SEQUENTIAL SAMPLING FOR PEST MANAGEMENT

by

George Fohner

May 1981 No. 81-12

Sequential Sampling for Pest Management

George Fohner[1/]

Introduction

With traditional methods of sampling and decision making, the number of observations to be made is predetermined prior to taking the sample. Sequential sampling is an alternative method in which the number of observations depends on the values that are observed as sampling proceeds. If the first observations strongly favor a particular conclusion then a decision is made without further sampling. When the purpose of sampling insect populations is to distinguish between population densities that warrant treatment and those that do not, sequential sampling has important advantages compared to sampling based on fixed sample sizes. The decision rule for drawing conclusions using sequential sampling applies to samples of all sizes rather than to only samples of a single fixed size. Consequently, decisions based on sequential sampling can be made with equal reliability and, on the average, with smaller samples than decisions based on fixed sample sizes. The purpose of this paper is to discuss sequential sampling in an example related to the Colorado Potato Beetle in potatoes. However, principles are applicable for other insects on other crops.

Sampling and Pest Management Decisions

As part of the New York State potato pest management program, fields are sampled for Colorado Potato Beetle to decide whether population density

---

1/ Graduate Research Assistant, Department of Agricultural Economics, Cornell University, Ithaca, NY 14853. Support for this research was provided by the U.S.D.A. under Cooperative Agreement Number 58-319V-9-2702.

exceeds two larvae per plant, the action threshold for insecticide treatment. Since the number of beetles varies from plant to plant, the average density on sampled plants will rarely equal the true mean population density in the field. Consequently, an estimate of population density based on the sample average may result in an erroneous conclusion about whether the true population density exceeds the treatment threshold and accordingly may result in inappropriate recommendations. The frequency of erroneous conclusions can be estimated using sampling theory and knowledge about the frequency distribution of beetles per plant. The negative binomial distribution has been shown to be a useful representation of the frequency distribution for counts of Colorado Potato Beetle and other insects that tend to have a "clumped" distribution (Anscombe, 1949; Harcourt, 1963). The negative binomial distribution has two parameters. Described in terms of insect counts, these parameters are (1) the mean population density, which is the average number of insects per plant, and (2) the index of aggregation, which reflects the degree to which the insects are spatially clumped. Once values of the parameters have been assigned or estimated, the negative binomial distribution can be used to calculate the probability that a sample will contain a particular number of insects. The probability of obtaining sample averages that result in erroneous conclusions can therefore be estimated for any particular true population density and index of aggregation.

For example, assume that the index of aggregation is 0.5 and the sample size is 30 plants, which was the sample size used in the potato pest management program in 1980. Also, assume that the true population density is 2.5 larvae per plant, a "high" density for which spraying should be recommended. In this situation, with a fixed sample size of 30 plants, sampling would result in an inappropriate recommendation not to spray

approximately 23% of the time (Appendix 1). With true densities nearer to the 2.0 threshold, the probability of obtaining a misleading sample average is higher, the closer to 2.0 the higher the probability of "error". Unless sample size is very large, true densities near 2.0 will have a high probability of producing sample averages on the "wrong" side of the 2.0 threshold. Viewed from another perspective, a large sample size will be necessary for distinguishing confidently between true densities 1.9 and 2.1 on the basis of sample averages. Fortunately, such distinctions between densities near 2.0 have little if any biological significance; the 2.0 threshold value is a breakpoint for decision making but is not a point of discontinuity in the effect of the beetle on the potato crop. Conceptually, then, it is useful to classify true population densities into three categories: (1) densities significantly below 2.0, (2) densities near 2.0, and (3) densities significantly above 2.0. In making recommendations, correct identification of categories (1) and (3) is important, while for category (2) it is less important that sample averages accurately indicate whether true density is above or below 2.0. The essence of designing a sampling program is to specify the boundaries between these three categories and the required probability of correct identification.

Suppose that when the true population density of Colorado Potato Beetle is less than 1.5 we want to correctly classify it as a low density at least 90% of the time, and when the true density is greater than 2.5 we want to correctly classify it as a high density at least 95% of the time. For other true densities, those between 1.5 and 2.5, we are less concerned about whether they are classified as high or low and, accordingly, the probability of a correct classification can be less than the 90% and 95% standards imposed for densities below 1.5 and above 2.5. How large a sample is

required to operate with these categories and probabilities of correct classification? Assuming a population having a negative binomial distribution and an index of aggregation of 0.5, a sample of 91 plants is required using traditional sampling with a fixed sample size (Appendix 2, and Table 1).

The same low and high categories and standards of accuracy can be adopted for sequential sampling. Again, the emphasis is on having a high probability of correctly identifying high and low densities while being less concerned about whether intermediate densities are classified as high or low. Since the sampling must result in either a recommendation to spray or not to spray, sampling continues until the population density is classified as high or low. If the sample average exceeds a predetermined upper limit the population density is classified as high. If it drops below a predetermined lower limit the population density is classified as low. With the categories and standards described above, the predetermined limits are chosen so that true density of 1.5 would have no more than a 10% chance of being misclassified and a true density of 2.5 would have no more than a 5% chance of being misclassified. Using terminology applied to statistical hypothesis testing, the .10 probability of misclassifying the low density is called the "α level" and the .05 probability of misclassifying the high density is called the "β level" (Wald, 1947). With these predetermined limits, true densities less than 1.5 would have less than a 10% chance of being wrongly classified as high. True densities greater than 2.5 have less than a 5% chance of being wrongly classified as low. For an intermediate true density, the probability of it being classified as high or low depends on its position in the intermediate range. The closer the intermediate true

density is to 2.5, for example, the higher the probability of it being classified as high.

The number of observations required or make a decision will depend on the true population density. If the true density is very low or very high, sample average leading to a decision will be obtained on the average with fewer observations than if the true density is intermediate. Also, for any one true density the number of observations required to make a decision will vary from one sample to another. Even when the true population density is very high or low, a particular sample may require an unusually large number of observations before a decision can be made with the chosen standards of accuracy. As a result, the number of observations needed for making decisions using sequential sampling is best described as an "average sample number" (ASN) associated with a particular true population density. For example, if the true population density is 2.5, then on the average 33 observations will be required to classify the population density as low or high using the categories and standards described above (Appendix 3, and Figure 1). Since the traditional sampling method required 91 observations, the sequential sample on the average would result in a decision with 58 fewer observations if the true density is 2.5. The larger traditional sample may provide a more precise estimate of the true population density but it is no better than the sequential sample for the purpose of deciding whether the field should be sprayed given the criterion for treatment that we have adopted.

Since the average sample number (ASN) depends on the true population density, the savings from using sequential sampling will also depend on the true density. Figure 1 presents the relationship between ASN and true

density and indicates the average savings compared to the traditional sample providing the same standards of accuracy.

It is possible that a particular sequential sample may require more observations than the comparable traditional sample. In fact, the only assurance provided by the sampling theorists is that sample size will not be infinite (!) and will virtually never exceed three times the ASN (Wald, 1947, p. 105). In practice, a common policy is to stop sampling at some selected maximum sample size if the sequential decision rule has not provided a decision by then. Adoption of such a policy, however, increases the probability of misclassifying the population density (Wald, 1947, p. 64).

Information Needed for Sequential Sampling

Recall that all of the preceding calculations of sample sizes and levels of accuracy depend on the assumption that the Colorado Potato Beetle population has a spatial distribution in the field that can be accurately represented by the negative binomial distribution with an index of aggregation of 0.5. The usefulness of the negative binomial distribution is well documented for representing the distribution of Colorado Potato Beetle and other species with similar dispersion patterns, although a test of goodness of fit using data from the study area is certainly warranted. The value of the index of aggregation is less easily assumed. It has been suggested that the index of aggregation is an intrinsic characteristic of a species, one that does not vary with changes in population density (Anscombe, 1949). If populations behaved in a manner strictly consistent with the negative binomial model then mean density and the index of aggregation would vary independently, but in field studies estimates of the index increase as population density increases (Harcourt, 1963). Within the range of beetle

population densities found in commercial potato fields, the index possibly can be assumed to be constant without significantly affecting the sampling procedure. However, values of the index reported in the literature for Colorado Potato Beetle at high densities may be inappropriate for use in the pest management program. Estimation of the index of aggregation for commercial potato fields should be an important objective for developing a sampling program. Table 1 indicates the effect of the index value on sampling requirements when the density categories and standards of accuracy are the same as those described in the previous section.

Clearly, if the index of aggregation is much below 1.0, and published estimates (Harcourt, 1963) suggest that for low densities it probably is, the density categories and standards of accuracy described above are not feasible, even with sequential sampling. The effect on sample size from relaxing the accuracy requirements is represented in Figure 2.

For adjusting sampling requirements, an alternative to changing the standards of accuracy is to change the classification of low and high densities. For example, with the original 90% and 95% standards of accuracy and an index of aggregation equal to 0.5, the boundary values of 1.5 and 2.5 result in a maximum average sample number equal to 65 (Table 1). If a more precise distinction between low and high densities had been desired and the boundary values of 1.75 and 2.25 were used, the maximum average sample number would jump to 258.

Conclusions and Proposed Action

The purpose of this discussion has been to suggest possible benefits from sequential sampling and to highlight some of the important considerations in using it (see Table 2). The mechanics of calculating and

using the sequential decision rule were not discussed but these aspects of the procedure are straight forward once levels of precision and acceptable probabilities of error have been specified (Onsager, 1976). The specification of these levels and probabilities should receive careful consideration when sampling procedures are being adopted. Costs of sampling, the cost of inappropriate recommendations, and the probability of various population densities should be included among the considerations when these standards are specified.

Prior to implementing a sequential sampling program, the following work is necessary:

(1) Intensive field sampling to determine the spatial distribution of pests in commercial fields,

(2) Specification of the degree of precision that sampling should provide and the frequency of errors that can be tolerated,

(3) An activity analysis of the sampling procedure: timing the scout as various steps are performed and investigating the interrelationship among these steps,

(4) Computer analyses and field testing of alternative sampling programs.

Table 1. Effect of Index of Aggregation on Sampling Requirements

Index of Aggregation	Maximum Average Sample Number Using Sequential Sampling*	Size of Comparable Traditional Sample**
0.1	269	382
0.3	99	140
0.5	65	91
1.0	39	54
1.5	31	42
2.0	26	36

Boundary Values 1.5 and 2.5
$\alpha = .10$ $\beta = .05$

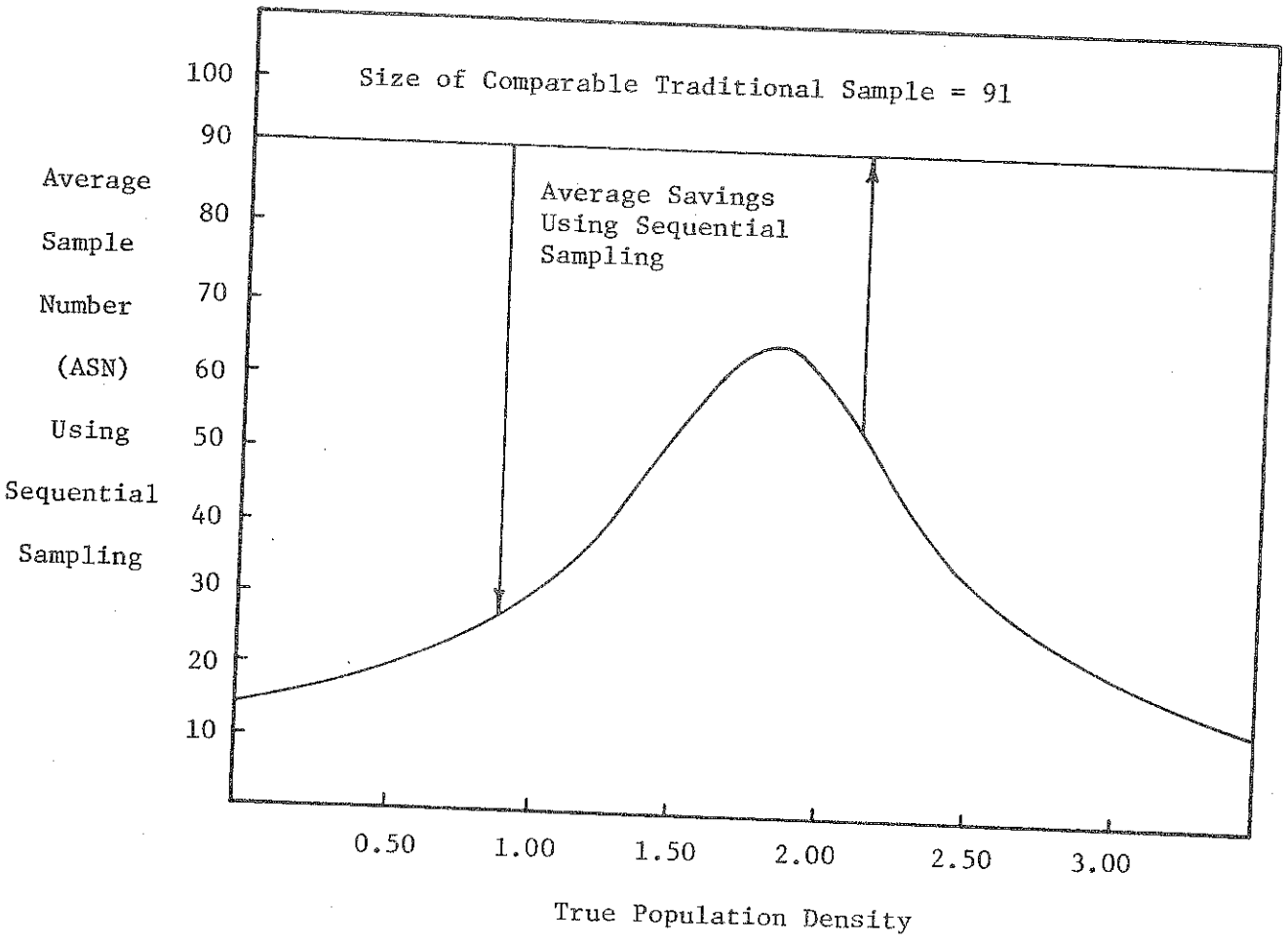
* Depending on the true population density, the sample size required using sequential sampling may be substantially less than this maximum (see Figure 1).

** See Appendix 2.

Table 2. Summary of the Relationship Between Sequential Sampling Requirements and Population and Decision Parameters

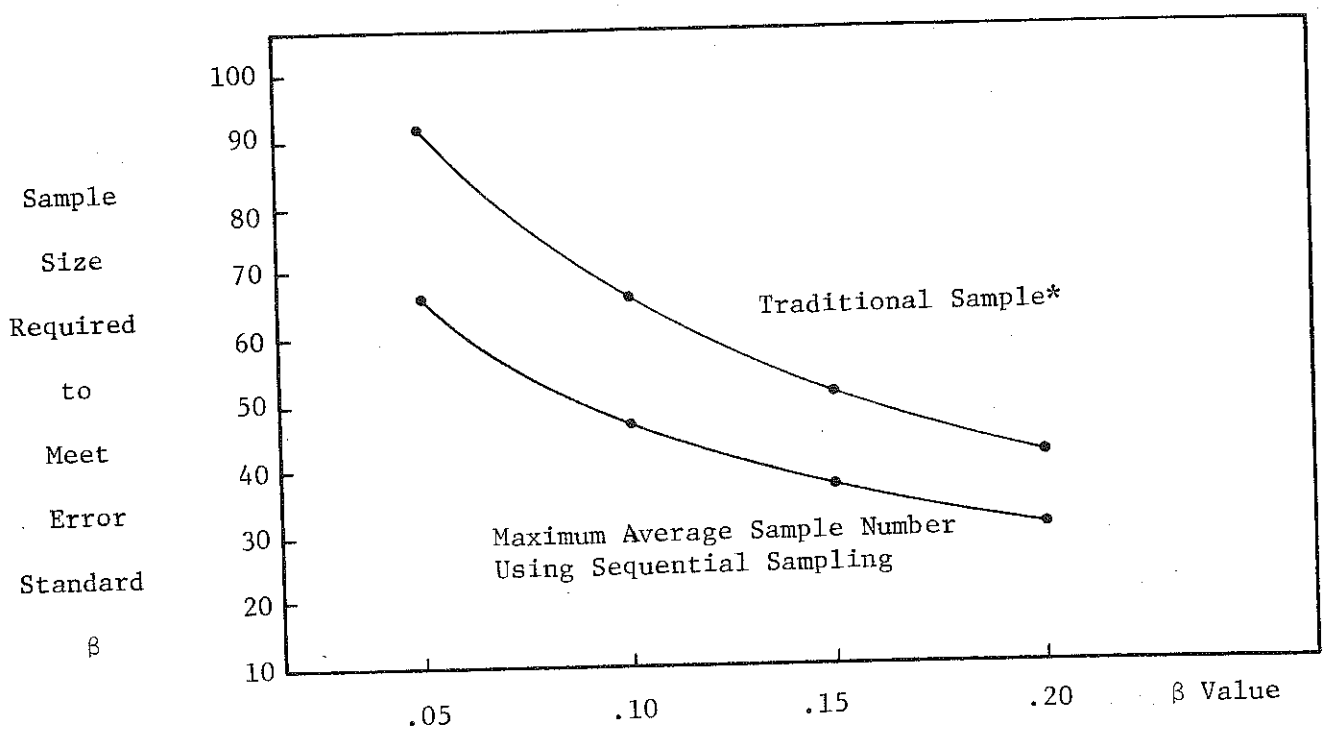
Change in Parameter	Effect on Average Sample Size
Higher Index of Aggregation	Decrease
True Density Farther Above Upper Boundary	Decrease
True Density Farther Below Lower Boundary	Decrease
Higher Tolerance for Misclassification	Decrease
Higher Upper Boundary Value	Decrease
Lower Lower Boundary Value	Decrease

Figure 1. Relationship Between True Population Density and Average Sample Number and Savings Using Sequential Sampling*.



* Population has negative binomial distribution, index of aggregation 0.5, Boundary Values 1.5 and 2.5, $\alpha = .10$, $\beta = .05$.

Figure 2. Standards of Accuracy and Sampling Requirements.



β = Probability that a High Density Population (density = 2.5 larvae/plant) will be Misclassified as a Low Density Population (less than 1.5 larvae/plant).

Probability of the converse misclassification: $\alpha = .10$

Population has a negative binomial distribution, index of aggregation 0.5

* See Appendix 2.

Appendix 1. Calculating the Probability that the Sample Average will be in a Particular Interval

An estimate of the probability distribution of sample averages can be obtained by relying on the Central Limit Theorem and approximating the distribution of the sample average using the normal distribution. This approximation is best when sample size is large and $\frac{m}{m+k}$ is close to .5. Using a continuity correction, the normal approximation for this problem is:

$$\text{probability } (\bar{x} < 2.0) = \text{prob.} \left(\frac{\bar{x} - m}{\sigma/\sqrt{n}} < \frac{2.0 - 2.5}{\sqrt{2.5 + \frac{(2.5)^2}{.5}}/\sqrt{30}}\right)$$

$$\doteq \text{prob.} \left(z \leq \frac{2.0-2.5-\frac{1}{2n}}{\sqrt{15}/\sqrt{30}}\right) = \text{prob.} (z \leq -.7307) = .23$$

An alternative to the normal approximation is to use the exact distribution of the number of insects in a sample. If the number of insects per plant has a negative binomial distribution with an index of aggregation (k) of 0.5 then the number of insects per 30 randomly selected plants has a negative binomial distribution with k = .5 x 30 = 15. Also, the threshold of 2.0 per plant is equivalent to 2 x 30 = 60 per 30 plants, and a mean density of 2.5 per plant is equivalent to 75 per 30 plants. The problem involving average number of insects per plant can therefore be translated into one involving insects per sample. The cumulative negative binomial distribution can then be used to calculate the probability that the number of insects in the sample is less than 60. A computer program from the International Mathematical and Statistical Library (IMSL) can be used to perform this calculation. The IMSL does not contain a program explicitly for the negative binomial distribution but does have programs for the binomial distribution and the incomplete beta distribution. Problems

involving the negative binomial distribution can be reformulated into binomial or beta expressions as follows:

Cumulative Negative Binomial Distribution, mean density = mn, index = kn:

$$\text{probability } (x < nc) = \sum_{r=0}^{nc-1} \binom{r+nk-1}{nk-1} \left(1 - \frac{m}{m+k}\right)^{nk} \left(\frac{m}{m+k}\right)^{r}$$

is equivalent to the Cumulative Binomial Distribution, number of trials

$$= (nc - 1) + nk, \quad p = \left(1 - \frac{m}{m + k}\right):$$

$$= \sum_{r=nk}^{(nc-1)+nk} \binom{(nc-1)+nk}{r} \left(1 - \frac{m}{m+k}\right)^{r} \left(\frac{m}{m+k}\right)^{(nc-1)+nk-r}$$

and equivalent to the Incomplete Beta Distribution,

$$X = \left(1 - \frac{m}{m + k}\right), \text{ parameters } a=nk, b=nc:$$

$$= \frac{\Gamma(nk+nc)}{\Gamma(nk)\Gamma(nc)} \int_0^{\left(1 - \frac{m}{m + k}\right)} u^{(nk-1)} (1-u)^{(nc-1)} \, du$$

($\Gamma(\cdot)$ is notation for a gamma integral).

The formulation in terms of the beta distribution is more flexible since noninteger values of nk and nc are permitted; nc and nk must be integers to use the program for the binomial distribution.

For the example described above, the probability calculated using the IMSL program is .24 that sampling will result in an inappropriate recommendation.

Appendix 2. Calculating Sample Size for Traditional Sampling

Sample size needed for traditional sampling can be calculated by approximating the distribution of the sampling average using the normal distribution. This approximation is best when the sample size (n) is large and $\frac{m}{m+k}$ is near .5 (where m is the mean density and k is the index of aggregation), but has performed well for the range of n, m, and k described in this paper. The results reported in this paper were checked by comparing the α and β levels suggested by the normal approximation with the α and β levels calculated using the true cumulative negative binomial distribution (Table A1). These calculations were made using a computer program contained in the International Mathematical and Statistical Library (IMSL), (see Appendix 1). The computer program could be adapted to iteratively calculate sample sizes for specified α and β levels but the time and expense to develop and use such an approach seems unjustified for our purposes given the adequate performance of the normal approximation.

Using the normal approximation, sample size for traditional sampling is estimated by specifying the two expressions implied by the α and β levels, and then using the two expressions to solve for two unknowns: n (sample size) and c (the critical value for testing the alternative hypotheses about population density) (Wald, 1947, p. 54). The two expressions implied by α and β are:

$$(1)\ \text{prob.}\ (\bar{x} \geq c \mid \text{given true mean density} = m_0) \doteq 1 - \text{prob.}\ \left(z \leq \frac{c - m_0 - \frac{1}{2n}}{\sigma_0/\sqrt{n}}\right) = \alpha$$

$$\text{so}\quad \frac{c - m_0 - \frac{1}{2n}}{\sigma_0/\sqrt{n}} = q_{1-\alpha}$$

Table A1. Comparison of β levels from Normal Approximation and Exact Distribution Calculated Using IMSL Program

Index of Aggregation	Traditional Sample Size	β Level Using Normal Approximation	β Level Using Exact Distribution
0.1	382	.05	.039
0.3	140	.05	.039
0.5	91	.05	.040
1.0	54	.05	.041
1.5	42	.05	.041
2.0	36	.05	.042

Data used for Figure 2 Traditional Sample Size	β Level Using Normal Approximation	β Level Using Exact Distribution
91	.05	.040
66	.10	.091
51	.15	.148
41	.20	.204

Index of Aggregation = 0.5

(2) prob. $(\bar{x} < c \mid \text{given true mean density} = m_1) \doteq \text{prob.}\ (z \leq \frac{c - m_1 - \frac{1}{2n}}{\sigma_{1/\sqrt{n}}}) = \beta$

$$\text{so} \quad \frac{c - m_1 - \frac{1}{2n}}{\sigma_{1/\sqrt{n}}} = q_\beta$$

where $q_{1-\alpha}$ and q_β are the quantiles of the normal distribution corresponding to cumulative probability of $1 - \alpha$ and β respectively. Substituting and solving for n and c:

$$\text{sample size } n = \left(\frac{q_{1-\alpha}\sigma_0 - q_\beta \sigma_1}{m_1 - m_0}\right)^2$$

$$\text{critical value } c = \frac{q_{1-\alpha}\sigma_0}{\sqrt{n}} + m_0 + \frac{1}{2n}$$

With a sample size n, and critical value c for testing the hypotheses of low versus high density, the probability is α that a population with a true density m_0 would be erroneously classified as high; the probability is β that a population with true density m_1, would be erroneously classified as low.

Appendix 3. Formula for Average Sample Number (ASN) for the Negative Binomial Distribution (Oakland, 1950; Onsager, 1976)

$$E_m(n) = \frac{h_1 + (h_0 - h_1)\, L(\frac{m}{k})}{m - s} = \text{ASN given true mean equal to } m$$

$$h_0 = \frac{\ln B}{\ln(\frac{p_1 q_0}{p_0 q_1})} \qquad h_1 = \frac{\ln A}{\ln(\frac{p_1 q_0}{p_0 q_1})}$$

$$S = k\, \frac{\ln(\frac{q_1}{q_0})}{\ln(\frac{p_1 q_0}{p_0 q_1})} \qquad B = \frac{\beta}{1-\alpha} \qquad p_i = m_i/k$$

$$A = \frac{1-\beta}{\alpha} \qquad q_i = 1 + p_i$$

m_0 = lower boundary value for mean density

m_1 = upper boundary value for mean density

k = index of aggregation

α, β = levels of significance

$$L(\frac{m}{k}) = \frac{A^d - 1}{A^d - B^d}$$

and

$$(\frac{m}{k}) = \frac{1 - (\frac{q_0}{q_1})^d}{(\frac{p_1 q_0}{p_0 q_1})^d - 1}$$

Values of the function $L(\cdot)$ of $\frac{m}{k}$ are found by varying d and solving for the corresponding values of $\frac{m}{k}$ and $L(\frac{m}{k})$.

d = 1 gives $L(\frac{m_0}{k})$

d = -1 gives $L(\frac{m_1}{k})$

References

(1) Anscombe, F. J. 1949. The statistical analysis of insect counts based on the negative binomial distribution. Biometrics (5): 165-173.

(2) Harcourt, D. G. 1963. Population dynamics of Leptinotarsa decemlineata (Say) in Eastern Ontario: I. Spatial pattern and transformation of field counts. The Canadian Entomologist (95): 813-820.

(3) Oakland, G. B. 1950. An application of sequential analysis to whitefish sampling. Biometrics (6): 59-67.

(4) Onsager, J. A. 1976. The rationale of sequential sampling, with emphasis on its use in pest management. U.S.D.A. Technical Bulletin No. 1526. 19p.

(5) Wald, A. 1947. Sequential Analysis. New York, Wiley & Sons. 212p.