Proceedings of the Institute of Acoustics

A NOTE ON THE ANALYSIS OF MARINE SURVEY DATA

G M Jolly(1) and S J Smith(2)

(1)38 Silverknowes Terrace, Edinburgh, UK
(2)Bedford Institute of Oceanography, Dartmouth, N.S. Canada

ABSTRACT

The question of whether transformations have a role in the analysis of marine survey data is discussed with particular reference to the Δ-distribution of Aitchison and Brown [1]. The distinction is made between design-based and model-based estimators of the mean and variance of a population sampled acoustically or otherwise. Analyses of real data on the assumption that the data follow a Δ-distribution are compared with standard methodology for the sampling of finite populations. The results appear to confirm theoretical arguments that the model-based method is less robust, being less efficient on average and subject to large unpredictable biases. Possibilities for future work are mentioned.

An approximative method is suggested for improving confidence limits for small samples of highly skewed data.

INTRODUCTION

An objective of surveyors in any field is to obtain information of an acceptable standard at minimum cost. In the fishery acoustic context Shotton and Bazigos [2] (pp 40,51) cite examples of transformations being applied to create more nearly normally distributed data. Sometimes there is an apparent assumption that normality is necessary if bias is to be avoided in estimation of the mean. In other instances the aim of a transformation with its associated estimators is to improve efficiency as demonstrated by Finney [3] in sampling from a distribution known to be lognormal. Jolly and Hampton [4] note that in classical sampling theory, e.g. Cochran [5], the property of unbiasedness, in a well-designed survey, exists independently of the distribution of the data; only in assigning confidence limits is a distributional assumption made, this being usually, but not necessarily, one of normality.

A distribution considered by some to have potential interest in improving the efficiency of marine surveys is the Δ-distribution, so-named by Aitchison and Brown [1]. In this the data elements are assumed to be either zero or positive, with the latter being lognormally distributed. Pennington [6] has applied and extended the results of Aitchison [7] to give formulae for the estimated mean and variance of the distribution, and for the estimated variance of the estimated mean. Pennington, while claiming some success in increasing efficiency, also notes that in other instances efficiency appears to be reduced. Myers and Pepin [8] have used simulations to test the robustness of the distribution. They conclude that the method is sensitive to small, and probably undetectable, violations in the model, and therefore cannot be recommended for general use in estimating population abundance.

A NOTE ON THE ANALYSIS OF MARINE SURVEY DATA

Following a discussion of the concepts of design-based and model-based sampling theory, the present paper examines the performance of the above method on real data from 1) an acoustic survey of South African anchovy and 2) a trawl survey of haddock on the Eastern Nova Scotian shelf.

THEORY

Design-based estimation

The methods of analysis conventionally employed for sample surveys in general are directly linked to the sampling design which in turn is based on classical sampling theory for finite populations. The concept of a finite population (of sampling units) is explained in the first chapter of Barnett [9]. The word "finite" distinguishes a real-life population of units comprised of individual animals, dwellings, fish shoals, etc. from an imagined infinite population of such units, with the added implication that the latter is usually envisaged as having a mathematically simple distributional form such as normal, Poisson, lognormal, etc.

The two dominant features of finite population sampling methodology, as made clear, for example, in Cochran [5], are 1) its basis in the random sample and 2) the fact that no distributional assumptions are required in respect of the surveyed population. Sampling designs appropriate to marine acoustic surveys of biomass are described briefly by Jolly and Hampton [4], and, along with methods of analysis, more fully in Jolly and Hampton [10].

Considering a simple random sample design for one stratum where the sampling units are parallel transects, the mean observed fish density is estimated as

$$\hat{Y} = \sum_{i=1}^{n} w_i \bar{y}_i / n \qquad (1)$$

where $w_i = L_i/\bar{L}$, $\bar{L} = \sum_{i=1}^{n} L_i/n$, L_i = length of i'th sample transect, $\bar{y}_i$ is the mean observed density for transect i, and n is the number of sample transects. The variance of $\hat{Y}$ is estimated as

$$\text{var}(\hat{Y}) = s^2/n \qquad (2)$$

where $s^2 = \sum_{i=1}^{n} w_i^2 (\bar{y}_i - \hat{Y})^2/(n-1)$.

In the formula for s^2 the finite population correction factor, 1-f, has been omitted since the sampling fraction, f, being the proportion sampled of all possible transects, is very small. The estimate of total biomass is $\hat{Y}$ multiplied by the stratum area.

The sample mean square, s^2, is an unbiased estimate of the population mean square, S^2, which is defined similarly to s^2 but with n replaced by N, the total number of possible transects. (The actual value of N in the present context would depend on the width assigned to a single transect so may appear somewhat of an abstraction, but this is unimportant.) Thus, the mean square,

S^2/n, with estimate, s^2/n, represents the exact variance that would occur among a series of means, $\bar{Y}$, from repeated independent samples of n transects. This leads to the concept of design-unbiasedness. Both $\bar{Y}$ and its estimated variance, var($\bar{Y}$), can be said to be "design-unbiased" estimators in contrast to Cochran's term "model-unbiased" for an estimator whose unbiasedness (slightly differently defined) is not absolute but depends on the assumptions of a model being met. The terms "unbiased" and "design-unbiased" will be used synonymously. Since S^2 is the exact transect variance for the given method of sampling, the efficiency of any alternative sampling scheme will be determined from a comparison of the latter's estimated variance of its estimator of $\bar{Y}$ with var($\bar{Y}$) = s^2/n.

Similar considerations hold for the mean. Although it can be shown that a model-based estimator of $\bar{Y}$ can be design-unbiased only if it equals the unbiased estimator, $\bar{Y}$, a model-based estimator could be acceptable if its variance were less than that of $\bar{Y}$, and its bias were small. Bias in the estimator would be assessed from a comparison with $\bar{Y}$.

Model-based estimation

What then are the implications of a model-based theory, in particular, one based on the Δ-distribution? If any benefit is to be derived from the model, the sample data must have properties resembling those of a random sample from a Δ-distribution. Immediately, a problem arises since it is already known that the random sample has the properties of one from a specific finite population. Thus, the finite population of sampling units must itself approximate to a Δ-distribution. The average of a large number of populations may do so (by the central limit theorem) during its approach to normality, but this is unlikely to be true of many single populations, especially if small. In the next section comparisons are made for two typical data sets in terms of bias and efficiency between estimates from conventional sampling methods and those from the assumption of a Δ-distribution.

The Δ-distribution is defined as follows. Suppose x to represent the non-zero elements of a sample from a Δ-distribution. Then, if $y = \ln(x)$, y will be normally distributed. For a sample of n elements of which m are non-zero, Pennington's [6] formulae (2) and (4) give respectively estimators c and d of the population mean and variance, and the estimated variance of c.

COMPARISON OF METHODS

Acoustic survey of anchovy

The data of Table 1 are from the annual survey of the spawning anchovy stock off the South Coast of South Africa conducted in November 1986. The survey design was a stratified random sample of parallel transects. Of six strata surveyed, four were in regions of low anchovy density and were consequently sampled less intensively. Since the remaining two strata provided the main bulk of data, only those are included in Table 1, as Strata 1 and 2.

TABLE 1

Cape anchovy acoustic survey, 1986

Estimation		Weighting for section transect (Sampling unit) and number of			Means (T/Km^2) STRATA 1	2	Weighted Means of Strata	Assuming† Single Stratum
A	S	L	L	Mean	55.6	30.0	44.9	45.0
		(Transect) 27		S.E.	8.3	9.0	6.1	6.7
				CV%	15.0	30.1	13.7	14.9
B	S	E	NS	Mean	62.6	30.5	49.2	49.5
		(Transect) 27		S.E.	10.7	9.6	7.4	8.1
				CV%	17.2	31.4	15.1	16.4
C	S	E	-	Mean	62.6	30.5	49.2	49.5
		(Section) 231		S.E.	11.3	8.7	7.5	7.7
				CV%	18.1	28.5	15.3	15.5
D	M	E	-	Mean	102.5	28.2	71.4	65.6
		(Section)		S.E.	31.8	21.0	20.5	14.2
				CV%	31.0	74.5	28.7	21.7
	M	Prop. Bias of D v C			1.6	0.9	1.5	1.3
		Efficiency* of D v C			0.34	0.15	0.28	0.51

S = Standard sampling method, M = Model-based
L = Length (of Section or Transect), NS = Number of Sections
E = Equal weighting
* Estimated as square of ratios of CV's
† Strata had same sampling intensity

In parts A, B and C of Table 1 the estimates of mean density and their standard errors (S.E.'s) are from formulae (1) and (2). In part A, the mean density for each integration section of transect was weighted by length of section to estimate the transect mean, and each transect mean was weighted by length of transect in forming the stratum mean. Total numbers of transects and sections are shown. The means in the penultimate column used the known stratum areas as weights. All S.E.'s in part A were calculated treating transects as the sampling units, so represented the appropriate form for the survey analysis. As the sampling intensity was the same for both strata, these could reasonably be analysed as a single stratum for present purposes as shown in the final column.

The analysis in B is the same as in A except that transect sections were given equal weight and number of sections replaced transect length in the transect weighting. Assuming section length to be uncorrelated with transect length, B also provides unbiased estimates of mean density but with the expectation of slightly increased S.E.'s, as the table shows. Part C is not a valid analysis unless between-and within-transect variation were equal, but is used to give as many degrees of freedom as possible for a comparison with D. Estimation in D is from Pennington's [6] formulae on the assumption that section mean densities approximate to a sample from a Δ-distribution. The comparison of C with D therefore gives a direct comparison of the model-based estimates with unbiased estimates from standard procedures independently of the survey design.

The last two lines of Table 1 show the proportional biases of the model-based method as the ratio of the estimated means of D to those of C, and the relative efficiencies of D to C. To remove dependency on bias, efficiency has been measured as the square of the ratio of coefficients of variation (CV) rather than of S.E.'s.

Stratum 1 shows a positive proportional bias of 60% and Stratum 2 a negative bias of 10%, with an average positive bias of 50%. The average efficiency of the model-based method is estimated at 28%, rising to 51% when the strata are treated as a single stratum, reflecting the relative instability of the model-based estimators.

Sample data from a trawl survey of haddock

The haddock estimates in Table 2 are from one region of a trawl survey on the Eastern Scotian Shelf conducted in March 1987 and designed to produce catch estimates for standardised effort. The data are from a stratified random sample of tows. Of the ten strata, only four are shown, the catch being zero in four and insignificantly small in two of the remaining strata.

Smith [11] has shown that, for the small sample sizes often encountered in trawl surveys, the large-sample approximation used by previous authors for the estimated variance of the estimates of the mean is inadequate; in its place he gives the exact form and this has been used for Table 2. The penultimate two rows of the table give the proportional bias and efficiency of the model-based method relative to standard sampling procedures. Both bias and efficiency fluctuate widely, partly due doubtless to the small sample sizes which range from 6 to 14. There is no evidence of a gain on average from the model-based method.

The last line of the table gives estimates of the efficiency that would be attained if the lognormal model were appropriate. This is calculated as the ratio of the estimated variance, v/n, to the estimated variance of the model-based estimate of the mean, where v is the estimate (assuming the model applies) of the ordinary sample variance, S^2.

TABLE 2

Nova Scotian haddock trawl survey, 1987

Estimation		Strata 1	2	3	4
	No. of tows	14	6	13	7
S	Mean catch	25.1	644.8	147.0	169.0
	S.E. of mean	24.8	516.9	73.8	73.6
	CV%	99.0	80.2	50.2	43.6
M	Mean catch	17.3	570.4	336.5	207.9
	S.E. of mean	16.5	466.0	281.6	115.1
	CV%	95.8	81.7	83.7	55.4
M	Prop. Bias of M v S	0.7	0.9	2.3	1.2
	Efficiency of M v S	2.3	1.2	0.1	0.4
	Effficiency of model fitted	1.1	1.4	1.8	1.1

S = Standard sampling method, M = Model-based

DISCUSSION

We hope to have shown by simple theoretical argument that, in a single, finite natural population, there is no reason to expect that sampling unit means should follow a lognormal or other simple distribution. Although the amount of data presented is not large, we believe it is sufficient to show that a method based on a lognormal assumption can be very misleading and cannot be recommended generally as a means of increasing efficiency. The results in Tables 1 and 2, in our opinion, amply support the findings of Myers and Pepin [8] from simulation.

A more promising line of investigation appears to be that exemplified by Brewer [12], Little [13] and Tam [14] who consider the property of "asymptotic design-consistency". One of us (S.J.S.) is currently examining possibilities within this family of estimators.

When sample size on occasion falls well below an acceptable level, asymmetry in the distribution of the mean may result in unrealistic confidence limits if these are based on the normal distribution. A simple solution to this problem is recognition that the distribution of the mean is likely to be better approximated by a Poisson than by the normal. On this basis Jolly and Hampton [4] suggest the more appropriate limits obtained by adding to each 95% normal-based confidence limit the quantity $\bar{x}c^2$, where $\bar{x}$ is the estimated mean and c its coefficient of variation.

ACKNOWLEDGEMENTS

We wish to thank Ian Hampton, Sea Fisheries, Cape Town for use of the anchovy data of Table 1.

REFERENCES

[1] Aitchison, J. and Brown, J.A.C.- 1957. The Lognormal Distribution. Cambridge: Cambridge University Press.

[2] Shotton, R. and Bazigos, G.P. - 1984. Techniques and considerations in the design of acoustic surveys. Rapp. P.-V Reun. Cons. int. Explor. Mer., 184, 34-57.

[3] Finney, D.J. - 1941. On the distribution of a variate whose logarithm is normally distributed. J.R. statist. Soc., Suppl. 7, 155-161.

[4] Jolly, G.M. and Hampton, I. (in press). Some problems in the statistical design and analysis of acoustic surveys to assess fish biomass. In Proceedings of the 1987 International Symposium on Fisheries Acoustics, Seattle, U.S.A.

[5] Cochran, W.G. - 1977. Sampling Techniques. New York: Wiley.

[6] Pennington, M. - 1983. Efficient estimators of abundance, for fish and plankton surveys. Biometrics, 39, 281-286.

[7] Aitchison, J. - 1955. On the distribution of a positive random variable having a discrete probability mass at the origin. J. Am. statist. Ass., 50, 901-908.

[8] Myers, R.A. and Pepin, P. (in draft). Dept. of Fisheries, Newfoundland, Canada.

[9] Barnett, V. - 1974. Elements of Sampling Theory. London: The English Universities Press Ltd.

[10] Jolly, G.M. and Hampton, I. (in draft). Acoustic survey methods to estimate biomass of a fish population.

[11] Smith, S.J.- 1988. Evaluating the efficiency of the distribution mean estimator. Biometrics, 44, 485-493.

[12] Brewer, K.R.W. - 1979. A class of robust sampling designs for large-scale surveys. J. Am. statist. Ass., 74, 911-915.

[13] Little, R.J.A. - 1983. Estimating a finite population mean from unequal probability samples. J. Am. statist. Ass., 78, 596-604.

[14] Tam, S.M. - 1988. Asymptotically design-unbiased predictors in survey sampling. Biometrika, 75, 175-177.