IMPROVING STABILITY OF ANNUAL STATE WATERFOWL HARVEST ESTIMATES IN HIGHLY-SKEWED DATA

Lois M. Moyer and Paul H. Geissler, Patuxent Wildlife Research Center

ABSTRACT

Questionnaires returned by waterfowl hunters to the Waterfowl Harvest Survey Section of the U. S. Fish and Wildlife Service sporadically contain records of extremely large harvests of ducks. Such reports can result in a state estimate for one year that is much larger than the estimates for other years. We approached the problem of unstable annual state estimates by Winsorizing the extreme reports of duck harvest and then adjusting the estimates to remove the bias resulting from the Winsorization. A simulation was conducted to find the method that best stabilizes the estimates as measured by the mean squared error. A computer program for calculating the estimators was developed and tested.

BACKGROUND

The U. S. Fish and Wildlife Service's Waterfowl Hunter Questionnaire Survey is an annual survey of waterfowl hunters that estimates the numbers of waterfowl harvested within states, flyways, and the nation (Department of the Interior 1979). Questionnaires are sent to approximately 100,000 potential hunters each year. On the average, about 70,000 questionnaires are returned. Hunters are contacted personally to verify figures they submitted if an unusually large harvest is reported. Harvest is estimated with a ratio estimate [(mean harvest per hunter)*(number of duck stamps sold)]. Impossibly large harvests are deleted from the files and appropriate corrections are made as a result of the hunter contacts. After preliminary screening, however, there still remain some reports of large duck harvests that are apparently correct but that also may have a large effect on the sample mean and consequently on the estimated harvest.

IDENTIFICATION OF LARGE HARVEST VALUES

The first step in stabilizing state harvest estimates is to identify the extreme values. The reported harvests are non-negative and usually small (Figures 1 and 2), with extreme values occurring only in the upper tail. The mode occurs at the minimum harvest of zero ducks. The extreme values in the upper tail appear to be accurate and a part of the underlying distribution. Rejection or Winsorization of values from the upper tail will result in a negative bias in the harvest estimates. However, these large values may have a great effect on the stability of the harvest estimates and should be investigated.

In this situation, we do not view the extreme values as outliers or discordant observations because we believe they are true values and part of the distribution. However, we use an outlier identification method to locate the observations that destabilize the harvest estimates. Instead of viewing these procedures as testing to see whether an extreme observation comes from the distribution, we view the test as simply identifying possible destabilizing observations. The significance level controls the number of extreme observations identified.

We require an outlier identification procedure that identifies multiple outliers in the upper tail of a gamma-like distribution and that is computationally feasible with fairly large sample sizes (up to 3000 per state).

In a review of outlier detection procedures, Barnett and Lewis (1978) suggested several methods that are suitable for multiple upper outliers in a gamma distribution. Unfortunately, most of the procedures are appropriate only for small or moderate sample sizes. Because our data look like a gamma distribution, we used a square root transformation on the observations and then applied an outlier detection procedure for a normal sample with unknown mean and variance (Barnett and Lewis 1978).

The generalized extreme studentized deviate (ESD) procedure for multiple outliers (Rosner 1975, 1983) was selected for identification of extreme observations. The ESD procedure is based on the extreme studentized deviates R_i, $i=1,2,\ldots,k$ that are computed from the successively reduced samples of size n, $n-1,\ldots,n-k+1$, respectively. If the data are ordered such that $x_{(i)} > x_{(i+1)}$ then

$R_i=(x_{(i)}-\bar{x}_i)/s_i$ where

$\bar{x}_i=(\Sigma_{j=i}^{n} x_{(j)})/(n-i+1)$ and

$s^2_i=\Sigma_{j=i}^{n}(x_{(j)}-\bar{x}_i)^2/(n-i)$.

We declare x_i to be an extreme value if

$R_i > \lambda$ where λ is a calculated

percentage point for $(n-i+1)$ observations for the ESD procedure (Rosner 1983). Approximate percentage points are provided for up to 10 outliers and a maximum sample size of 500. For values of n greater than 500, the normal distribution can be used to obtain approximate percentiles, with

$$\lambda = [z_p(n-i-1)]/[(n-i-2+z_p^2)(n-i)]^{1/2} \quad (1)$$

where $i=0,1,\ldots,k-1$, $p=1-[\alpha/(n-i)]$, α is the probability of a Type I error, and z_p is the p^{th} percentile of a normal distribution. This procedure is appropriate for normally distributed data and is an improvement over Rosner's earlier ESD procedure. The procedure controls the Type I error under both the hypothesis of no outliers and the alternative hypothesis of $1,2,\ldots,k-1$ outliers.

The ESD procedure appears to perform better than other procedures when outliers are all on the same side of the mean; furthermore, it is computationally simple (Rosner 1975). Simonoff (1982) notes that the sequential ESD techniques are preferable over other methods such as Johns' adaptive estimator because the ESD procedure is consistently effective for both large and small data sets. He demonstrates that the ESD technique is the preferred method over robust estimators for asymmetric outliers.

Rosner's ESD test statistics for the ESD procedures are not prone to the effect of masking, which is the inability of a testing procedure to identify a single outlier in the presence of several suspected values. Also, his procedure requires only the knowledge of k, the maximum number of outliers present (Beckman and Cook 1983). Rosner (1975) asserts that the greatest increase in overall power occurs when the outliers are on the same side of the mean (as in our data) and are approximately of the same magnitude. This is also the situation where the masking problem is most serious.

WINSORIZATION

After identifying the extreme observations, we wanted to reduce the effect of these values and stabilize the waterfowl harvest estimates. Early accommodation procedures included a suggestion by Riders (1933) that an observation which differs widely from the rest should be retained but assigned a smaller weight or be replaced with a value closer to the mean. A robust estimator of the mean can be obtained by Winsorizing, or replacing the extreme observations with the nearest retained neighbors and taking an unweighted average from the modified sample. Dixon (1960) showed the efficiency of Winsorized estimators. Tukey (1960) favored Winsorized means on the grounds that long-tailed distributions are more common than short. We used a form of Winsorization called semi-Winsorization or the S-rule (Guttman and Smith 1969).

If $R_i=(x_{(i)}-\bar{x}_i)/s_i > \lambda$ and $R_{i+1} < \lambda$

for $1 \leq i < k$, we replace $x_{(1)},\ldots,x_{(i)}$ with $x_{i+1}+\lambda s_{i+1}$.

In other words, we replace any observation greater than $\bar{x}_i+\lambda s_i$ with that value rather than its nearest neighbor. The estimates are then based on the modified sample.

BIAS ADJUSTMENT

Winsorizing extreme values from the upper tail but not from the lower tail of the distribution biases the estimated annual state harvests to be too low. National estimates are not as seriously affected by the extreme values. Therefore, we adjusted Winsorized state estimates by a proportion so that they sum to the un-Winsorized national estimate as in:

$$A_i = W_i \; (\Sigma U_j/\Sigma W_j) \qquad (2)$$

where A_i is the adjusted estimate for state i, U_j is the un-Winsorized state estimate over all states and W_j is the Winsorized state estimate over all states. This adjustment takes the extreme harvest peaks that were removed by Winsorization and redistributes the peak harvest among the states, raising the lower values and decreasing the extreme values.

SIMULATION STUDY

A simulation was performed to compare three methods for estimating state harvests. These were the unadjusted estimates, the semi-Winsorized estimates and the semi-Winsorized estimates with a bias adjustment. From these three methods, we determined which method best stabilizes the estimates as measured by the mean squared error (MSE).

Fifty states in the nation were classified into 15 high success states and 35 low success states. Two empirical frequency distributions were constructed to represent these high and low success states using Waterfowl Harvest Survey data for the years 1978 through 1980. For each simulated year, a sample of 100 hunters was selected from the empirical distribution for each of the high success states and a sample of 50 hunters was selected for each of the low success states. Each year we obtained unadjusted estimates, semi-Winsorized estimates and bias-adjusted semi-Winsorized estimates. Because we only wanted to Winsorize the most extreme values in the tail, we set $\alpha = 0.001$. Mean biases and mean squared errors were calculated separately for high and low success states each year. Bias and MSE estimates were standardized by dividing by the true value and the true value squared, respectively. One hundred simulated years constitute the independent replicates. Estimates of the mean bias and mean squared error and their standard errors were calculated from the annual means.

Because many extreme values were already removed before the data were keypunched, we also contaminated some of the samples with the most extreme observations in the two true distributions. We substituted the largest true value for one other value in one high and one low success state each year (replicate).

SIMULATION RESULTS AND DISCUSSION

We found that our procedure only minimally reduced the MSE in the uncontaminated data ($P>0.05$, Table 1). With these data, extreme values do not occur frequently enough to be detected in our simulation. On the other hand, use of our procedure when it is not needed, does not increase the MSE.

As expected, Winsorization does result in a negative bias. The unadjusted estimates had the smallest bias, the bias-adjusted Winsorized values were next, and the Winsorized values had the largest bias. The adjustment removes the bias for all states combined ($P>0.05$); however, the high success states had a positive bias whereas the low success states had a negative bias ($P<0.01$). This bias of about 1% is not considered to be serious. It occurs because proportionally more harvest is removed by Winsorization in the low success states as compared to the high success states.

The contamination assured that some extreme values occured in each replication. Both the Winsorized and adjusted Winsorized estimates in the contaminated data reduced the MSE ($P<0.01$) but no difference could be found between these procedures ($P>0.05$, Duncan's multiple range test). In the contaminated samples, as in the uncontaminated samples, the bias adjustment corrected any significant bias resulting from the semi-Winsorization.

In conclusion, our methods to stabilize the annual waterfowl state harvest estimates appear to decrease the MSE whenever there are extreme values in the data, and do not increase it when extreme values are not present. The adjusted Winsorized estimates are recommended because they reduce the bias that is introduced by the Winsorization. The procedure does not introduce any serious bias into the estimates.

The authors would like to thank Christine M. Bunck and Dr. B. Kenneth Williams for their review of various drafts of the paper and many helpful suggestions.

REFERENCES

BECKMAN, R.J. and R.D. Cook (1983), "Outlier.....s," Technometrics, 25, 119-149.

BARNETT, V. and T. Lewis (1978), Outliers in Statistical Data, John Wiley and Sons.

DIXON, W. J. (1960), "Simplified Estimation from Censored Normal Samples," Annals of of Mathematical Statistics, 31, 385-391.

GUTTMAN, I. and D.E. Smith (1969), "Investigation of Rules for Dealing with Outliers in Small Samples from the Normal Distribution. I: Estimator of the Mean," Technometrics, 11, 527-550.

HUBER, P.J. (1972), "Robust Statistics: A Review (The 1972 Wald Lecture)," Annals of Mathematical Statistics, 43, 1041-1067.

RIDER, P.R. (1933), "Criteria for Rejection of Observations," Washington University Studies--New Series, Science and Technology, 8, 3-23.

ROSNER, B. (1975), "On the Detection of Many Outliers," Technometrics, 17, 221-227.

ROSNER, B. (1983), "Percentage Points for a Generalized ESD Many-Outlier Procedure," Technometrics, 25, 165-172.

SIMONOFF, J.S. (1982), "A Comparison of Robust Methods and Detection of Outliers Techniques When Estimating a Location Parameter," Proceedings of the SAS Users Group International Conference, 278-281.

TUKEY, J.W. (1960), "A Survey of Sampling from Contaminated Distributions," in Olkin (1960).

UNITED STATES DEPARTMENT OF THE INTERIOR (1982), Fish and Wildlife Special Scientific Report, Wildlife No. 246, Waterfowl Status Report 1979, Washington, D. C.

TABLE 1

RELATIVE MSE AND BIAS ESTIMATES FOR WATERFOWL HARVEST SURVEY

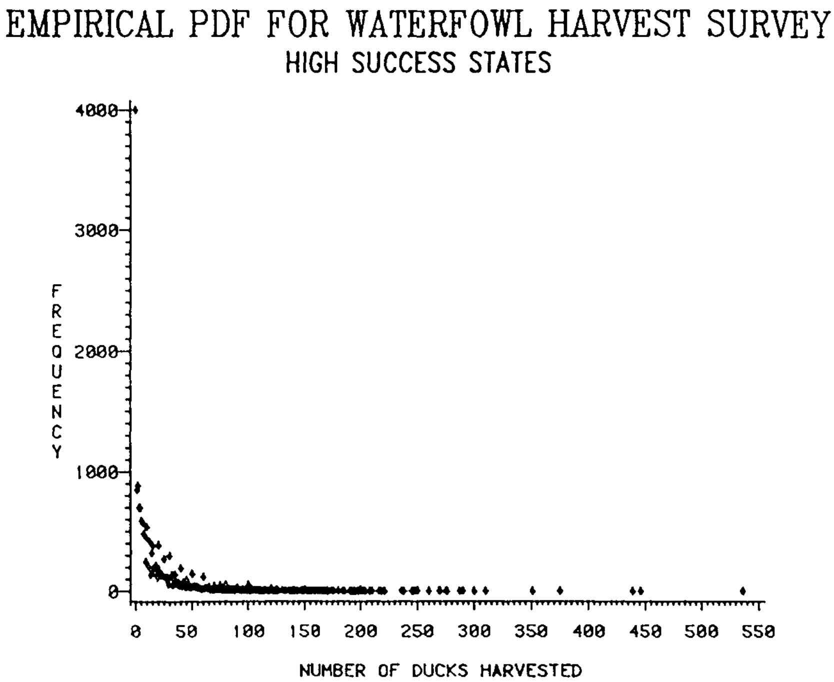
UNCONTAMINATED SURVEY DATA

HIGH/LOW	METHOD	MEAN BIAS	MEAN MSE	SE BIAS	SE MSE
H	NONE	.0082	.0266	.0039	.0011
H	WIN	.0016	.0304	.0039	.0011
H	ADJ	.0137	.0263	.0039	.0011
L	NONE	.0044	.0815	.0047	.0021
L	WIN	-.0264	.0770	.0046	.0020
L	ADJ	-.0147	.0784	.0047	.0020
COMBINED	NONE	.0063	.0541	.0031	.0023
COMBINED	WIN	-.0124	.0537	.0032	.0020
COMBINED	ADJ	-.0005	.0523	.0032	.0022

CONTAMINATED SURVEY DATA

HIGH/LOW	METHOD	MEAN BIAS	MEAN MSE	SE BIAS	SE MSE
H	NONE	.0278	.0332	.0040	.0014
H	WIN	.0101	.0329	.0040	.0011
H	ADJ	.0340	.0301	.0041	.0012
L	NONE	.0283	.1036	.0049	.0027
L	WIN	-.0162	.0820	.0049	.0023
L	ADJ	.0071	.0857	.0050	.0025
COMBINED	NONE	.0281	.0684	.0032	.0029
COMBINED	WIN	-.0031	.0575	.0033	.0022
COMBINED	ADJ	.0205	.0579	.0034	.0024

FIGURE 1


EMPIRICAL PDF FOR WATERFOWL HARVEST SURVEY
HIGH SUCCESS STATES
FREQUENCY
4000
3000
2000
1000
0
0 50 100 150 200 250 300 350 400 450 500 550
NUMBER OF DUCKS HARVESTED

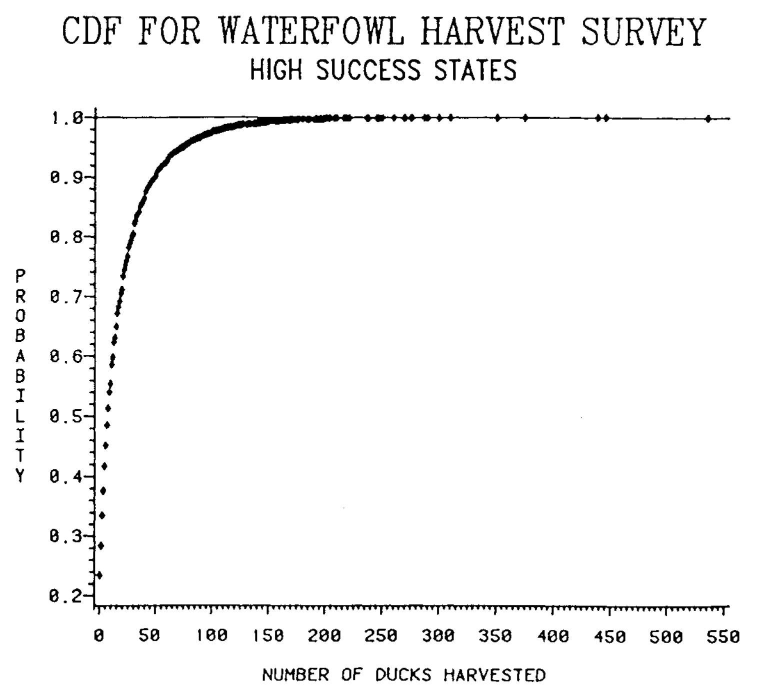
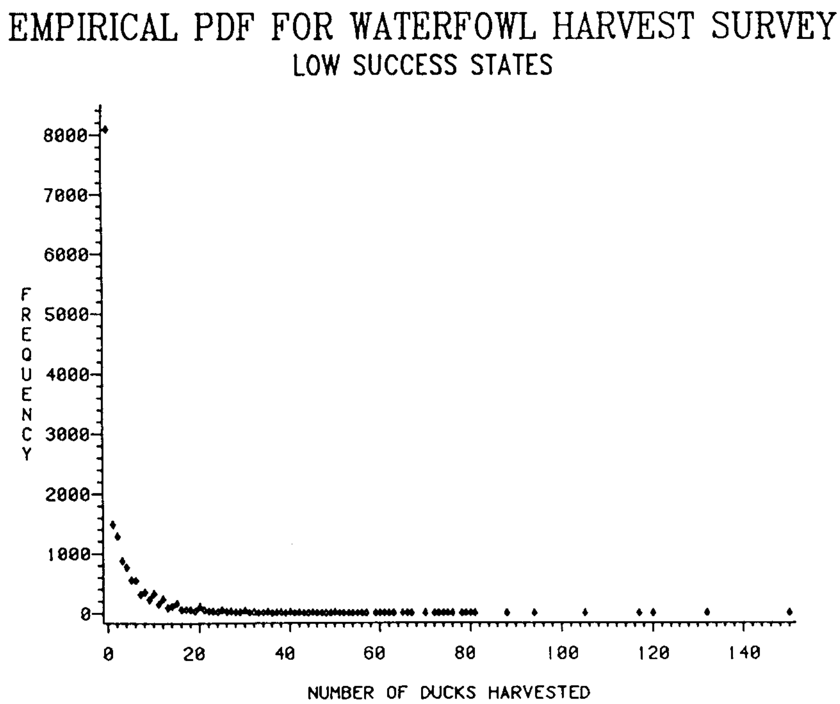
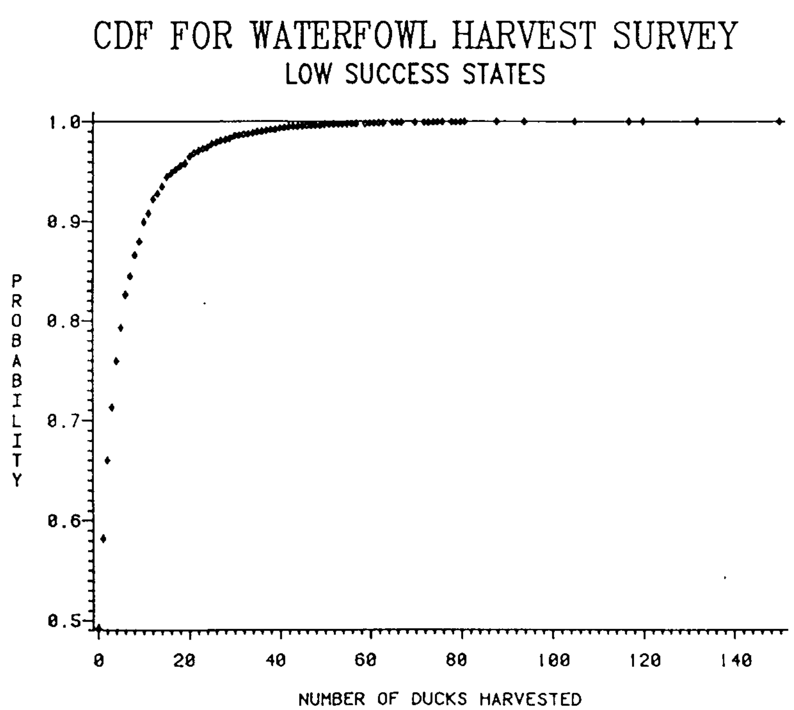

CDF FOR WATERFOWL HARVEST SURVEY
HIGH SUCCESS STATES
PROBABILITY
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0 50 100 150 200 250 300 350 400 450 500 550
NUMBER OF DUCKS HARVESTED

FIGURE 2



EMPIRICAL PDF FOR WATERFOWL HARVEST SURVEY
LOW SUCCESS STATES
FREQUENCY
8000
7000
6000
5000
4000
3000
2000
1000
0
0
20
40
60
80
100
120
140
NUMBER OF DUCKS HARVESTED



CDF FOR WATERFOWL HARVEST SURVEY
LOW SUCCESS STATES
PROBABILITY
1.0
0.9
0.8
0.7
0.6
0.5
0
20
40
60
80
100
120
140
NUMBER OF DUCKS HARVESTED