# EMPIRICAL METHODS in the EVALUATION of ESTIMATORS

A case study based on insect density and survival data

by Gerald S. Walton
and C. J. DeMars


USDA FOREST SERVICE RESEARCH PAPER NE-272
1973
FOREST SERVICE, U.S. DEPARTMENT OF AGRICULTURE
NORTHEASTERN FOREST EXPERIMENT STATION
6816 MARKET STREET, UPPER DARBY, PA. 19082
WARREN T. DOOLITTLE, DIRECTOR

## The Authors

GERALD S. WALTON has served as a mathematical statistician in the biometrics branch of the Northeastern Forest Experiment Station since 1969. He received a bachelor of arts degree in mathematics from the University of California at Berkeley in 1960 and a master's degree in forest science from Harvard University in 1965.

C. J. DeMARS, a research entomologist, has been engaged in the study of bark beetles at the Pacific Southwest Forest and Range Experiment Station since 1957. He received a bachelor of science degree in forest science at the University of Georgia in 1953 and a Ph.D. degree in entomology at the University of California at Berkeley in 1966.

MANUSCRIPT RECEIVED FOR PUBLICATION 21 JUNE 1971

# EMPIRICAL METHODS in the EVALUATION of ESTIMATORS

A case study based on insect density and survival data


## ABSTRACT

The authors discuss the problem of selecting estimators of density and survival by making use of data on a forest-defoliating larva, the spruce budworm. Various estimators are compared. The results show that, among the estimators considered, ratio-type estimators are superior in terms of bias and variance. The methods used in making comparisons, particularly simulation, are generally applicable to situations in which data appropriate for making estimates are on hand.


THE EFFICIENCY of estimates of population parameters is affected by the choices of both sampling plan and estimators. When a population parameter to be estimated is the mean of some variable, there is generally a best estimator for a given well-defined sampling method. It is then possible to compare different sampling methods in terms of the precision with which the best estimator for each sampling method estimates the population parameter of interest. However, when the population parameter to be estimated is not a weighted sum, over the population, of a variable, there will be a "best" estimator only under very restrictive assumptions about the population—assumptions that, in practice, cannot be readily verified.

In this paper, we are concerned with the estimation of density and survival in a population of forest-defoliating larvae under simple sampling methods. Generally "best" estimators do not exist for these parameters, so our investigations, of necessity, are pragmatic. We have used actual data for the spruce budworm *(Choristoneura occidentalis* Freeman).

We have also tried, in an informal way, to see why the better estimators performed as they did, so the results described might be generalized to some extent. And we hope the reader will find the methods we have used for comparing estimators, using a particular data base, of value. Certainly it is demonstrated that the choice of estimators is a matter of some consequence. The best estimators of survival and density were always at least 60 percent more efficient than the next best estimator for the data and sampling designs used in this study.

The selection of estimators for comparison would seem to be a straightforward problem. There are methods, such as maximum likelihood and minimum square error, which gener-

ally produce reasonably good estimators. The rub is that the joint distribution of the random variables involved must be known, if not to obtain good estimators to begin with, then to determine the bias and variance of any estimator.

## THE PROBLEM

Look at the example. A number of trees are randomly selected from a presumably infinite population of trees. From each tree, on each of two sampling dates, a fixed number of branches are randomly selected from the presumably infinite number of branches on the tree. Two random variables are measured: the number of insects on the branch and the branch surface area. These random variables will be written

$Y_{tij}$ for the number of insects

$X_{tij}$ for the branch surface area

where $t = 1, 2$ denotes the first or second sampling date; $i = 1, \ldots, n$ denotes the $i^{th}$ tree in the sample and $j = 1, \ldots, n_t$ denotes the number of the branch removed on sampling date t. The same trees are sampled at each date.

Our object is to estimate survival and density. But before we proceed, we must be specific in our definitions of these parameters. Clearly, the way in which the parameters are defined will affect the choice of estimators. Morris *(1955)* discussed the biological interpretation of certain insect population parameters. In our paper, we choose to define the density parameter at sampling period t by

$$D_t = Y_{t\cdot\cdot}/X_{t\cdot\cdot}$$

Y represents insects; and X, the branch surface area, is the product of the foliated branch length and greatest branch width. The dots denote sums over branches and trees in the population. Survival between time periods 1 and 2 is defined in terms of density, using the equation

$$S = D_2/D_1$$

At this point we inject a brief digression on the notational conventions peculiar to this paper. We will ignore the distinction between the true and measured values of the random variables X and Y, since they are presumed to be measured without error. A dot in place of a subscript denotes summation of the variable over that subscript. The standard $\Sigma$ operator will be used for more complex sums. The ranges of summations are denoted explicitly only when these are not clear from the context. That is, sums used in defining population parameters are taken over the whole population while sums used in defining estimates are taken over a sample from the population. We have dispensed with formal limiting procedures when dealing with functions of sums over infinite populations (e.g., $D_t$). Population parameters are represented by uppercase letters (S,D), while the corresponding estimators use lowercase letters (s,d). We differentiate between different estimators of the same population parameter by the use of uppercase subscripts ($s_M$,$s_R$). "t" always stands for sampling period. It is inevitable, since we will be discussing several estimators of the same parameter, that the reader will have trouble keeping the various estimators straight. In order to rectify this problem, all estimators are assembled in appendix III for ready reference. We return to the main theme of this paper by considering some implications of the definitions of population density and survival.

A common alternative to $D_t$ as a definition of the population density is

$$D_t^* = (1/N_{t\cdot}) \sum_{i,j} (Y_{tij}/X_{tij})$$

for a finite population with $N_{t\cdot}$ total branches.

Our preference for $D_t$ is based on the intuitive feeling that insect density should not change if the number of insects and total branch surface area is the same at both sampling periods. If, in such a situation, the insects migrate to branches with greater surface area, $D_t^*$ decreases, while $D_t$ remains constant. Since $D_t^*$ can be affected by the spacial distribution of the insects, we elect to estimate $D_t$.

The survival parameter, S, as defined, is not the number of survivors divided by the number of insects initially present, although it does reduce to this if branch surface area is unchanged between sampling dates. Rather, S might be termed "survival relative to branch

surface area." The remainder of the paper is devoted to the estimation of S and $D_t$, to the exclusion of other population parameters.

## TRIAL COMPARISON OF SOME ESTIMATORS

One method of comparing estimators is simply to use each one on an appropriate set of data and observe the results. We have two examples of this in table 1. These examples are based on part of the data on spruce budworm collected in Montana in 1965 by the Insecticide Evaluation Project of the Pacific Southwest Forest and Range Experiment Station, USDA Forest Service (*Williams and Walton 1968*). The data are from two crown levels of 32 trees. Two branches were taken from each level at the two sampling dates.

The estimates in table 1 were computed as follows. The mean density was calculated, using

$$d_{Mt} = 1/(n\, n_t) \sum_{i,j} (Y_{tij} / X_{tij})$$

and the mean survival, using

$$s_M = 1/n \sum_{i} ( (Y_{2i.} / X_{2i.}) / (Y_{1i.} / X_{1i.}) ).$$

The ratio estimates of survival and density were got by

$$d_{Rt} = Y_{t..} / X_{t..} \text{ and}$$

$$s_R = \sum_{i} (Y_{2i.} X_{1i.}) / \sum_{i} (Y_{1i.} X_{2i.}).$$

Properties of the estimator $s_R$ are discussed in appendix I.

Density estimates were obtained from transformations of the form $\log (Y_{tij} / X_{tij} + k)$. One estimate, $d_{Lt}$, is the antilog of the sample mean of the transformed variables. The other estimate, $d_{Ft}$, is based on a method attributable to Finney (*1941*). This method gives unbiased estimates of the untransformed population mean if the transformed data are independent and normally distributed. Survival estimates based on the transformation

$$\log [(Y_{2i.} X_{1i.}) / (Y_{1i.} X_{2i.}) + k]$$

were also obtained using antilogs ($s_L$) and Finney's method ($s_F$).

The estimates in table 1 show a tendency to decrease, from left to right, in each row. The only exception is in the results for the lower crown, where $s_M$ and $s_R$ are in reverse order. The ordering implies that the estimates have different expectations or, in other words, that they are unbiased estimates of different population parameters. It is also evident that the constant, k, in the log-transformation has a pronounced effect on the estimates, especially at low densities. Since the population parame-

Table 1.—Estimates of density[1] and survival, and standard errors[2] for different estimators on the same data

| Item | Mean | Ratio | Logarithmic transformation | | | |
|---|---|---|---|---|---|---|
| | | | K = 1.0 | | K = .01 | |
| | | | F | L | F | L |
| | | LOWER CROWN | | | | |
| Density (1st date) | 2.555 | 2.541 | 1.988 | 1.926 | 1.281 | 1.049 |
| | ±.327 | ±.339 | ±.308 | ±.298 | ±.364 | ±.298 |
| Density (2nd date) | .311 | .267 | .215 | .205 | .029 | .022 |
| | ±.121 | ±.089 | ±.078 | ±.074 | ±.012 | ±.009 |
| Percent survival | 10.618 | 12.538 | 2.330 | 1.840 | .260 | .108 |
| | ±3.508 | ±4.056 | ±.862 | ±.681 | ±.105 | ±.044 |
| | | UPPER CROWN | | | | |
| Density (1st date) | 5.434 | 4.994 | 4.191 | 4.060 | 3.730 | 3.494 |
| | ±.784 | ±.718 | ±.587 | ±.569 | ±.658 | ±.616 |
| Density (2nd date) | .674 | .579 | .430 | .411 | .076 | .056 |
| | ±.203 | ±.169 | ±.118 | ±.113 | ±.027 | ±.020 |
| Percent survival | 19.526 | 12.891 | 4.554 | 3.652 | 1.217 | .498 |
| | ±7.515 | ±4.291 | ±1.501 | ±1.204 | ±.475 | ±.194 |

[1]Density is expressed in insects per 100 square inches of branch surface area.
[2]The standard error of estimate is shown below the estimate.

ters are unknown, it is impossible to draw any conclusion about the degree of bias of any of the estimators.

The standard errors for the mean and ratio estimates in table 1 were estimated by conventional formulas, and those for the ratio estimates by using the combined ratio formulae (*Cochran 1953*, section 6.11) treating each tree as a strata with the same (large) number of branches. The errors for the log-transformation estimates were estimated on the assumption that the transformed data are normally distributed, as indicated by Finney. The resulting estimates are not valid unless the normality assumption is true. The data were roughly "normal" for $k = 1.0$, but very skew for $k = .01$. Comparison of the coefficients of variation of the estimates gives no evidence that any particular estimator is superior by this criterion. However, because of the notoriously high variability of error estimates and the small number of data sets considered, no firm conclusion can be drawn from this comparison.

## THEORETICAL PROPERTIES OF ESTIMATORS

A more orderly procedure is to consider the conditions under which an estimator is optimal in some sense and to discard those estimators for which the optimum conditions are obviously not met by the data. For instance, the sample mean is an unbiased estimate of the population mean under random sampling. Now, the population means of $Y_{tij}/X_{tij}$ and $(Y_{2i.}/X_{2i.}) / (Y_{1i.}/X_{1i.})$ are not population density ($D_t$) and survival (S). Therefore, $d_{Mt}$ and $s_M$ are biased estimators of $D_t$ and S. A sample mean is a minimum variance estimator of the population mean if the sample mean and its variance are independent. This condition was not met by the data in table 1, so $d_{Mt}$ and $s_M$ are probably not minimum variance estimators.

From scatter diagrams, based on trees, the variances of both $Y_{tij}/X_{tij}$ and $(Y_{2i.}/X_{2i.}) / (Y_{1i.}/X_{1i.})$ appeared to be roughly proportional to the square of their means. In this case, a transformation of the form

$$\log (Z + k)$$

tends to induce independence between mean and variance on the transformed scale. k is needed to ensure the existence of the transform when $Z = 0$, and affects the resulting distribution. Anscombe [1949] discusses selection of values for k. If the transformation does result in independence of mean and variance on the transformed scale, the mean on the transformed scale would be the minimum-variance linear estimator of the population mean of the transformed variables. One would expect estimates on the untransformed scale which were based on the mean of the transformed data to be low variance estimators. The antilog of the mean of the log-transformed data estimates the geometric mean of $(Z+k)$. The estimates $d_{Lt}$ and $s_L$ were computed by subtracting k from the antilogs of the transformed means. Thus $d_{Lt}$ and $s_L$ should have reasonably small standard errors, but there is no reason why they should be unbiased estimates of $D_t$ and S.

Transformations that induce independence between mean and variance often tend to make the data more normally distributed. If the log-transformed data are in fact normal, Finney's method gives unbiased estimates of the population mean of the untransformed data. In this case, $d_{Ft}$ and $s_F$ have the same expectation as $d_{Mt}$ and $s_M$, so they are biased estimators of $D_t$ and S. The estimates $d_{Ft}$ and $s_F$ should have smaller variances than $d_{Mt}$ and $s_M$.

Scatter diagrams may not unambiguously indicate variance-mean relationships. For the data used in table 1, we could easily have argued that the variance of the mean for a tree was proportional to $\mu + \lambda^2 \mu^2$, where $\mu$ is the tree mean and $\lambda$ is a constant. Beall (*1942*) showed that this relation leads to the variance stabilization transformation $T(z) = \lambda^{-1} \sinh^{-1} (\lambda \sqrt{z})$ in the same way that the assumption that the variance is proportional to $\mu^2$ leads to the log-transformation we just considered. We therefore introduce two new estimators of $D_t$ and S, written $d_{tB}$ and $s_B$, based on the inverse of the transformation T applied to the mean of the T-transformed data. Neyman and Scott (1960) gave a method for obtaining unbiased estimators of

the mean on the untransformed scale when the transformed data are normally distributed. This method, which is analogous to Finney's method for the log transformation, led us to consider the estimators, $d_{Nt}$ and $s_N$, of $D_t$ and S. All the comments we have made about the estimators based on the log transformation are, by analogy, appropriate to the corresponding etsimators based on the T transformation.

The sample ratio, $d_{Rt} = Y_{t..}/X_{t..}$, is an unbiased estimator of $D_t$ if

$$Y_{tij} = D_t X_{tij} + e_{tij}$$

where the $e_{tij}$ are independent errors with zero expectations ($E(e_{tij}) = 0$) for any $X_{tij}$. We assume the $X_{tij}$ are known exactly and that $E(e_{tij}|X_{tij}) = 0$ within the sample. $d_{Rt}$ is an unbiased estimate of $D_t$ under weaker conditions, as shown in appendix I. Further, $d_{Rt}$ is a minimum variance estimator if $E(e^2_{tij}) = hX_{tij}$ for all $X_{tij}$ and some constant, h. The conditions for unbiasedness and minimum variance for ratio estimators are given in several texts, including Sukhatme (*1954, p. 154*) and Cochran (*1953, p. 123*). These conditions appeared to be satisfied by the test data.

In table 1, $s_R = (\Sigma Y_{2i.}X_{1i.}) / (\Sigma Y_{1i.}X_{2i.})$ was used to estimate survival. It should be noted that this is not the estimator $(Y_{2..}X_{1..}) / (Y_{1..}X_{2..})$. In appendix I, conditions under which $s_R$ is an unbiased estimate of the population survival, S, are shown. Roughly, the assumptions include:

- The conditions for $d_{Rt}$ to be an unbiased estimator hold for both sampling periods 1 and 2.
- The $X_{ti.}$ are independent random variables within any sampling period. (Note that for unbiased estimates of density, only the error terms were required to be independent. It need not even be assumed that branch surface area is a random variable.)
- Various assumptions about the independence of the random variables.

The mathematical conditions for $s_R$ to be minimum variance can be established. However, unlike the conditions for unbiasedness given above, they are too complex to be subject to biological interpretation.

Tin (1965) gave an adjustment to ratio estimators that reduces bias when the conditions for unbiasedness do not hold. The adjustment applied to the density estimator may, for infinite populations, be written

$$d_{At} = \frac{Y_{t..}}{X_{t..}}\left[1 - \frac{n}{n-1}\left\{\frac{\Sigma X^2_{tij}}{(\Sigma X_{tij})^2} - \frac{\Sigma X_{tij}Y_{tij}}{(\Sigma X_{tij})\,(\Sigma Y_{tij})}\right\}\right]$$

The price paid for the reduction in bias is a higher variance, although generally not much higher than for $d_{Rt}$. $d_{At}$ was not included in table 1, but will be considered later.

Using a modification of Tin's approach, we obtained correction term for the "within–tree" bias for the estimator, $s_R$. For the $i^{th}$ tree, the resulting correction factor is

$$C_i = \left[1 - \frac{n_1}{n_1-1}\left\{\frac{\sum_j Y^2_{1ij}}{(Y_{1i.})^2} - \frac{\sum_j X_{1ij}Y_{1ij}}{(X_{1i.})\,(Y_{1i.})}\right\}\right] \times \left[1 - \frac{n_2}{n_2-1}\left\{\frac{\sum_j X^2_{2ij}}{(X_{2i.})^2} - \frac{\sum_j X_{2ij}Y_{2ij}}{(X_{2i.})\,(Y_{2i.})}\right\}\right]$$

If the variance of the correction term is independent of the random variables on the right side of the equation, the "corrected" estimate of S would best be got by $s_R(C./N)$. This condition was well satisfied by our test data.

In considering which of the estimators just discussed should be used, the ratio estimators appear to have the edge. Under certain conditions, the other estimators give unbiased estimates of population parameters other than density and survival as we have defined them. The ratio estimators are unbiased estimators of density and survival under ideal conditions. These conditions are fairly restrictive, so bias correction might be needed.

## INVESTIGATIONS BASED ON SIMULATION

In practice, one would not expect specific optimality conditions required for an estimator to hold precisely. Therefore, we have used Monte Carlo simulation in order to study the behavior of the estimators over a range of con-

ditions likely to be encountered in applications. To construct a simulator that generates realistic data, we considered diverse sets of budworm data and established characteristics which these sets had in common.

We concluded, in agreement with Morris *(1955)*, that the number of insects on a branch closely follows the negative binomial frequency distribution, regardless of the intensity of the populations we observed. Further, by plotting, we found that

$$\frac{Y_{tij} - D_t X_{tij}}{\sqrt{X_{tij}}}$$

was distributed roughly normally for these populations. These conditions were used to define the joint distribution of insects and branch surface area, given the means, variances, and covariance of these variables at one sampling date (appendix II).

A computer program was written that generated simulated insect counts and branch surface areas. The computer was supplied the means, variances, and covariance of X and Y at each sampling date. To study the properties of the estimators on an individual tree basis, the same characteristics were used for 500 independent tree samplings with 4 branches sampled at the first period, 8 at the second.

The results of these simulations for the ratio estimators are given in table 2. From this table we concluded that the ratio is serviceable as an estimator of population density. For survival estimates, the bias-corrected ratio estimator, $s_A$, should be used.

The mean and uncorrected transform based estimates were all high. Both bias and variance were reduced by corrections to these estimates, but the corrections were too great for low density or low survival rates, and too small for high density or survival. The square

Table 2.—The tree mean, standard error, and bias of the ratio estimators, estimated by simulation

| Case[1] | True parameter | Average estimate | Standard error of tree estimate | Estimate of bias | t-test for bias = 0 |
|---|---|---|---|---|---|
| DENSITY IN INSECTS PER 1,000 SQUARE INCHES OF FOLIAGE | | | | | |
| 1 | 44.064 | [2]46.288 | 16.422 | 2.224 | 3.03 |
| | | 44.762 | 14.450 | .698 | 1.08 |
| 2 | 33.256 | 33.405 | 12.661 | .149 | .26 |
| | | 32.633 | 11.905 | —.623 | —1.17 |
| 3 | 28.091 | 28.932 | 26.593 | .840 | .72 |
| | | 24.627 | 27.521 | —3.464 | —2.82 |
| 4 | 22.485 | 24.543 | 23.850 | 2.058 | 1.93 |
| | | 18.049 | 41.808 | —4.436 | —2.37 |
| 5 | 11.877 | 12.005 | 8.074 | .127 | .35 |
| | | 11.634 | 7.513 | —.243 | —.72 |
| 6 | 5.706 | 5.635 | 3.632 | —.071 | —.44 |
| | | 5.537 | 3.494 | —.170 | —1.09 |
| 7 | 2.480 | 2.476 | 2.114 | —.403 | —.04 |
| | | 2.433 | 2.036 | —.047 | —.52 |
| 8 | 1.363 | 1.342 | 1.458 | —.021 | —.32 |
| | | 1.319 | 1.409 | —.044 | —.69 |
| PERCENT SURVIVAL | | | | | |
| 1-8 | 3.093 | 3.254 | 3.833 | 0.161 | 0.04 |
| | | 2.915 | 3.466 | —.178 | —1.15 |
| 2-7 | 7.458 | 8.653 | 8.671 | 1.195 | 3.08 |
| | | 7.319 | 7.011 | —.139 | —.44 |
| 3-6 | 20.314 | 45.170 | 77.837 | 24.856 | 7.15 |
| | | 17.420 | 24.068 | —2.894 | —2.69 |
| 4-5 | 52.822 | 116.787 | 210.391 | 63.965 | 6.80 |
| | | 49.715 | 82.832 | —3.107 | —.84 |

[1]Cases 1 to 4 are based on 4 branch samples per tree, cases 5 to 8 on 8 branch samples. 500 trees were sampled for each case.

[2]For each case, the first row is not corrected for bias, while the second row is.

of the coefficient of variation of all but the mean estimators is roughly in inverse proportion to the sample size. Thus a measure of efficiency, which approximates the ratio of sample sizes required for equal coefficients of variation, is the square of the ratio of coefficients of variation. With this measure of efficiency, one of the ratio estimates is always at least 60 percent more efficient than any transform based estimator of the same parameter. We have given detailed results only on the ratio estimators and their vaiances, since none of the other estimators are serious competitors.

However, data are generally transformed to induce independence of mean and variance and, one hopes, to make the data more normally distributed, in order to satisfy assumptions required in analysis of variance. Two rough measures of normality are skew ($\gamma_1$) and kurtosis ($\gamma_2$). Both are zero for normal distributions. They are defined, for a random variable Z, by

$$\gamma_1 = E(Z\text{-}E(Z))^3 / [E(Z\text{-}E(Z))^2]^{3/2} \text{ and}$$

$$\gamma_2 = E(Z\text{-}E(Z))^4 / [E(Z\text{-}E(Z))^2] - 3.$$

Table 3 gives estimates of the skew and kurtosis of the means of the transformed data on each tree.

The skew and kurtosis are of an order such that the transformed variables would ordinarily be considered normally distributed in practice. Scheffé *(1959, p. 333)* gives values of $\gamma_1$ and $\gamma_2$ typically occurring in applications. From table 3, there is no clear difference between the log and Beall's transformation. Beall's transformation appears to give the better results, but this might be because the constant $\lambda$ in the transformation was determined exactly from the parameters of the simulator. We did not check the sensitivity of the transformation to the choice of $\lambda$. The constant K in the log transformation was estimated from all 500 trees for each case, and so is much more adequately estimated than it would ordinarily be. Where 1 was added before taking logs, the results are still quite good, indicating the choice of a constant for the log transformation is not critical if reasonable care is taken to ensure normality.

It should be noted that the simulator was not constructed so that any of the above transformations would induce normality. However, these transformations have been known to do so with similar insect data *(Morris 1955)*. The fact that the simulator reproduced this distribution property would certainly support the realism of the assumptions that were used in constructing the simulator.

In any interpretation of the simulation results, one should recognize that all of the 500 trees sampled in a simulation case had identically distributed insects and branch surface areas. Equivalently, each simulation case represents 500 independent samplings of the same tree. The results of the simulation indicate that the ratio estimators were very satisfactory when individual trees were treated as the population sampled. We cannot, however, on the basis of the simulation alone, extend the results of the simulation to groups of trees where the distribution of insects and trees are

**Table 3.—Estimates of skew and kurtosis of simulated data after transformation**

| Case[1] | Skewness | | | Kurtosis | | |
|---|---|---|---|---|---|---|
| | log(Y/X+1) | log(Y/X+K)[2] | Beall | log(Y/X+1) | log(Y/X+K)[2] | Beall |
| 1 | 0.138 | 0.569 | 0.438 | 0.941 | 1.423 | 1.310 |
| 2 | —.636 | .037 | —.327 | .946 | .574 | .783 |
| 3 | .088 | .020 | —.073 | —.317 | —.314 | —.336 |
| 4 | .067 | .066 | —.175 | —.175 | —.176 | —.210 |
| 5 | .286 | .279 | .169 | .258 | .249 | .130 |
| 6 | .510 | .555 | .436 | .583 | .724 | .380 |
| 7 | .667 | .650 | .595 | .342 | .290 | .149 |
| 8 | .755 | .700 | .667 | .111 | .002 | —.490 |

[1]For each case, the same data are used as for the corresponding case in Table 2.
[2]K was estimated from the simulated data.

not identical, since the behavior of the various estimators will be affected by the differences between trees.

## SUMMARY AND CONCLUSIONS

Our objective was to find estimators of well-defined population parameters for use on specific data. The behaviors of selected estimators were compared in three steps.

First, estimates were computed on sample data. From this step we concluded that there were important differences between estimators, but we were unable to decide which estimators were better in any sense.

In the second step, we checked to see what optimality criteria might be roughly satisfied by the data. The results concerning variance stabilizing transformations were inconclusive. However, the data did conform reasonably well to the conditions for unbiasedness and minimum variance for the ratio estimators.

In the third step, simulation was used to appraise the influence of departures from ideal conditions on the ratio estimators. Bias, which was our major concern, is generally most apparent in small samples for ratio estimators. We therefore sampled a single tree many times, using the simulator. The distributional characteristics of the simulated data were based on the corresponding characteristics of actual data. In addition to information on the behavior of the ratio estimators under diverse conditions, the simulator gave us insight on the effectiveness of variance stabilization transformations in "normalizing" the data, and the sensitivity of these transformations to the choice of constants.

Variations on the steps given in the preceding paragraphs would be applicable to the problem of selecting an estimator of any population parameter from an adequate data base. For the specific problem we considered, that of estimating insect density ($D_t$) and survival (S), we came to the following conclusions.

The ratio estimators ($d_{Rt}$, $d_{At}$, $s_R$ and $s_A$) were superior to the other estimators considered in terms of bias, as established by individual tree simulation. In these simulations, the ratio estimators were always at least 60 percent more efficient than the other estimators. For estimating survival, there is a slight advantage in using the bias-adjusted estimator, $s_A$. For estimating density, the unadjusted ratio estimator, $d_{Rt}$, seems adequate. These results from simulation, plus the correspondence of observed data to the conditions required for optimality of the ratio estimators, leads us to conclude that $d_{Rt}$ and $s_R$ should be used to estimate $D_t$ and S.

Estimates of skew and kurtosis of the ratio estimators based on simulation indicate these estimators are far from normally distributed on an individual tree basis. Therefore, for hypothesis testing within an analysis of variance framework, we recommend using either log or $\sinh^{-1}$ transformations to "normalize" the data. The skew and kurtosis of either transformation was fairly insensitive to small changes in the constants involved in the transformations. Moderate care in setting these constants should suffice in practice.

## REFERENCES


Anscombe, F. J.
1949. THE STATISTICAL ANALYSIS OF INSECT COUNTS BASED ON THE NEGATIVE BINOMIAL DISTRIBUTION. Biometrics 5:165.

Beall, G.
1942. THE TRANSFORMATION OF DATA FROM ENTOMOLOGICAL FIELD EXPERIMENTS SO THAT THE ANALYSIS OF VARIANCE BECOMES APPLICABLE. Biometrika 32:243.

Cochran, W. G.
1953. SAMPLING TECHNIQUES. John Wiley & Sons, Inc., New York.

Finney, D. J.
1941. ON THE DISTRIBUTION OF A VARIATE WHOSE LOGARITHM IS NORMALLY DISTRIBUTED. J. Royal Stat. Soc. Supp. Ser. B 7:155.

Morris, R. F.
1955. THE DEVELOPMENT OF SAMPLING TECHNIQUES FOR FOREST INSECT DEFOLIATORS, WITH PARTICULAR REFERENCE TO THE SPRUCE BUDWORM. Can. J. Zool. 33:225.

Neyman, J., and Scott, Elizabeth L.
1960. CORRECTIONS FOR BIAS INTRODUCED BY A TRANSFORMATION OF VARIABLES. Ann. Math. Stat. 31:643.

Sukhatme, P. V.
1954. SAMPLING THEORY OF SURVEYS WITH APPLICATIONS. Iowa State College Press, Ames, Iowa.

Scheffé, H.
1959. THE ANALYSIS OF VARIANCE. John Wiley & Sons, Inc., New York.

Tin, M.
1965. COMPARISON OF SOME RATIO ESTIMATORS. J. Amer. Stat. Assoc. 60:309.

Williams, C. B., and Walton, G. S.
1968. EFFECTS OF ZETRAN ON THE BUDWORM, *Choristoneura occidentalis* AND ASSOCIATED INSECTS IN MONTANA. J. Econ. Entomol. 61:784.

# APPENDIX I

## Conditions for Unbiasedness of the Double Ratio Estimator

The object of this appendix is to present a set of conditions under which the estimator $s_R$ yields unbiased estimates of S. The conditions imposed cannot be expected to hold exactly in practice. The utility of this demonstration will be as a guide to conditions that must be approximately satisfied if $s_R$ is to be trusted as an estimate of S. And $s_R$ may be unbiased under other, and perhaps less stringent, conditions.

The conditions are labeled with roman numerals for easy reference. The conditions are stated as they are used in the demonstration partly because this makes it easier to see how the demonstration turns on the assumptions, partly because the conditions are more comprehensible at that point in the development, and partly to emphasize that the conditions are imposed in order to procede with the demonstration.

To begin, we write the number of insects on the $j^{th}$ branch of the $i^{th}$ tree at time t as

$$Y_{tij}=D_tX_{tij}-e_{tij}$$

If the sampling of the branches on the $i^{th}$ tree at time t is such that $E(e_{tij})=0$ then, summing over the branch samples, we have

$$Y_{ti.}=D_tX_{ti.}+e_{ti.} \text{ and } E(e_{ti.})=0.$$

If the N trees are sampled independently, the $e_{ti.}$ will be independent between trees for fixed t. With this motivation we state the following conditions, which are sufficient for $d_{tR}$ to be an unbiased estimator of $D_t$, namely:

(I) $Y_{ti.}=D_tX_{ti.}+e_{ti.}, E(e_{ti.})=0$ and $e_{ti.}$ is independent of $e_{tj.}$ for $i \neq j$.

The estimator $s_R$ has been defined as

(2) $s_R=(\Sigma Y_{2i.}X_{1i.})/\Sigma(Y_{1i.}X_{2i.})=\Sigma Y_i^*/\Sigma X_i^*$

where

(3) $Y_i^*=Y_{2i.}X_{1i.}$ and $X_i^*=Y_{1i.}X_{2i.}$

Define the population parameter $S^*$ by

(4) $S^*=(\Sigma Y_{2i.}X_{1i.})/(\Sigma Y_{1i.}X_{2i.})$

and $e_i^*$ by the relation

(5) $Y_i^*=S^*X_i^*+e_i^*$

We will first establish conditions under which $E(e_i^*)=0$ and the $e_i^*$ are independent, which is tantamount to showing $s_R$ is an unbiased estimate of $S^*$. Then we will establish conditions under which $S^*=S$, where

(6) $S=(\Sigma Y_{2i.})(\Sigma X_{1i.})/(\Sigma Y_{1i.})(\Sigma X_{2i.})=D_2/D_1$

Using (I), (2), (3), (4) and (5) we may write

$$e_i^*=Y_i^*-S^*X_i^*=Y_{2i.}X_{1i.}-(D_2/D_1)\,Y_{1i.}X_{2i.}$$

$$=(D_2X_{2i.}+e_{2i.})X_{1i.}-(D_2/D_1)(D_1X_{1i.}+e_{1i.})X_{2i.}$$

$$=e_{2i.}X_{1i.}-(D_2/D_1)e_{1i.}X_{2i.} \text{ or}$$

(7) $e_i^*=(D_1e_{2i.}X_{1i.}-D_2e_{1i.}X_{2i.})/D_1$

From (7),

$$(8)\quad E(e_i^*) = (D_1E(e_{2i}.X_{1i}.) - D_2E(e_{1i}.X_{2i}.)\ ) / D_1$$

We impose the condition that the "tree error" at one sampling date is not correlated with the total branch surface area from the same tree at the other sampling date. Formally,

$$(IX)\quad E(e_{ti}.X_{si}.) = E(e_{ti}.)\ E(X_{si}.) \text{ for } t \neq s.$$

Then since $E(e_{ti}.)=0$ from (I), $E(e_i^*)=0$, which is what we set out to show.

We will now establish additional conditions so $E(e_i^*\ e_j^*) = E(e_i^*)\ E(e_j^*)$ for $i \neq j$. From (7),

$$\begin{aligned} E(e_i^*\ e_j^*) = E\ \Big[ & (D_1X_{1i}.e_{2i}. - D_2X_{2i}.e_{1i}.)\ /D_1 \\ & (D_1X_{1j}.e_{2j}. - D_2X_{2j}.e_{1j}.)\ /D_1 \Big] \\ = E\ \Big[ & D_1^2X_{1i}.e_{2i}.X_{1j}.e_{2j}. \\ & -D_1D_2X_{1i}.e_{2i}.X_{2j}.e_{1j}. \\ & -D_1D_2X_{2i}.e_{1i}.X_{1j}.e_{2j}. \\ & +D_2^2X_{2i}.e_{1i}.X_{2j}.e_{1j}.\Big] / D_1^2 \end{aligned}$$

Now, from (8),

$$\begin{aligned} E(e_i^*)E(e_j^*) = \ & D_1^2E(X_{1i}.e_{2i}.)\ E\ (X_{1j}.e_{2j}.) \\ & -D_1D_2E(X_{1i}.e_{2i}.)\ E\ (X_{2j}.e_{1j}.) \\ & -D_1D_2E(X_{2i}.e_{1i}.)\ E\ (X_{1j}.e_{2j}.) \\ & +D_2^2E(X_{2i}.e_{1i}.)\ E\ (X_{2j}.e_{1j}.)\ /D_1^2 \end{aligned}$$

It is apparent that the last two equations are equivalent if

$$(X)\quad E(X_{1i}.e_{2i}.X_{1j}.e_{2j}.) = E(X_{1i}.e_{2i}.)E(X_{1j}.e_{2j}.)$$

$$E(X_{1i}.e_{2i}.X_{2j}.e_{1j}.) = E(X_{1i}.e_{2i}.)E(X_{2j}.e_{1j}.)$$

$$E(X_{2i}.e_{1i}.X_{1j}.e_{2j}.) = E(X_{2i}.e_{1i}.)E(X_{1j}.e_{2j}.)$$

$$E(X_{2i}.e_{1i}.X_{2j}.e_{1j}.) = E(X_{2i}.e_{1i}.)E(X_{2j}.e_{1j}.)$$

for $i \neq j$.

In words, the above assumption is that the product of the branch surface area for a tree at one sampling date and "tree error" at the other sampling date are uncorrelated between different trees, and it allowed us to show $E(e_i^*\ e_j) = E(e_i^*)\ E(e_j^*)$ for $i \neq j$. That is, $e_i^*$ and $e_j^*$ are uncorrelated.

Showing two variables are independent is sufficient to show they are uncorrelated. Showing two variables are uncorrelated is not sufficient to prove they are independent although it certainly makes independence a more palatable assumption. For $s_R$ to be an unbiased estimator of $S^*$, we need independence, and therefore must make independence an assumption:

(XI) $e_i^*$ and $e_j^*$ are independent for $i \neq j$.

We still have to justify our claim S and $S^*$ are the same population parameter, where $S^* = (\Sigma\ Y_{2i}.X_{1i}.)\ /\ (\Sigma\ Y_{1i}.X_{2i}.)$ and S is as shown in (6).

Up to this point in this appendix we have made no conditions on the number of trees or branches.

Let there be a finite population of, say, $i = 1, M$ trees such that $Y_{ti.} = D_t X_{ti.} + e_{ti.}$ Write

$$S_M^* = \left(\sum_{i=1}^{M} Y_{2i.} X_{1i.}\right) / \left(\sum_{i=1}^{M} Y_{1i.} X_{2i.}\right)$$

$$= \frac{D_2 \Sigma (X_{2i.} + e_{2i.}/D_2)/X_{1i.}}{D_1 \Sigma (X_{1i.} + e_{1i.}/D_1)/X_{2i.}}$$

$$= \frac{D_2}{D_1} \left[\frac{(\Sigma X_{1i.} X_{2i.}) + (\Sigma X_{1i.} e_{2i.}/D_2)}{(\Sigma X_{1i.} X_{2i.}) + (\Sigma X_{2i.} e_{1i.}/D_1)}\right]$$

$$(12) \quad S_M^* = \frac{D_2}{D_1} \left[1 + \frac{(\Sigma e_{2i.} X_{1i.}/D_2)/M - (\Sigma e_{1i.} X_{2i.}/D_1)/M}{(\Sigma X_{1i.} X_{2i.})/M + (\Sigma X_{2i.} e_{1i.}/D_2)/M}\right]$$

We write $\mu_M(Z_i)$ for $\Sigma Z_i/M$, since this is the mean of the M- population. We assume that the M-population can be augmented indefinitely so that in the limit as $M \to \infty$, $\mu_M(e_{ti.}) = 0$ for all $X_{ti.}$ as implied by (I), and $\mu_M(X_{ti.} e_{sj.}) = \mu_M(X_{ti.})\,\mu_M(e_{sj.})$ for $i \neq j$ and $s \neq t$, as implied by (3). Since $X_{1i.}$ and $X_{2i.} > 0$, we have, from (12),

$$\lim_{M \to \infty} S_M^* = S^* = \frac{D_2}{D_1} = S.$$

It should be proven that such a limiting process exists for any infinite population meeting conditions [I], [IX], and (XI). The proof is beyond the scope of the authors.

# APPENDIX II

## Discussion of the Simulation Model Used in Generating Insect Counts and Branch Surface Areas

We start with the assumption that the insect counts (Y) follow the negative binomial distribution, and that the population means, variances, and covariance for the branch surface area (X) and insect counts are known. Because the mean and variance of a negative binomial variable completely determine its distribution, the $Y_i$ are easily generated using a psuedo-random number generator for the rectangular distribution.

We also want $(Y_i - dX_i) / \sqrt{X_i}$ to be roughly normally distributed. Since $Y_i$ is determined by its mean and variance, and $X_i$ appears twice, it is inconvenient to have the statistic exactly normally distributed. However, because $Y_i \doteq dX_i$, we chose to consider the statistic $(X_i - CY_i) / \sqrt{Y_i}$ in place of the former. This statistic is not entirely satisfactory either, because $P[Y_i = 0] \neq 0$.
Consequently, we add the $R(0,1)$ variable $r_i$ to $Y_i$. But, branch surface area may never be negative or zero so a constant K is introduced to reduce $P[X_i \leq 0]$ to nearly zero for all $Y_i$.

The final model selected is

$$(1)\quad [\,(X_i - K) - C(Y_i + r_i)\,] \,/\, \sqrt{\sigma(Y_i + r_i)} = e_i$$

where $e_i \sim N(0,1)$, $r_i \sim R(0,1)$ and $Y_i \sim$ N.B.$(\mu_Y, V_Y)$ and all three variables are independently distributed.

Once $e_i, r_i$ and $Y_i$ are generated by the pseudo-random number generator it can be seen from (1) that $X_i$ is determined by the equation

$$(2)\quad X_i = K + C(Y_i + r_i) + \sigma e_i (Y_i + r_i)^{1/2}.$$

We must now demonstrate that the constants K, $\sigma$, and C are uniquely determined by the population means ($\mu_y$ and $\mu_x$) variances ($V_y$ and $V_x$) and covariance Cov(X,Y).
From (2),

$$(3)\quad \mu_x = E(X_i) = K + C(\mu_y + 1/2) \text{ and}$$

$$\begin{aligned} E(X_i^2) = E\Big[ & K^2 + \sigma^2 e_i^2 (Y + r_i) + C^2(Y_i^2 + 2Y_i r_i + r_i^2) \\ & + 2KC(Y_i + r_i) + \sigma e_i (Y_i + r_i)^{1/2} (K + C(Y_i + r_i)\Big] \\ = {} & K^2 + \sigma^2(\mu_y + 1/2) + C^2(V_y + \mu_y^2 + \mu_y + 1/3) \\ & + 2KC(\mu_y + 1/2) \end{aligned}$$

so, $V(X) = E(X_i^2) - E^2(X_i)$

$$(4)\quad V(X) = \sigma^2 (\mu_y + 1/2) + C^2(V_y + 1/12)$$

For Cov(X,Y) we compute, using (2),

$$\begin{aligned} E(X_i, Y_i) &= E\Big[KY_i + C(Y_i^2 + Y_i r_i) + \sigma e_i (Y_i + r_i)^{1/2} Y_i\Big] \\ &= K\mu_y + C(V_y + \mu_y^2 + \mu_y/2), \end{aligned}$$

since $E(e_i) = 0$ independently of $Y_i$ and $r_i$.

But then

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E(X_1Y_1) - E(X_1)\,E(Y_1)\\ &= K\mu_y + C(V_y + \mu_{\bar{y}}^2 + \mu_y/2)\\ &\quad - K\mu_y - C(\mu_{\bar{y}}^2 + \mu_y/2)\end{aligned}$$

$$(5)\quad \mathrm{Cov}(X,Y) = CV_y$$

It can easily be seen that the constants C, $\sigma$ and K are uniquely determined. C can be got from (5). Given C, $\sigma$ is available from (4), and then K can be found using (3).

# APPENDIX III

## Summary of Estimators

1. The "mean" estimators were defined by

$$d_{Mt} = 1/(nn_t) \sum_{i,j} (Y_{tij}/X_{tij}) \text{ and}$$

$$s_M = 1/n \sum \left[ (Y_{2i.}/X_{2i.})/(Y_{1i}/X_{2i.}) \right]$$

2. The simple ratio estimators are

$$d_{Rt} = Y_{t..}/X_{t..} \text{ and}$$

$$s_R = \left[ \sum_i (Y_{2i.}X_{1i.}) \right] / \left[ \sum_i (Y_{1i.}X_{2i.}) \right]$$

3. Adjusting the simple ratio estimate for bias leads to the adjusted ratio estimates

$$d_{At} = d_{Rt} \left[ 1 - \frac{n}{n-1} \left\{ \frac{\Sigma X_{tij}^2}{(\Sigma X_{tij})^2} - \frac{\Sigma X_{tij}Y_{tij}}{(\Sigma X_{tij})(\Sigma Y_{tij})} \right\} \right]$$

$$\text{and } s_A = s_R \left[ 1 - \frac{n_1}{n_1-1} \left\{ \frac{\sum_j Y_{1ij}^2}{(Y_{1i.})^2} - \frac{\sum_j X_{1ij}Y_{1ij}}{(X_{1i.}Y_{1i.})} \right\} \right] \times$$

$$\left[ 1 - \frac{n_2}{n_2-1} \left\{ \frac{\sum_j X_{2ij}^2}{(X_{2i.})^2} - \frac{\sum_j X_{2ij}Y_{2ij}}{(X_{2i.}Y_{2i.})} \right\} \right]$$

The equation for $s_A$ is appropriate in the above form for a single tree. See the discussion for clarification.

4. Earlier, estimators based on log-tranformations were introduced.

Let

$$\mu_{tij} = \log \left[ Y_{tij}/X_{tij} + K \right]$$

Then,

$$d_{Lt} = \exp \left[ \mu_{t..}/nn_t \right] - K$$

Also, let

$$v_i = \log \left[ (Y_{2i.}/X_{2i.})/(Y_{1i.}/X_{1i.}) + K \right] .$$

Then,

$$s_L = \exp \left[ v./n \right] - K$$

5. We presented estimators based on Finney's method of correcting for bias in estimators of the mean of log-normal populations.

Let

$$\bar{U}_t = U_{t.}/nn_t \text{ and}$$

$$\hat{\sigma}^2_{\bar{U}t} = {}_1\Sigma_j (U_{tij} - \bar{U}_t)^2 / \left[(nn_t - 1)nn_t\right] .$$

Define

$$g(t) = 1 + \frac{n-1}{n} t + \frac{(n-1)^3 t^2}{2!n^2(n+1)} + \frac{(n-1)^5 t^3}{3!n^3(n+1)(n+2)} + \cdots$$

Then,

$$d_{Ft} = \exp(\bar{U}_t)\, g(\tfrac{1}{2} \hat{\sigma}^2_{\bar{U}t}) - K.$$

Similarly, let

$$\bar{V} = V./n \text{ and } \hat{\sigma}^2_{\bar{V}} = \Sigma(\bar{V} - V_i)^2 / \left[n(n-1)\right] .$$

Then

$$s_F = \exp(\bar{V})\, g(\tfrac{1}{2} \hat{\sigma}^2_{\bar{V}})$$

6. Beall [1942] studied a transformation which we adapted for estimation.

Let

$$P_{tij} = \lambda^{-1} \sinh^{-1} \lambda\left(\sqrt{Y_{tij}/X_{tij}}\right)$$

Then,

$$d_{Bt} = \lambda^{-1} \sinh^2(\lambda P_{t..}/nn_t).$$

Let

$$q_i = \lambda^{-1} \sinh^{-1} \lambda\left(\sqrt{(Y_{2i.}/X_{2i.})\ (Y_{1i.}/X_{2i.})}\right)$$

Then,

$$s_B = \lambda^{-1} \sinh^2(\lambda q./n)$$

7. Neyman and Scott (1960) extended Finney's method for bias correction to other transformations. We introduced a "bias corrected" estimator based on this work for Beall's transformation. Let

$$\bar{P}_t = P_{t..}/nn_t \text{ and } \hat{\sigma}^2_{\bar{P}t} = \frac{\Sigma\Sigma(\bar{P}_t - P_{tij})^2}{nn_t(nn_t - 1)} .$$

$$d_{Bt} = \frac{1}{2\lambda^2} \left[\cosh(2\lambda \bar{P}_t)\, g(2\lambda^2 \hat{\sigma}^2_{\bar{P}t}) - 1\right],$$

where $g(t)$ is given in part 5.

Also, let $\bar{q} = q./n$, $\hat{\sigma}^2_{\bar{q}} = \dfrac{\Sigma(\bar{q} - q_i)^2}{n(n-1)}$

Then

$$s_B = \frac{1}{2\lambda^2} \left[\cosh(2\lambda \bar{q})\, g(2\lambda^2 \hat{\sigma}^2_{\bar{q}}) - 1\right] .$$


THE FOREST SERVICE of the U. S. Department of Agriculture is dedicated to the principle of multiple use management of the Nation's forest resources for sustained yields of wood, water, forage, wildlife, and recreation. Through forestry research, cooperation with the States and private forest owners, and management of the National Forests and National Grasslands, it strives — as directed by Congress — to provide increasingly greater service to a growing Nation.