FISKERIDIREKTORATETS SKRIFTER
Serie Havundersøkelser
(Report on Norwegian Fishery and Marine Investigations Vol. IV, No. 2)

Published by the Director of Fisheries

Connected Frequency-Distributions

A Preliminar Account

By

EINAR LEA

1 9 3 3

A.s John Griegs Boktrykkeri, Bergen

In the course of some work concerned with the improvement of the methods for evaluating the individual growth-history of herrings from measurements on their scales the necessity arose to examine in more detail than has been commonly done, whether or not the statistical formulas and ways for treating the observational data, usually employed in such work, were really well suited for the purpose to be attained, which is a calculation formula giving better numerical results than the approximative formula published in 1910 (Lea, I).

Most of the results of the work carried out in this field of research since that time have been attained by procuring a set of paired observations of the total length of the fish and the linear size of one of its scales and treating these observations according to the prescriptions set out in the mathematical theory of correlation, based upon Bravais' formula, or after the method of least squares as developed by Legendre and Gauss, which method, by the way, will give numerical results which are identical with those obtained by means of the formulas of the theory of correlation, when the same presuppositions are made in either case. In some cases these methods have been used in a more empirical way, quite as Galton did, without taking recourse to the arithmetical formulas most commonly known under the designation "regression equations". The results of such empirical determinations are, in principle at least, similar to those obtained by means of the equations.

It soon became evident that the indiscriminate use of the said statistical methods would lead, and this with an apparent stringency, to erroneous conclusions concerning the functional connection between the size of the fish and the size of the scale, or scales, used in the investigation, and thereby also concerning the mathematical form of the equation to be used in growth-calculation. The real reason for this was found to be the presuppositions, upon which the calculations have been based, concerning the significance of the constants of the regression equations. These presuppositions, which are most often tacitly accepted, prove to be fallacious in the case under consideration. When, for instance, the regression equation expressing the average values of the total length L of the fish as a function of the size S of the scale is calculated, or determined in a more empirical way, all individual deviations from these average values are measured as if they had arisen exclusively from variation in the total length L. If, on the other hand, the other regression equation, expressing the average values of S as a function of L, is determined, then all deviations are measured as if they had now arisen solely by variation in S. These assumptions concerning the nature or modus of the deviations found are, as Sverdrup (II) has shown, implied in the very equations for the regression lines, and they are also made, if not expressly stated, when average values for one variable are calculated for definite values of the other variable, and a line connecting these values is found in one way or the other (Galtons own manner).

Undoubtedly cases occur, where one or the other of these assumptions may be considered as justified and approximately fulfilled for a particular purpose. But in no case this justification ought to be taken as granted or self-evident, and the greater the deviations are, that is the more the two regression lines diverge, the more a close examination of the fundamental assumptions will be needed. In the case under discussion, concerning the connection between the size (and growth) of the scale and that of the whole animal as expressed by the total length, this critical examination has been lacking. In view of the fact that results of calculations along the said lines are published rather frequently, it has been deemed advisable to write and publish this little preliminar account of a part of the writers researches into the matter, in the hope that further futile work may be thereby prevented. The present account is concerned mainly with the reasons, why the common manner of treatment cannot lead up to tenable results, and it has as a consequence a rather negative character. However, the task of rendering a positive contribution towards the proper solution of the problems has been achieved, or so it is believed. But as the exposition of these results is a much more complicated affair, involving the extensive use of observations, it cannot well be done in a short preliminar account and must be postponed until the complete report now under preparation can be published.

Whether the results to be described have been arrived at before or not is, of course, a question of minor importance, which the writer cannot decide, as he has not had access to enough literature on statistical theory. He has, however, found nothing similar in Polya's "Wahrscheinlichkeitsrechnung" or in Reibesell's "Biometrik" (Abderhalden's "Handbuch der Biologischen Arbeitsmethoden", Lieferung 165, Berlin-Wien, 1925), and he considers this as an indication that similar results as those expounded below had not, until five years ago, found their way to textbooks consulted by biologists, to whom the results ought to be of a particular interest.

The writer is aware that the conclusions reached may be applied outside the narrow field of research considered here. But as they are essentially quite simple, it will be an easy task to extend their use in the treatment of other problems. The main results achieved are, of course, independent of the particular kind of observations treated here, and they could equally well have been reached through abstract operations with pure mathematical fictions.

If a sample of herring, comprising a certain number of individuals of equal age, for instance ten years, is examined as to the total length (L). of the animals and the linear size (S) of a particular scale, the result will be a set of paired numerical observations, each pair of which can be represented in a cartesian co-ordinate system by a point, the position of which is defined by the particular value of L and S, and all points together will form what may be termed a *point-cluster* (P, or for the particular case P_{10}). Such a point-cluster is essentially the same as a tri-dimensional frequency-distribution, only the points or observations are not taken together in size-groups, but remain in their exact geometrical position.

Every particular herring in the sample has had a history leading up to the moment, when it was caught and came under observation. And if it had been possible to follow these histories as partly expressed by the total length and the size of the scale, point-clusters could have been formed for every age less than ten years (P_9, $P_{7.3}$ etc.)..

All these infinitely many point-clusters are connected in such a way that every particular point in one of the clusters corresponds to one particular point in every other of them. This is really the same as to say that every individual herring can be considered as characterized (with regard to L and S) by a line, the *individual line,* which terminates in a point belonging to the last point-cluster, the end-point-cluster. The individual lines stretch backwards (in time) towards the origin of the co-ordinate system.

Concerning the herrings it is now assumed that they have, during the whole course of their development and with respect to the two characters observed, behaved like isomorph geometrical figures. This assumption involves that the fraction or index L/S has had a constant value for every individual herring, although this value may be, and is assumed to be, different from one individual to another. This assumption involves that all individual lines are straight and point strictly towards the origin of the co-ordinate system, while their slopes are different according to the numerical value of the index L/S.

For such "isomorph" herrings the simple proportion-formula for growth-calculation is strictly valid. Using the terms of the correlation routine, the possible paired observations of L and S for anyone particular individual will be perfectly correlated ($r = 1$), and the two regression lines, in other cases divergent, will coincide and be represented by the individual line.

The point-cluster comprising the terminating points of the individual lines (in our case P_{10}) can always be constructed so as to be identical with any given real point-cluster representing a set of observations. Concerning other clusters belonging to the same system of individual lines, for instance P_9 or P_2, they cannot as a rule be determined empirically (unless it should become possible to keep a shoal of herring in a tank and examine lengths and scale-sizes from time to time). But by considering the lengths l_1, l_2 etc., calculated by means of the simple proportion-formula, together with the corresponding scale-sizes s_1, s_2 etc., actually measurable on the scales, as sets of paired observations, a kind of semi-empirical point-clusters can be constructed, where the distribution of the s-values conforms strictly with actual observations, and where the assumption concerning the isomorphism is at the same time fulfilled. In how far these semi-empirical clusters will be similar to the corresponding, but unknown, actual clusters, will depend upon the degree of approximation with which the calculation formula renders the true lengths of the animals.

For these point-clusters correlation coefficients, regression lines etc. can be calculated exactly as if they represented real observations throughout. And now it happens that the regression lines for such clusters, as well as for actual end-point-clusters always intersect the co-ordinate axes at some distance from the origin and may be represented by equations of the form $L = a + bS$, or $S = c + dL$, where the constants a and c have definite values, for instance 14, 36 or 213 millimetres for a. These values prove to be in a very high degree dependent upon the distance of the point-cluster as a whole from the origin, and also, of course, of the "constellation" or "configuration" of the points in the cluster relative to each other. The configuration will determine the numerical value of the correlation coefficient and thereby also the divergence of the regression lines.

In addition to these real or semi-empirical point-clusters an infinite number of other clusters may be imagined. Very many of these cannot well be realised in nature, but an infinite number of them will be so similar to real, empirical clusters, that no one could have any suspicion about their reality, if they were not otherwise known to be constructions. They may be so arranged as to be considered as copies of most of the real frequency-distributions hitherto published with regard to the constant terms of the regression equations, the co-ordinates of the averages, the numerical values of the dispersion measures etc.

All these clusters may be most correctly regarded as real or imagined descriptions of a momentary situation or status of the individuals with respect to the characters observed, as something similar to a single picture cut out of a cinema film. If the time is not the decissive argument for the "exposure", but the passage of the individuals through one or the other developmental stage, for instance the stage of complete sexual maturity, or the stage, when the third summers growth commences on the scales, then new kinds of clusters arise. All these possible point-clusters have the common property that every individual is represented by one point in each of them, and further that one single cluster is really only descriptive of a kind of momentary situation (when these words are taken in a wide sense), a differential element of a greater unity, which consists of all individual lines.

A single individual line, on the other hand, also represents an element, but of another kind, of this greater unity, which may be termed the *total description.*

In fig. 1 is shown a very simple total description for ten individuals, with the end-point-cluster and a cluster for the commencement of a new summers' growth on the scales. For the purpose of constructing this figure the data for every tenth herring in a sample of 101 5 years old herrings were used. For the empirical end-point-cluster as well as for the semi-empirical „age-ring-cluster" the correlation coefficient and the regression line for average total length as a function of scale-size were determined and included in the graph.

A concideration of total descriptions as that pictured in the figure immediately leads to the conclusion that the regression lines for the point- clusters and the individual lines may well be very different but co-existing things, and that nothing at all prevents the simple proportion formula to be valid for the events which take place along any of the individual lines, while the regression lines, with their constant terms a, are a sort of shorthand descriptions of the point-clusters. But when such is the case, then the application of the regression equations to individual growth calculation will appear to lack a rational foundation and will have to be looked upon as erroneous.

This conclusion is not restricted to the most simple case, when the herrings are assumed to behave as isomorph bodies. The indivi-

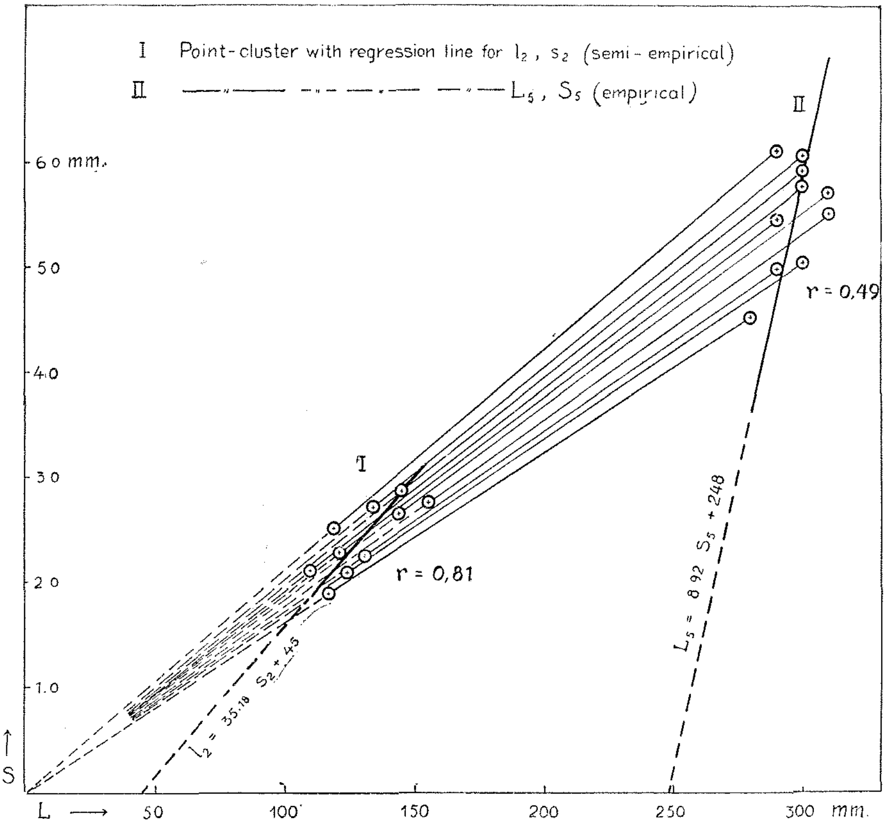

Fig. 1. Individual lines, empirical end-point-cluster and semi-empirical "age-ring-cluster" for the commencement of the second summers growth with regression lines for average total length as a function of scale-size. Material: sub-sample consisting of ten herrings out of a lot of 101 5 years old individuals belonging to a sample of Norwegian spring herring caught in 1917.

dual lines may, for instance, be assumed to intersect with the L-axis at a point outside the origin. This assumption conforms with a formula for individual growth-calculation having the form:

$$l = \frac{s}{S} (L \div a) + a,$$

where l denotes the length corresponding to a scale-size s. Even the assumption that the individual lines converge towards a common point on the axis may be exchanged for a new one that the intersection points are crowded around a certain mean-point on the axis a. s. o. It becomes clear that a point-cluster, otherwise termed a frequency distribution for L and S, is about as instructive for the proper understanding of

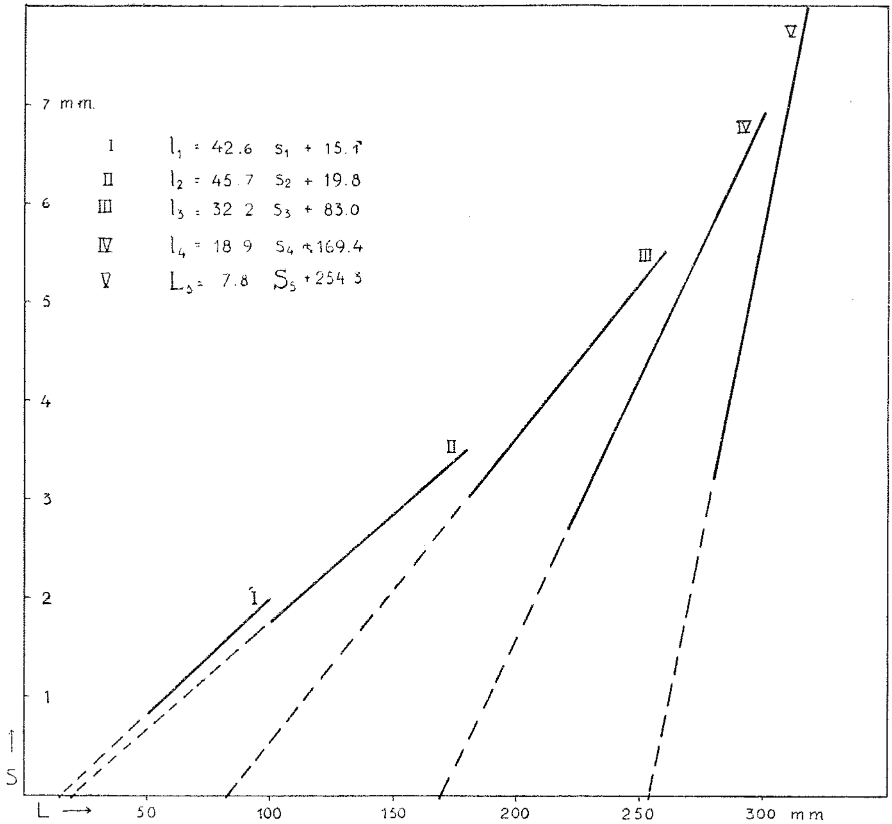

Fig. 2. Regression lines for average L as a function of S for one actual end-point-cluster and four semi-empirical age-ring-clusters. Material: the whole lot of 101 5 years herring before mentioned.

the problems under discussion as a single film picture is for the happenings in a film story. They both just give a situation, nothing more, and in both cases several trains of events may be imagined to lead up to that situation. Which of these is the actual one cannot then be decided.

In fig. 2 one of the regression lines for each of five connected point-clusters are shown. They represent one empirical end-point-cluster and four semi-empirical age-ring-clusters for the before mentioned 101 herrings.

From this graph is seen that the slope of the regression lines as well as their intersection points with the L-axis alter in a definite manner according to the distance of the clusters from the origin. If the lines are represented by equations of the form $L = a + bS$, then b will tend to decrease and a to increase with increasing distance of the clusters. This rule has a wide application also for empirical clusters of an approximate normal configuration, but it will break down if the clusters are widely different in this respect. For the proper understanding of this rule it is necessary to bear in mind that the frequency distribution for L has a tendency to acquire a narrower range with increasing age, when the annual growth-increment diminishes, while the same does not happen with the frequency distribution for S. The herrings grow, during the later years of their life, to a certain degree asymptotically towards a common maximal length, at the same time as the variation in relative scale size (expressed by the index S/L) remains essentially constant, just as does the variation in the relative size of the head, operculum or any other „morphological" character. This difference in the modus of variation of the two characters is to a great extent responsible for the alteration with increasing age in the configuration of the age-clusters. If the fish actually did grow to the same final size, then the regression line for average L as a function of S would run parallel with the S-axis. This ideal case is not realised in samples of real herrings or other fish, but the approximation is in samples of old, Norwegian herring, great enough for the effect to be very clearly observable. The numerical value of the correlation coefficient dwindles down from about 0.9 to 0.2, the constant b in the regression equation from about 45 to 4, while the constant term a increases from a value of about 30 to more than 300 millimetres.

For the case of old fish the situation can, perhaps, be most shortly described by saying that the age-clusters are at the same time approximate length-clusters, when by that term is understood such clusters, as are obtained by observing the L's and the S's for each individual herring at the moment, when it acquires a fixed, definite total length. In such specific clusters the L-S-points will lie on a straight line running parallel to the S-axis, and the L's will be completely independent of the S's. The Bravais formula will not be applicable for such clusters, as it gives an indefinite numerical value for r.

Age-clusters which are at the same time more or less approximate length-clusters may also arise as a consequence of natural or artificial selection. Norwegian herring, spawning at the early age of three years, will, for instance, form such naturally size-selected groups showing a low value of the correlation coefficient, while samples from catches made with highly selective gear may furnish examples of clusters exhibiting artificial size-selection.

Selective processes and asymptotic growth will thus account for low correlation and highly divergent regression lines. On the other hand will the correlation be high and the regression lines show little divergence in such cases, where the L's are distributed over a wide range. This will happen in a natural way, when the external conditions influencing the growth of the fish are widely different for the different members, or groups of members, of a population. Several fine examples of this kind may be found in the Norwegian material of observations. Similiar results will ensue from an artificial lumping together of observations from two or more natural but differing samples. In short, the course of the regression lines and the numerical value of the correlation coefficient may give clues to interesting interpretations of the samples, for which they are descriptive, but they will decidedly lead astray, when they are used for the determination of the properties of the individual lines, which are the sole requisite for the formulas for individual growth-calculation.

In comparison with these conclusions it becomes a matter of minor importance to mention that two point-clusters, partially covering the same space in the L-S-plane, and otherwise resembling each other will exhibit different average S-values for a definite L-value common to both clusters. The cluster nearest to the origin will furnish the lowest S-values. A whole series of such clusters embracing a definite common L-value will give a more or less regular progression of increasing average S-values, and if such a series is actually procured by observing during a period of time the mean values of S for a definite value of L in individuals belonging to two year-classes of a population of rapidly growing fish, then an apparent annual periodicity in these average values may be observed. The results of the ordinary treatment of such observations have been unwarranted conclusions in the direction that the scales of the fish grow faster than the total length, or that there are periodic discrepancies, or that there is no rigid connection at all in the growth of the two characters a. s. o.

Still more immaterial is the fact that some authors, in their search after the "relation between the scale-size and the total length" have treated their paired observations of L's and S's as if the task were to arrive at a method for calculating the size of the scale, when the total length of the fish is known for different life-epochs. The regression lines for the average values of S as a function of L are calculated or otherwise determined in these cases. However, no essentially greater harm has been thereby done. For the really wrong thing to do in this matter is to let the conclusions concerning this relation and concerning the formulas for individual growth-calculation depend upon an uncritically executed statistical treatment of the observations, whichever of the regression lines is chosen for the purpose.

LITERATURE.

I. Lea, E.: On the Methods used in the Herring Investigations, Publ. de Circonstance, No. 53, Copenhagen 1910.

II. Sverdrup, H. U.: Druckgradient, Wind und Reibung an der Erdoberfläche, Annalen d. Hydrographie u. Maritimen Meteorologie, 1916, H. VIII. Berlin 1916.