EMPIRICAL COMPARISON OF UNIFORM AND NON-UNIFORM PROBABILITY SAMPLING FOR ESTIMATING NUMBERS OF RED-COCKADED WOODPECKER COLONIES

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

ABSTRACT

Four sampling and estimation methods for estimating the number of red-cockaded woodpecker colonies on National Forests in the Southeast were compared, using samples chosen from simulated populations based on the observed sample. The methods included (1) simple random sampling without replacement using a mean per sampling unit estimator, (2) simple random sampling without replacement with a ratio per pine area estimator, (3) probability proportional to "size" sampling with replacement, and (4) probability proportional to "size" without replacement using Murthy's estimator. The survey sample of 274 National Forest compartments (1000 acres each) constituted a superpopulation from which simulated stratum populations were selected with probability inversely proportional to the original probability of selection. Compartments were originally sampled with probabilities proportional to the probabilities that the compartments contained woodpeckers ("size"). These probabilities were estimated with a discriminant analysis based on tree species and tree age. The ratio estimator would have been the best estimator for this survey based on the mean square error. However, if more accurate predictions of woodpecker presence had been available, Murthy's estimator would have been the best. A subroutine to calculate Murthy's estimates is included; it is computationally feasible to analyze up to 10 samples per stratum.

SURVEY

We had the task of designing a survey to estimate the number of red-cockaded woodpecker colonies on National Forests (Lennartz et al. in press). This is an endangered woodpecker that lives in old pine trees in the Southeastern States. It is standard forestry practice to harvest pine trees before they are old enough to support red-cockaded woodpeckers; therefore a conflict has resulted.

The survey was stratified by ranger districts within national forests. The sampling units were the approximately 1000 acre compartments used for managing the national forests. There were 30 strata with 43 to 207 compartments per stratum (mean=98, s=48). From 4 to 21 compartments were randomly selected from each stratum based on an optimal allocation (mean=9.1, s=4.8). Teams of biologists searched selected compartments for red-cockaded woodpecker colonies. Woodpecker colonies are easily visible to the searcher because of the white gummy substance that cascades from the woodpecker holes in living pine trees.

We attempted to increase the efficiency of the survey by selecting compartments with probability proportional to size. "Size" was the probability that the compartment contained a woodpecker colony. These probabilities ("size" of the compartments) were estimated using a discriminant analysis based on tree species and tree age. If a compartment had been previously searched for woodpeckers, the "size" was doubled if woodpeckers had been found and halved if they had not.

Murthy's estimator (Cochran, 1977: 263-265) was used based on a review by Rao (1978: 75) who cited studies "which indicate that Murthy's method might be preferable over other methods ... when a stable estimator as well as a stable variance estimator are required." Because of the computational requirements of Murthy's estimator, no more than 10 samples per stratum could be analyzed. CPU times on a Hewlett Packard 3000 minicomputer were:

Samples	4	5	6	7	8	9	10
Seconds	27	29	30	37	85	533	5231

CPU seconds can be estimated by 26 + 0.0014339 n! where n is the number of samples. It took about 1.5 hours to analyze 10 samples, and would have taken about 16 hours to analyze 11 samples. While it is clear from the variance formula that computational times are related to n!, guidelines on reasonable sample sizes are not readily available. When more than 10 samples had been drawn, the national forest-ranger district strata were poststratified into approximately equal sized substrata. Six of the 30 strata had more than 10 samples and had to be poststratified, 4 strata into 2 substrata and 2 into 3 substrata. Poststratification increases the variance because samples are not distributed proportionally to the poststrata (Cochran 1976: 135). The increase is about (L-1)/Ln where L is the number of poststrata and n is the mean number of sampling units per poststratum. For 2 poststrata with 6 sampling units per strata, this is about an 8% increase.

EMPIRICAL STUDY

With the advantage of hindsight, we compared four sampling and analytic methods that we could have used for the survey:

1. Mean per unit estimator with simple random sampling, equal probability and without replacement (see Cochran 1977: 21-26).

$$\hat{Y}_1 = N \, \Sigma_i^n \, y_i / n \qquad (1)$$

$$v(\hat{Y}_1) = N^2 \, (1 - n/N) \, \Sigma_i^n \, (y_i - \bar{y})^2 \, / \, n(n-1) \qquad (2)$$

where
Y = total number of colonies in stratum
y_i = number of colonies in ith compartment
$\bar{y} = \Sigma_i^n \, y_i / n$
N = number of compartments in stratum
n = number of compartments in sample

2. Ratio per pine area with simple random sampling, equal probability and without replacement (see Cochran 1977: 150-156).

$$\hat{Y}_2 = \hat{R} X \quad (3)$$

$$v(\hat{Y}_2) = N^2 (1-n/N) \Sigma_i^n (y_i - \hat{R}x_i)^2 / n(n-1) \quad (4)$$

where
X = total pine acres in stratum
x_i = pine acres in ith compartment
$\hat{R} = \Sigma_i^n y_i / \Sigma_i^n x_i$

3. Probability proportional to "size" with replacement (see Cochran 1977: 252-255).

$$\hat{Y}_3 = \Sigma_i^n y_i/z_i n \quad (5)$$

$$v(\hat{Y}_3) = \Sigma_i^n (y_i/z_i - \hat{Y}_3)^2/ n(n-1) \quad (6)$$

where
z_i = probability of selecting ith compartment

4. Murthy's estimator with probability proportional to "size" and without replacement (see Cochran 1977: 263-265).

$$\hat{Y}_4 = \Sigma_i^n P(s|i) y_i / P(s) \quad (7)$$

$$v(\hat{Y}_4) = \Sigma_i^n \Sigma_{j>i}^n [P(s)P(s|i,j)- P(s|i)P(s|j)] z_i z_j (y_i/z_i - y_j/z_j)^2/ P(s)^2 \quad (8)$$

where
$P(s) = \Sigma_{i\neq j\neq k}^n z_i z_j/(1-z_i) z_k/(1-z_i-z_j)$

unconditional probability of drawing sample (for n=3)

$P(s|i)= \Sigma_{j\neq k}^n z_j/(1-z_i) z_k/(1-z_i-z_j)$

conditional probability of drawing sample, given that the ith compartment was drawn first (for n=3)

$P(s|i,j)= \Sigma_k^n z_k/(1-z_i-z_j)$

conditional probability of drawing sample, given that the ith and jth compartments were selected (in either order) in the first 2 draws (n=3)

An empirical approach was used to investigate these 4 methods for the specific conditions of the red-cockaded woodpecker survey. The actual sampled compartments from all strata were used as a combined superpopulation from which the artificial stratum populations were drawn. A Tausworthe random number generator was used (Kennedy and Gentle, 1980: 155). Sampling from the superpopulation was with replacement and inversely proportional to the probability that the original compartments were selected. This selection reversed the over-representation of compartments that were originally assigned high selection probabilities so that the simulated populations were as similar to the real population as possible. Fifty trials each with 25 replications (1250 replications in all) were run for 9 situations. The situations included combinations of 3 population sizes (20, 100, and 500 compartments), and 3 numbers of strata (1, 10, and 20); each strata had 4 sample compartments. For each trial, the bias, estimated variance and mean square error, and the proportion of 90% and 95% confidence intervals that enclosed the true value were output to a disk file for summarization. These five values were used as criteria for evaluating the four sampling and estimation methods. The relative bias, relative mean square error and relative variance were defined as

$$\text{Rel. bias} = \Sigma_r^m (\hat{Y}_r-Y) / Y m \quad (9)$$

$$\text{Rel. MSE} = \Sigma_r^m (\hat{Y}_r-Y)^2 n / Y^2 m \quad (10)$$

$$\text{Rel. var.} = \Sigma_r^m v(Y6_r) n / m \quad (11)$$

where r=1,...,m indexes replications.

Note that the relative mean square error and relative variance have been multiplied by the number of observations n so that they would be comparable over trials with different sample sizes.

Unfortunately the discriminant anaylsis predictions of which compartments contained woodpeckers were not as good as we had hoped. Additional trials were conducted to see how the non-uniform probability methods would perform if accurate predictions of the probability that a compartment contains woodpeckers were available. In these trials, the sampling units contained woodpeckers only when the posterior probability from the discriminant analysis was greater than 0.5. In that case, the probabilities of compartments containing 1, 2, ..., or 11 woodpecker colonies were 0.44, 0.17, 0.14, 0.11, 0.03, 0.05, 0.02, 0.02, 0.01, 0, 0.01, respectively, based on the observed frequency distribution.

RESULTS

For the red-cockaded woodpecker survey, equal probability sampling with a ratio to pine area would have been the best method as judged by the mean square error (Table 1). Unequal probability sampling with Murthy's method that we used for the survey was second best. However, Murthy's method requires extensive computations which effectively limit one to 10 or fewer samples per stratum. Although none of the estimators is seriously biased, the ratio estimator underestimated the number of colonies by about 2 percent. Note that unbiased and reduced bias ratio estimators are available (see Cochran 1977: 174-177). The variance of the ratio estimator underestimated the mean square error more than the other methods.

The disappointing performance of Murthy's method may be due to our inability to predict the presence of woodpeckers (estimate the "size") as well as we had hoped. The trials with improved "size" estimates showed greatly improved performance with Murthy's method. If more accurate predictions of woodpecker presence had been available, Murthy's estimator would have been the best. The squared correlation between the probability and the number of colonies in the sample was only 9.7%. The discriminant function had the following classification rates with the original training compartments that were available for planning the survey and with the National Forest sample compartments (using modified probabilities if the compartment had been previously searched):

Training Compartments

	Predicted		
Observed	Absent	Present	Total
Absent	244(76%)	77(24%)	321(100%)
Present	99(39%)	152(61%)	251(100%)
Total	343(60%)	229(40%)	572(100%)

Sample Compartments

	Predicted		
Observed	Absent	Present	Total
Absent	66(46%)	78(54%)	144(100%)
Present	21(16%)	109(84%)	130(100%)
Total	87(32%)	187(68%)	274(100%)

The 64% correct classification rate for the sample compartments compares favorably with the 69% correct classification rate for the training compartments. Although 84% of the sample compartments with woodpeckers were correctly classified, only 46% of the compartments without woodpeckers were correctly classified. The low correct classification rates for compartments without woodpeckers results in poor overall predictions because few compartments have woodpeckers. We can only speculate on the reason for our poor predictions. It is possible that the compartments that were predicted to have woodpeckers have good habitat but that much of the habitat is not occupied by this endangered species. It is also possible that we did not have a representative training sample of compartments without woodpeckers because the training compartments were not selected randomly.

Confidence interval widths were underestimated especially with small sample sizes. With larger sample sizes the widths were improved. With 60 degrees of freedom for error, the 90% confidence intervals included the true value about 87% of the samples.

The authors would like to thank Dr. B. Kenneth Williams, also of Patuxent Wildlife Research Center, for his review of earlier drafts of this paper and his many helpful suggestions.

REFERENCES

COCHRAN, W.G. (1977), Sampling Techniques, New York: John Wiley.

KENNEDY, W.J. and J.E. Gentle (1980), Statistical Computing, New York: Marcel Dekker.

KERNIGHAN, B.W. and P.J. Plauger (1976), Software Tools, Addison-Wesley.

LENNARTZ, M.R., P.H. Geissler, R.F. Harlow, R.C. Long, K.M. Chitwood, and J.A. Jackson (in press), "An Estimate of Red-cockaded Woodpecker Populations on Federal Lands in the South," in Proceedings Red-cockaded Woodpecker Symposium II., Jan. 27-29, 1982, Panama City, FL.

RAO, J.N.K. (1978), "Sampling Designs Involving Unequal Probabilities of Selection and Robust Estimation of a Finite Population Total," in Contributions to Survey Sampling and Applied Statistics, ed. H.A. David, New York: Academic Press, 69-87.

Table 1. Results of empirical study of red-cockaded woodpecker survey, sampling from a superpopulation consisting of the observed samples. Marginals for population size and number of strata (each with 4 samples) are shown. Estimates are followed by their standard errors which were calculated among trials each of which had 50 replicates. In some trials, the estimate of "size" was improved by generating a nonzero number of woodpecker colonies whenever the discriminant analysis predicted their presence.

Population size	Number strata	Number trials	Equal-----probability: Mean per unit	Equal-----probability: Ratio to pine area	Unequal---probability: Probability proportional to "size"	Unequal---probability: Murthy's method
			Relative mean square error, standard error with actual "size" estimates			
		450	2.59,.07	2.20,.06	2.63,.07	2.43,.07
20		150	2.10,.15	1.84,.11	2.38,.17	1.89,.13
100		150	2.78,.10	2.34,.10	2.70,.11	2.60,.11
500		150	2.88,.08	2.41,.09	2.80,.09	2.80,.09
	1	150	2.56,.14	2.18,.12	2.69,.16	2.48,.14
	10	150	2.76,.12	2.38,.11	2.54,.10	2.31,.09
	20	150	2.44,.10	2.04,.07	2.64,.11	2.48,.10
			with improved "size" estimates			
		450	3.91,.09		2.55,.07	2.35,.06
			Estimated relative variance, standard error with actual "size" estimates			
		450	2.81,.07	1.81,.04	2.74,.07	2.52,.06
20		150	2.38,.18	1.54,.08	2.59,.18	2.06,.14
100		150	2.94,.09	1.95,.06	2.81,.11	2.71,.10
500		150	3.10,.08	1.95,.04	2.82,.06	2.80,.06
	1	150	3.23,.17	2.03,.08	3.09,.17	2.80,.14
	10	150	2.66,.10	1.74,.06	2.57,.09	2.39,.08
	20	150	2.52,.09	1.66,.05	2.56,.09	2.38,.09
			with improved "size" estimates			
		450	4.02,.08		2.50,.06	2.29,.05
			Relative bias, standard error with actual "size" estimates			
		450	.001,.005	-.018,.005	.010,.005	.011,.005
20		150	.005,.007	-.013,.008	.009,.008	.010,.007
100		150	-.003,.009	-.024,.009	.011,.011	.012,.011
500		150	.002,.009	-.017,.009	.011,.009	.011,.009
	1	150	.004,.014	-.010,.013	.038,.015	.036,.014
	10	150	-.005,.005	-.025,.007	-.010,.005	-.008,.004
	20	150	.005,.003	-.019,.004	.004,.003	.005,.003
			with improved "size" estimates			
		450	-.014,.008		-.006,.005	-.006,.005
			Proportion of 90% confidence intervals with true value, standard error with actual "size" estimates			
	1	150	.945,.008	.909,.007	.792,.009	.796,.009
	10	150	.862,.006	.816,.007	.867,.005	.873,.005
	20	150	.892,.005	.840,.006	.877,.005	.882,.005
			with improved "size" estimates			
	1	150	.821,.009		.796,.008	.800,.008
	10	150	.877,.006		.873,.006	.871,.006
	20	150	.892,.005		.894,.005	.893,.005
			Proportion of 95% confidence intervals with true value, standard error with actual "size" estimates			
	1	150	.990,.004	.962,.004	.849,.008	.847,.009
	10	150	.909,.005	.873,.005	.910,.004	.918,.004
	20	150	.934,.004	.900,.005	.928,.004	.929,.004
			with improved "size" estimates			
	1	150	.886,.008		.854,.007	.858,.007
	10	150	.925,.004		.923,.005	.924,.005
	20	150	.939,.004		.940,.004	.942,.004

```
#SUBROUTINE FOR MURTHY'S METHOD
#
# Subroutine to calculate Murthy's estimates,
# written in RATFOR (Kernignon and Plauger,
# 1976), a FORTRAN preprocessor that translates
# the structured source into a FORTRAN
# subroutine.  A copy of the resulting FORTRAN
# program in available on request.  A "#"
# indicates that the remainder of the line is
# a comment.  A "_" signals the continuation of
# a statement.  "DO I=1,N statement" specifies
# that the statement is to be executed N times
# with I=1,2,...,N. Compound statements may be
# used.  They are indicated by "$( statement-1
# statement-2 ... $)".  This construction is
# similar to the PL/I "DO; ... END;" and the
# Pascal "BEGIN ... END;".  "BREAK" transfers
# control out of the current loop, while "NEXT"
# transfers control to the next iteration of the
# current loop.

SUBROUTINE MURTHY(NSAMP,Y,Z,ESTIMATE,VARIANCE)
# NSAMP = no. of samples (2<=NSAMP<=10)
#     (input, integer)
#     NSAMP>10 requires excessive computing
# Y = array of observed totals for sampling
#     units 1, 2, ..., NSAMP
#     (input, double precision)
# Z = array of probabilities that sampling units
#     1, 2, ..., NSAMP are selected if a single
#     sample was drawn (input, double precision)
# ESTIMATE = estimated total
#     (output, double precision)
# VARIANCE = estimated variance of total
#     (output, double precision)
INTEGER NSAMP,I,I1,I2,I3,I4,I5,I6,I7,I8,I9,I10,
   ISU(10)
DOUBLE PRECISION Y(10),Z(10),
   ESTIMATE,VARIANCE,TOTALZ,DEN1,
   DEN2,DEN3,DEN4,DEN5,DEN6,DEN7,DEN8,DEN9,
   DEN10,
   PROB,P1,P2,P3,P4,P5,P6,P7,P8,P9,P10,
   P,PIJ(10,10)
LOGICAL INSAMP(10),EOF10,
   NSAMP2,NSAMP3,NSAMP4,NSAMP5,NSAMP6,NSAMP7,
   NSAMP8,NSAMP9,
   NSAMP10 # TRUE IF SAMPLE SIZE
IF (NSAMP>10)
   $(
   WRITE (6,*) ' NSAMP FOR MURTHYS METHOD',
      NSAMP,' RETURN ZEROS'
   ESTIMATE=0D0
   VARIANCE=0D0
   RETURN
   $)
IF (NSAMP==0)
   $(
   WRITE(6,*) ' NSAMP=0 FOR MURTHY''S METHOD'
   ESTIMATE=0.0D0
   VARIANCE=0.0D0
   RETURN
   $)
IF (NSAMP==1)
   $(
   WRITE(6,*) ' NSAMP=1 FOR MURTHY''S METHOD'
   ESTIMATE=Y(1)/Z(1)
   VARIANCE=0.0D0
   RETURN
   $)
IF (NSAMP==2) NSAMP2=.TRUE.
      ELSE    NSAMP2=.FALSE.
IF (NSAMP==3) NSAMP3=.TRUE.
      ELSE    NSAMP3=.FALSE.
IF (NSAMP==4) NSAMP4=.TRUE.
      ELSE    NSAMP4=.FALSE.
IF (NSAMP==5) NSAMP5=.TRUE.
      ELSE    NSAMP5=.FALSE.
IF (NSAMP==6) NSAMP6=.TRUE.
      ELSE    NSAMP6=.FALSE.
IF (NSAMP==7) NSAMP7=.TRUE.
      ELSE    NSAMP7=.FALSE.
IF (NSAMP==8) NSAMP8=.TRUE.
      ELSE    NSAMP8=.FALSE.
IF (NSAMP==9) NSAMP9=.TRUE.
      ELSE    NSAMP9=.FALSE.
IF (NSAMP==10) NSAMP10=.TRUE.
      ELSE    NSAMP10=.FALSE.
DO I=1,NSAMP
   $(
   INSAMP(I)=.FALSE.
DO J=1,NSAMP
   $(
   PIJ(I,J)=0.0D0
$) $)

# Start sample loops to compute Murthy's
# estimator with probability proportional
# to "size", without replacement (see
# equations 7 and 8 in text).  There is
# 1 nested loop for each sampling unit.
# INSAMP(i) is true if the ith
# sampling unit is already in the sample and
# control should be transferred to the next
# iteration of the loop.  NSAMPn is true if
# there are n sampling units and the inner loops
# should be skipped.
P1=0.0D0
DO I1=1,NSAMP
   $(
   INSAMP(I1)=.TRUE.
   DEN1=1.0D0-Z(I1)
   P2=0.0D0
DO I2=1,NSAMP
   $(
   IF (INSAMP(I2)) NEXT
   IF (NSAMP2)
      $(
      P2=Z(I2)
      PIJ(1,2)=1.0D0
      BREAK
      $)
   INSAMP(I2)=.TRUE.
   DEN2=DEN1-Z(I2)
   P3=0.0D0
DO I3=1,NSAMP
   $(
   IF (INSAMP(I3)) NEXT
   IF (NSAMP3)
      $(
      P3=Z(I3)
      BREAK
      $)
   INSAMP(I3)=.TRUE.
   DEN3=DEN2-Z(I3)
   P4=0.0D0
DO I4=1,NSAMP
   $(
   IF (INSAMP(I4)) NEXT
```

```
    IF (NSAMP4)
       $(
       P4=Z(I4)
       BREAK
       $)
    INSAMP(I4)=.TRUE.
    DEN4=DEN3-Z(I4)
    P5=0.0D0
 DO I5=1,NSAMP
    $(
    IF (INSAMP(I5)) NEXT
    IF (NSAMP5)
       $(
       P5=Z(I5)
       BREAK
       $)
    INSAMP(I5)=.TRUE.
    DEN5=DEN4-Z(I5)
    P6=0.0D0
 DO I6=1,NSAMP
    $(
    IF (INSAMP(I6)) NEXT
    IF (NSAMP6)
       $(
       P6=Z(I6)
       BREAK
       $)
    INSAMP(I6)=.TRUE.
    DEN6=DEN5-Z(I6)
    P7=0.0D0
 DO I7=1,NSAMP
    $(
    IF (INSAMP(I7)) NEXT
    IF (NSAMP7)
       $(
       P7=Z(I7)
       BREAK
       $)
    INSAMP(I7)=.TRUE.
    DEN7=DEN6-Z(I7)
    P8=0.0D0
 DO I8=1,NSAMP
    $(
    IF (INSAMP(I8)) NEXT
    IF (NSAMP8)
       $(
       P8=Z(I8)
       BREAK
       $)
    INSAMP(I8)=.TRUE.
    DEN8=DEN7-Z(I8)
    P9=0.0D0
 DO I9=1,NSAMP
    $(
    IF (INSAMP(I9)) NEXT
    IF (NSAMP9)
       $(
       P9=Z(I9)
       BREAK
       $)
    INSAMP(I9)=.TRUE.
    DEN9=DEN8-Z(I9)
 DO I10=1,NSAMP
    $(
    IF (INSAMP(I10)) NEXT
       BREAK
 $) # I10
    P9=P9+Z(I9)/DEN9*Z(I10)
    INSAMP(I9)=.FALSE.
 $) # I9
    P8=P8+Z(I8)/DEN8*P9
    INSAMP(I8)=.FALSE.
 $) # I8
    P7=P7+Z(I7)/DEN7*P8
    INSAMP(I7)=.FALSE.
 $) # I7
    P6=P6+Z(I6)/DEN6*P7
    INSAMP(I6)=.FALSE.
 $) # I6
    P5=P5+Z(I5)/DEN5*P6
    INSAMP(I5)=.FALSE.
 $) # I5
    P4=P4+Z(I4)/DEN4*P5
    INSAMP(I4)=.FALSE.
 $) # I4
    P3=P3+Z(I3)/DEN3*P4
    INSAMP(I3)=.FALSE.
 $) # I3
    P2=P2+Z(I2)/DEN2*P3
    PIJ(I1,I2)=P3/DEN2
    INSAMP(I2)=.FALSE.
 $) # I2
    P1=P1+Z(I1)/DEN1*P2
    PIJ(I1,I1)=P2/DEN1
    INSAMP(I1)=.FALSE.
 $) # END OF I1 LOOP

 # Calculate estimate of total and its variance
 # PROB = unconditional probability of drawing
 # sample.
 # PIJ(I,I) = conditional probability of drawing
 # sample given that the Ith sampling unit was
 # drawn first.
 # PIJ(I,J) = conditional probability of drawing
 # sample given that the Ith and Jth sampling
 # unit were drawn first.
 ESTIMATE=0
 DO I=1,NSAMP
    ESTIMATE=ESTIMATE+PIJ(I,I)*Y(I)
 PROB=P1
 VARIANCE=0
 DO I=1,NSAMP-1
    $(
 DO J=I+1,NSAMP
    $(
    P=Y(I)/Z(I)-Y(J)/Z(J)
    VARIANCE=VARIANCE + _
       (PROB*PIJ(I,J)-PIJ(I,I)*PIJ(J,J)) * _
       Z(I)*Z(J)*P*P
 $) $)
 ESTIMATE=ESTIMATE/PROB
 VARIANCE=VARIANCE/PROB/PROB
 RETURN
 END
```