NOT TO BE CITED WITHOUT PRIOR
REFERENCE TO THE AUTHOR(S)

International Commission for the Northwest Atlantic Fisheries

Serial No. 5453

ICNAF Res. Doc. 79/VI/91

ANNUAL MEETING - JUNE 1979

Results of Application of Robust Algorithms for Estimation of Abundance Indices from the Trawl Survey Data (After the Example of Georges Bank Silver Hake)

by

P. S. Gasjukov
Atlantic Research Institute of Marine Fisheries and Oceanography (AtlantNIRO)
Kaliningrad, USSR

Abstract

Instability of abundance indices resulting from anomalies available in measuring methods (after the example of Georges Bank silver hake) can be reduced by using robust algorithms for the estimation of the mean. The efficiency of algorithms is illustrated by a model. The Georges Bank hake abundance indices calculated both by a standard method and using robust algorithms are given. Their comparison has exhibited a certain lack of correspondence in time.

Introduction

Trawl surveys are one of the widespread methods of fish stock assessment. The application of this method is aimed at determination of summed abundance and biomass by means of direct observations.

The method is based on estimation of the mean catch size per hauling, the accuracy of which estimate is a significant factor in stock assessment. In trawl surveys conducted in the ICNAF Convention Area such estimate is derived using the stratified-selective analysis (Cochran, 1976). The area under study is subdivided into subareas-strata - from which simple random samples are taken by trawl in randomly selected points. Resulting data for each species in the catch are summed abundance, total weight and length frequency.

In conducting a survey it is assumed that the positions of strata and the number of observations by stratum are determined in advance. The method allows for estimating the accuracy of the observation results by means of calculation of confidence intervals for the abundance index, that is, the mean catch per haul by area.

Some approaches have been attempted to estimate the actual accuracy of the method relative to different fish species (Grosslein, 1971; Gasjukov a. Dorovskikh, 1978). It has been found that the accuracy of the abundance index is relatively low, namely, of the order of 40-50%. Naturally, the question arises on the possibility of increasing the precision of the method. For this purpose a few approaches are outlined by Pennington and Grosslein (1978). Below, the possibility of increasing the accuracy of the method by means of some modification of method processing is considered.

Some features of the abundance index

Statistical characteristics of the abundance index - the mean catch size per haul given in numbers and by weight by stratum - are derived from the following formulae:

$$\bar{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi} \qquad (1)$$

$$S_h^2 = \frac{1}{n_h - 1} \sum^{n_h} (x_{hi} - \bar{x}_h)^2 \qquad (2)$$

$$S_{\bar{x}_h}^2 = \frac{1}{n_h} S_h^2 \qquad (3)$$

where $\bar{x}_h$ is the mean catch size from stratum h, S_h^2 is the catch size variance per haul by stratum, $S_{\bar{x}_h}^2$ are the values of catch size variance per haul by stratum, x_{hi} is the catch size in *i* haul in stratum h.

To process the observation data collected in a number of strata, the following formulae of selected stratified analysis are applied:

$$\bar{x}_{st} = \frac{1}{N} \sum_{h=1}^{K} N_h \cdot \bar{x}_h \qquad (4)$$

$$S^2_{(\bar{x}_{st})} = \frac{1}{N^2} \sum_{h=1}^{K} \frac{N_h^2 \; S_h^2}{n_h} \qquad (5)$$

where $\bar{x}_{st}$ is the abundance index for a number of strata, $S^2_{(\bar{x}_{st})}$ is the variance of abundance index for a number of strata, N is the total square of the area, N_h is the square of the stratum h, K is the stratum number.

A stratified analysis is an effective method for reduction of variance. The mean catch per haul by stratum, however, is estimated by a classical method using the arithmetical mean. As is evident from a number of investigation,the latter is extremely sensitive to any departure from the assumptions that are taken as a principle of its application. The arithmetical mean, in particular, is readily subject to large "splashes" characteristic of real phenomena.

Let us consider a few examples. During the trawl survey of stratum 13 conducted by the USA research vessel "ALBATROSS-IV" in 1974 on Georges Bank the silver hake was caught (table 1).

The abundance index for this stratum calculated using a routine method was 370. If one haul containing 2 078 specimens were omitted, then the abundance index would be 178, that is it would reduce twice.

This was also the case in 1976, when the Soviet research vessel made a survey of stratum 13 on Georges Bank. The silver hake catches by haul are given in Table 2. On this occasion the abundance index for 8 hauls was 372 or 144 (!) if the 8th haul is excluded. This is a direct evidence of sensibility of a standard processing method to the catch size per haul.

Robust methods for estimates of an average haul

All the above-mentioned cases are evidently rather a rule than an exception. Therefore, in the recent years a new approach is being intensively developed in the mathematical statistics, which is known as the robust or stable estimation method.

The efficiency of the stable estimation method is somewhat lower compared with standard classical methods, and, in the absence of anomalous dimensions, give the results that are in close agreement with those yielded by standard methods. In the presence of discards, however, their usage makes it possible to avoid negative consequences: the obtained estimates are stable.

The simplest algorithms applied for estimating the average values and based on stable methods are given by Tukey and McLaughlin (1963), Andrews, Bickel, Hampel, Huber, Rogers and Tukey (1972) and Ershov (1978).

Let us assume that the value of α is $0 < \alpha < 0.5$ and then regulate the observation results with regard for increasing values of $y_1 \leqslant y_2 \leqslant \dots \leqslant y_n$, where y_i is the i-th observation by size and n is the total observation number.

On rejection of $100 \cdot \alpha\%$ values on the right and on the left of the regulated observation sequence, a standard selected mean may be derived from the elements maintained:

$$C_t \;(\alpha_1 n) = \frac{1}{n-2[\alpha n]} \sum_{i=[\alpha n]+1}^{n-[\alpha n]} y_i \qquad (6)$$

Such an estimate of the mean is called an α-trimmed mean. In this case the value $n^{\frac{1}{2}}$ $(C\;(\alpha_1 n) - C^*)$ has an approximately normal distribution with the zero mathematical expectation and asymptotic variance

---

C^* is the actual value

$$\sigma_t^2(\alpha) = \left(\int_{Z_\alpha}^{Z_{1-\alpha}} Z^2 P(Z)dZ + 2\alpha Z^2_\alpha\right) / (1 - 2\alpha)^2 \qquad (7)$$

where P(Z) is the function of the distribution of the random estimate possibilities, P(Z) is the possibilities density and $F(Z_\alpha) = \beta$.

The second algorithm is derived in a similar way, however, the extreme members of observations regulated with an increase are not rejected but projected on the nearest left point:

$$C_w(\alpha_1 n) = \frac{1}{n} \sum_{i=[\alpha n]+2}^{n-[\alpha n]-1} y_i + [\alpha n] (y_{[\alpha n]+1} + y_{n-[\alpha n]}) \qquad (8)$$

For this estimate named the α-winsorized mean the following expression is used for asymptotic variance (Ershov, 1978):

$$\sigma_w^2(\alpha) = \int_{Z_\alpha}^{Z_{1-\alpha}} Z^2 P(Z)dZ + 2\alpha [Z_{1-\alpha} + \alpha/P(Z_\alpha)]^2 \qquad (9)$$

However, the possibility of deducing stable estimates of random value variance by sample is of greater interest. Here Tukey and McLanghlin (1963) recommended the usage of winsorized variance.

The estimates derived from the formulae (6) and (8) are obviously dependent on the trimming parameter α, the choice of which, however, lacks certainty. As soon as the possibility of choice is offered, the adaptive stable procedures for estimating the mean value can be suggested. Thus, Andrews, Bickel, Hampel Huber, Rogers and Tukey (1972) recommended the choice of such a value of the trimming parameter within the $0 < \alpha < 0.25$ interval that would provide the minimum asymptotic variance of the α-trimmed mean. The usage of the α-winsorized variance for this purpose is recommended by Jaeckal L.A. (Ershov, 1978).

Strictly speaking, the suggested expressions should be used for the symmetrical distribution function. If the distribution function is asymmetric, which is typical of the distributional properties of the catch-per-tow, then the estimates would be biased, although stable algorithms would maintain their principal characteristics (Tsypkin, Poliak, 1977; Meshalkin, Smirnov, Sosnovsky, 1969). One way to reduce the bias would be the usage of the "jackknife" procedure (Ershov, 1978; Kendall, Stewart, 1973) The following formulae are applied to derive the estimates using the "jackknife" procedure:

$$C_n^o = \sum_{i=1}^{n} C_{ni}^o/n \qquad (10)$$

$$C_{ni}^o = n \cdot Cn(y_1, y_2, \ldots, y_n) - (n-1) \cdot Cn-1(y_1, y_2, \ldots y_{i-1}, y_{i+1}, \ldots y_n) \qquad (11)$$

where C_n^o is the mean calculated by one or another method from the n sample. If the bias of the major estimate is of the order of n^{-1}, then the estimate (10) has a n^{-2} bias.

Comparative analysis of different algorithms for calculation of abundance indices using the method of statistical modeling

Statistical modeling was used to estimate the accuracy of abundance indices derived by different data processing methods including stable methods.

As in the paper by Gasjukov and Dorovskikh (1978), the negative-binominal distribution (NBD) has been used as a model of spacial distribution of the fish. The function of the NBD distribution is as follows:

$$P_r = \frac{\Gamma(r+k) \cdot P^r}{r! \, \Gamma(k) \cdot Q^{k+r}} \qquad (12)$$

where Γ is a gamma-function, $Q=P+1$, $P=m/k$, m is the mean distribution and K is the parameter. Unlike other functions used for description of the regularities in spacial distribution, the given distribution function may have biological considerations (Pielon, 1969; Taylor, 1953): such a distributional pattern is possible if the distribution of groups of individuals is random (under Poisson law), and that of individuals within the group submits to the logarithmic law.

Figs. 1 and 2 show histograms of the abundance index derived using the algorithm of the arithmetical mean and α-trimmed mean, which were obtained from modeling NBD with the parameters: $m = 100$, $K = 0.9$, and $m = 500$, $K = 0.9$.

As is evident from the analysis of histograms, the scatter of the abundance index derived from the α-trimmed mean method is lesser than that produced by the standard method. The variance of the index resulting from application of the first modeling version is 785 for the normal algorithm and 588 for the α-trimmed one, whereas the second version produced the values 21 500 and 15 606 respectively.

Thus, the use of the α-trimmed mean promotes a decrease in the abundance index variance by 20-25%, however, the resulting estimates are biased (approximately by 20) which can be seen easily from histograms.

A cross-checking of the observation data for different fish species (Taylor, 1953; Gasjukov and Dorovskikh, 1978) showed a close agreement between these data and NBD, thus suggesting that the results of modeling may be considered as representative ones. However, this hypothesis did not hold for all cases. The explanation for this phenomenon appears to be based on the assumption that in the nature there hardly exists a "pure" law of distribution. The hypothesis of "contaminated" distribution with the model generating large scale "splashes" or anomalous dimensions seems to be more realistic.

Analytically, the model of "contaminated" distribution may be written as

$$P_r = (1 - \beta)\, P_r^0 + \beta P_r^1 \qquad (13)$$

where P_r is the possibility of realization of the event, which consists in omitting r points, P_r^0 is the principal law distribution which affects the studied phenomenon, P_r^1 is the law of distribution of noises and anomalous dimensions, which can be realized with low β possibility.

There are two ways of interpreting the distribution function (Collins, 1976). The spacial distribution may be described either by the P_r^1 model using β possibility or by the law of distribution using the $(1 - \beta)$ possibility. This may be illustrated by two examples: (1) the aggregation of the fish over the spawning ground that does not coincide with the stratum area, and (2) the formation of temporary local fish aggregations of higher density promoted by the environmental factors. Certainly, it is hardly possible to distinguish between the above cases from the observation data, but an indirect evidence in favour of this hypothesis is given below.

Results of processing the observation data on Georges Bank silver hake, 1971-1976

The results of processing of the observation data on Georges Bank silver hake by stratum (no less than 6 tows within each stratum) using different algorithms for calculating the abundance index are given in Tables 3 and 4. The data were obtained over the 1971 to 1976 period by the USA research vessel "ALBATROSS-IV" and Soviet research vessels.

Conditionally, these results can be subdivided into three classes. The first class contains a group of observations that has not exhibited any significant difference as a result of calculation of the mean by various methods. These are, for example, the observations made in stratum 13 (Soviet research vessel) in 1972; in stratum 13 in 1971, and in stratum 13 ("ALBATROSS-IV") in 1971.

The second class combines the observations that have exhibited different mean values as a result of treatment by a standard method and other methods. These values, however, are rather uniform. This can be examplified by the observations made in stratum 20 in 1974, in stratum 24 in 1973, in stratum 16 (Soviet research vessel), in stratum 16 in 1971, in stratum 20 in 1972, and in stratum 19 ("ALBATROSS-IV") in 1976.

The third class combines the observations that have exhibited a marked difference as a result of calculation of the mean by a standard method and other methods. These are in particular the observations made in stratum 13 in 1974, in stratum 16 in 1973, in stratum 19 ("ALBATROSS-IV") in 1975, in stratum 21 in 1973, in stratum 19 in 1975 and in stratum 24 in 1976 (Soviet research vessel).

The observations representing the first class seem to be rather uniform without obvious anomalies, whereas those of the second class exhibit anomalous variations that can be allowed for by stable assessment algorithms. The mean values of the third class observations do not demonstrate any stability, and no data are available to judge on desirability of one or another value.

It should be noted here that the second and third classes were not involved in modeling of a simple NBD. These classes occurred in modeling of "contaminated" distribution, which to a certain degree, supports the hypothesis of its validity. When processing data at sea, it may be suggested to operatively increase the number of hauls within a stratum, until the stability of values is attained.

Table 5 shows the results of treatment of the observation data on Georges Bank silver hake (for a number of strata) over a series of years using the standard method of estimating the abundance index for a stratum, and the α-winsorized mean derived from the "jackknife" procedure at the optimum value of α, with respect to minimum variance. The latter algorithm is used only for processing the data obtained in strata, where more than 6 hauls were made. For other strata the common arithmetic mean was used. The abundance index for a number of strata was derived from the equation (4), and the variance of the mean from the equation (5), where the α-winsorized variance means by stratum were used.

The treatment results fall into two groups. The first group combines the data that have not exhibited any significant difference with application of stable assessment algorithms (1971, 1972, 1973, 1976). The second group includes the observations that have exhibited marked differences as a result of treatment (1974, 1975). The differences are 30% and 20% respectively. The variance means, however, should be considered as relative ones, since they were derived without allowance for possible asymmetry of the distribution function.

Summary

The variation of processing algorithms is one of the ways to improve the assessment accuracy when using a trawl survey method. Such a variation was pre-conditioned by considerable "splashes", anomalous observations that had occurred occasionally.

That the anomalous measurements result in unstable treatment values can best be explained by referring to an example of Georges Bank silver hake. In order to avoid this instability we suggest to apply robust algorithms for the assessment of abundance indices. Among these algorithms there are α-trimmed and α-winsorized means. A "jackknifing" procedure can be recommended for the asymmetric distribution to adjust the bias of the value.

The efficiency of robust algorithms used for estimating the abundance index is demonstrated by a test model, with NBD and "contaminated" NBD assumed as a model of spacial distribution.

The analysis of the modeling results indicates that the application of the algorithms of the α-trimmed and α-winsorized means gives a smaller error in the abundance index estimate compared with that generated by the arithmetical mean method provided, however, that there is a certain bias. A "jackknifing" procedure considerably reduces that bias or practically eliminates it if applied to the algorithm of the α-winsorized mean. This results in smaller variance of the value compared with that obtained by the general method.

The algorithms given in this paper have been used in estimating abundance indices for Georges Bank silver hake by stratum and a number of strata over a series of years. The comparison of these indices with those estimated by the arithmetical mean method has revealed a few discrepancies that can be attributed to the influence of anomalous measurements on the treatment results.

The use of robust algorithms for estimating the mean of random value with an asymmetrical distribution function (Collins, 1976) may be of interest in the future studies.

References

1. ANDREWS D.F., BICKEL P.I., HAMPEL F.R., HUBER P.I., ROGERS W.H., TUKEY I.W., 1972. Robust estimates of location. Survey and advances. Princebon, Princebon Univ. Press., p. 373.

2. CENDALL M.J., STEWART A., 1973. Statistical canclusions and correlations., M., "Nauka", p. 899.

3. COLLINS I.S., 1976. Robust estimation of a location parameter in the presence of asymmetry. Ann.Stat., v.4, pp. 68-85.

4. ERSHOV A.A., 1978. Stable methods for estimating parameters (review). In: Avtomatika i Telemekhanika, No 8, pp. 66-100.

5. GASJUKOV P.S., DOROVSKIKH R.S. On estimating the accuracy of silver hake Soviet research vessel trawling surveys of the Emerald Deep area in 1972-1976. Int.Comm.Northw.Atlant.Fish.Res.Doc. 78/VI/76.

6. GROSSLEIN M.D., Some observation on accuracy of abundance indices derived from research vessel surveys. Int. Comm.Northw.Atlant.Fish.Res.Doc. 71/59.

7. COCHRAN I,, 1976. Methods of selective investigation., M., "Statistika", p.440.

8. MESHALKIN L.D., SMIRNOV N.P., SOSNOVSKY N.N., 1969. On stability of estimates in the distribution center. "Zavodskaya laboratorija", No 5, pp. 417-422.

9. PENNINGTON M.R., GROSSLEIN M.D. Accuracy of Abundance Indices Based on Stratified-Random Trawl Surveys. Int. Comm.Northw.Atlant.Fish.Res.Doc. 78/VI/77.

10. PIELON E.C., 1969. An Introduction to the Mathematical Ecology. N.I. John Wiley and Sons, p.286.

11. TAYLOR C.C., 1953. Nature of variability in trawl catches. Fish.Bull. U.S. 58, 145-166.

12. TSYPKIN Ja.Z., POLIAK B.T., 1977. Rough maximum probability method. In: "System Dynamics. Mathematical methods for oscillation theory", No 12, Gorky State University, pp. 22-46.

13. TUKEY I.W., McLANGHLIN D.H., 1963. Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization. I. Sankhja, Ser. A., v. 25, pp. 331-352.

Table 1. Silver hake catch in numbers for Georges Bank stratum 13, 1974 ("Albatross-IV").

Stratum	Haul No								
	1	2	3	4	5	6	7	8	9
13	1	67	61	4	165	2 048	811	72	69

Table 2. Silver hake catch in numbers from Georges Bank stratum 13, 1976 ("Belogorsk").

Stratum	Haul No							
	1	2	3	4	5	6	7	8
13	0	12	77	167	107	172	307	2 134

Table 3. Results of treatment of observation data on silver hake using different algorithms for estimating the mean (from the data of trawl surveys of Georges Bank area for 1971-1976, "Albatross-IV").

Year	Stratum	Haul nos.	Arithm. mean	α-trimmed mean	α-winsorized mean	α-trimmed mean with "jackknifing"	α-winsorized mean with "jackknifing"
1	2	3	4	5	6	7	8
71	13	9	72	52	57	58	61
71	16	10	59	49	56	57	76
71	19	7	49	38	38	38	37
71	24	6	11	10	11	11	13
72	13	9	12	10	11	12	13
72	16	10	1.3	0.5	0.7	0.7	1.1
72	19	8	11	11	11	11	11
72	24	6	4.5	4.5	4.2	3.9	2.8
73	13	7	35	31	35	36	45
73	16	12	30	15	16	16	18
73	21	6	91	61	70	76	108
74	13	9	90.1	62.1	66.4	67.9	75.1
74	16	11	36.8	34.6	36.4	36.8	40.5
74	20	6	12.8	1.5	1.8	2.0	3.2
74	24	7	57.4	36.4	37.6	38.1	41.1
75	13	7	999.7	957.6	977.7	987.7	1 023.1
75	16	8	221.5	167.2	171.1	172.7	178.9
75	19	8	149.6	64.5	93.5	105.1	178.5
75	24	8	1 492.0	1 340	1 362	1 371	1 542
76	13	8	372.5	141	146	148	164.6
76	16	11	162.6	152.7	153.2	153	146
76	19	6	30.2	13.3	16.2	18.1	27.8
76	24	8	3 663	1 818	2 291	2 481	1 676

Table 4. Results of treatment of observation data on silver hake using different algorithms for estimating the mean (from the data of trawl surveys of Georges Bank area for 1971-1976, Soviet research vessels).

Year	Stratum	Haul nos.	Arithm. mean	α-trimmed mean	α-winsorized mean	α-trimmed mean with "jackknifing"	α-winsorized mean with "jackknifing"
1	2	3	4	5	6	7	8
71	13	9	7	6	6	7	8
71	16	12	31	13	15	15	19
71	19	8	9	2	2	2	3
71	20	6	22	16	20	22	32
71	24	7	9	4	4	5	5
72	13	9	27	22	26	27	34
72	16	13	13	7.3	8.2	8.4	8.5
72	19	10	6.4	5.5	5.9	6.0	6.5
72	20	6	3.3	0.5	0.5	0.5	0.5
72	24	6	55.2	57.5	56	55	50
73	13	9	6.5	6.1	6.4	6.5	7.5
73	16	11	19.2	12.7	15.3	16.0	21.1
73	19	9	11.7	9.1	9.8	10.0	9.8
73	20	6	12.0	2.0	2.3	2.6	3.7
74	13	9	370	173	229	246	360
74	16	12	122	74.5	96	101	130
74	19	9	365	314	343	353	409
74	20	6	9.7	9.0	8.5	8.2	6.5
75	13	9	9.1	8.9	9.0	9.0	9.4
75	16	11	19.1	12.8	13.3	13.4	13.1
75	19	9	70.4	32.0	45.7	50.2	80.1
75	20	6	277	89.7	107	117.9	174.3
75	24	7	24.8	23.4	23.1	23.0	21.7
76	13	9	7.6	6.7	6.9	6.9	7.3
76	16	10	11.2	6.8	8.0	8.4	11.2
76	19	9	36.3	14.6	16.4	17.1	20.0

Table 5. Abundance indices for Georges Bank silver hake estimated by the method of arithmetical mean and α-winsorized mean with the "Jackknifing" procedure ("Albatross-IV" data).

Year	Method of arithmetical mean		Method of α-winsorized mean with "jackknifing"	
	mean	variance of the mean	mean	variance of the mean
1971	36.2	204.3	32.9	187.2
1972	24.0	11.9	23.1	7.7
1973	19.2	20.9	18.7	18.2
1974	152.8	1 972.7	107.3	197.0
1975	48.0	358.7	39.5	52.4
1976	31.6	50.3	28.9	34.7

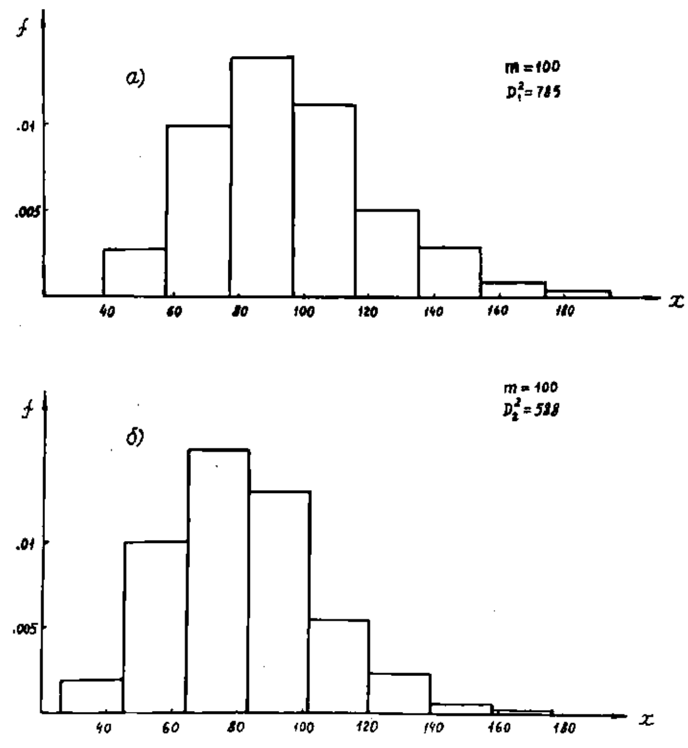

Fig. 1. Histogram of abundance index from the results of statistical modeling (negative-binomial distribution, m = 100, k = 0.9).
a) arithmetical mean b) α-trimmed mean.

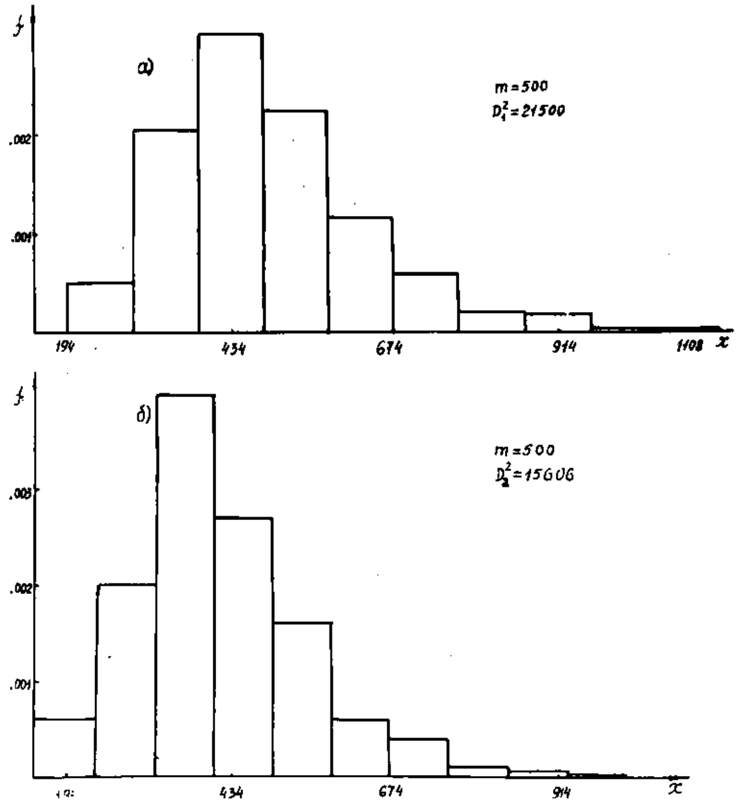

Fig. 2. Histogram of abundance index from the results of statistical modeling (negative-binomial distribution, m = 500, k = 0.9).
a) arithmetical mean b) α-trimmed mean.

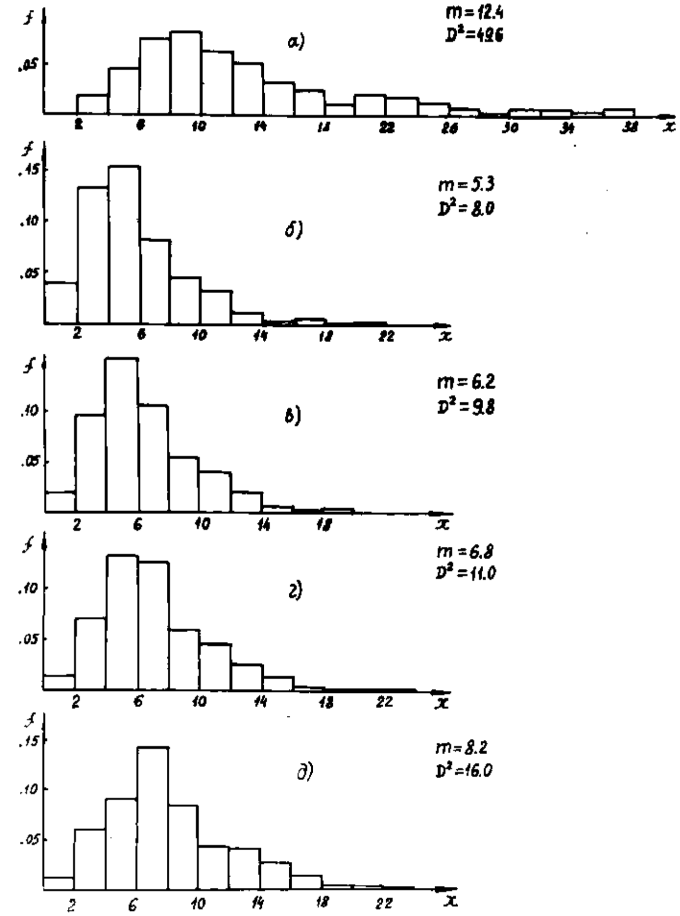

Fig. 3. Histogram of abundance index from the results of statistical modeling of "contaminated" negative-binomial distribution.
a) arithmetical mean; b) trimmed mean;
c) α winsorized mean; d) α-trimmed mean wtih the "jackknifing" procedure;
d) α-winsorized mean with the "jackknifing" procedure.