)	0.37 (0.15)	0.36 (0.15)	0.42 (0.11)
1978	0.40	0.38 (0.01)	0.38 (0.01)	0.37 (0.09)	0.36 (0.10)	0.43 (0.07)
1979	0.53	0.26 (0.12)	0.26 (0.12)	0.26 (0.08)	0.26 (0.08)	0.27 (0.14)
1980	0.24	0.50 (0.17)	0.50 (0.17)	0.49 (0.09)	0.49 (0.09)	0.45 (0.08)
1981	0.10	0.11 (0.06)	0.12 (0.06)	0.13 (0.10)	0.13 (0.10)	0.09 (0.06)
			Canvasback, Stratum 30/31			
1977	3.76	5.68 (1.59)	5.95 (1.29)	5.05 (0.72)	5.17 (0.71)	4.06 (0.87)
1978	1.60	0.64 (0.26)	0.68 (0.26)	1.14 (0.81)	1.21 (0.85)	0.79 (0.36)
1979	3.43	2.53 (0.37)	2.49 (0.43)	2.61 (0.78)	2.45 (0.78)	2.06 (0.33)
1980	3.20	3.83 (1.40)	4.38 (1.45)	3.58 (0.78)	3.78 (0.78)	4.74 (2.04)
1981	1.76	1.06 (0.20)	1.05 (0.23)	1.51 (0.85)	1.92 (0.96)	1.15 (0.28)

Table 4. Performance criteria for five estimators applied to two species and two strata.

		Mallard		Canvasback	
Criterion	Estimation	28/29	30/31	28/29	30/31
RMSE	$\hat{\theta}^{(1)}$	1.233	5.231	0.180	1.123
	$\hat{\theta}^{(2)}$	1.234	5.320	0.180	1.299
	$\hat{\theta}^{(3)}$	1.156	1.515	0.180	0.743
	$\hat{\theta}^{(4)}$	1.144	2.160	0.181	0.833
	$\hat{\theta}^{(5)}$	0.843	2.339	0.155	1.039
Average SE	$\hat{\theta}^{(1)}$	0.869	2.376	0.093	0.764
	$\hat{\theta}^{(2)}$	0.853	2.269	0.094	0.730
	$\hat{\theta}^{(3)}$	1.359	4.357	0.103	0.789
	$\hat{\theta}^{(4)}$	1.361	4.405	0.105	0.815
	$\hat{\theta}^{(5)}$	0.841	3.059	0.093	0.778
Coverage	$\hat{\theta}^{(1)}$	0.600	0.800	0.800	0.400
	$\hat{\theta}^{(2)}$	0.600	0.800	0.600	0.400
	$\hat{\theta}^{(3)}$	1.000	1.000	0.600	1.000
	$\hat{\theta}^{(4)}$	1.000	1.000	0.600	1.000
	$\hat{\theta}^{(5)}$	0.800	0.800	0.800	0.400