Some Remarks on the Graduation of Measurement Data.

By

H. J. Buchanan-Wollaston,
Fisheries Laboratory, Lowestoft.

This paper deals particularly with some methods of treating ichthyometric data. The methods are, however, applicable to many other cases of frequency distributions. They have special reference to the graduation of length-measurements of fish made by means of a pin on celluloid slips[1]), in which case the measurements are originally ungraduated and may be graduated subsequently in any way desired. The methods are, however, applicable in modified form to the case of measurements *made originally in class-intervals, for instance centimetre* intervals.

To avoid too theoretical a discussion it is proposed to explain the treatment of the data entirely by means of examples. The first examples given are taken from measurements of plaice made on celluloid slips. The fish were taken in an inshore trawl in early August (Aug. 1st—8th) 1928 on the inshore grounds near Beer in Devonshire. The plaice taken on each day were measured on one slip, with the exception of those taken in a few hauls which were for some reason or another not comparable with the remainder. Fig. (1 *a*) shows the distribution of the pin-pricks on the slip for $^3/_8$ 28. On all slips a line was drawn at 15 cms. from the block against which the noses of the fish were pressed. The frequency distribution of the pin-pricks is not easy to grasp when they are dispersed widely on the slips and therefore the pin-pricks were projected with a parallel ruler near to a line at right angles to the 15 cms. line by which each slip was orientated. Fig. 1 (*b*) to (*h*) shows the distri-

[1]) Journ. du Conseil, Vol. III, No. 3, p. 380.

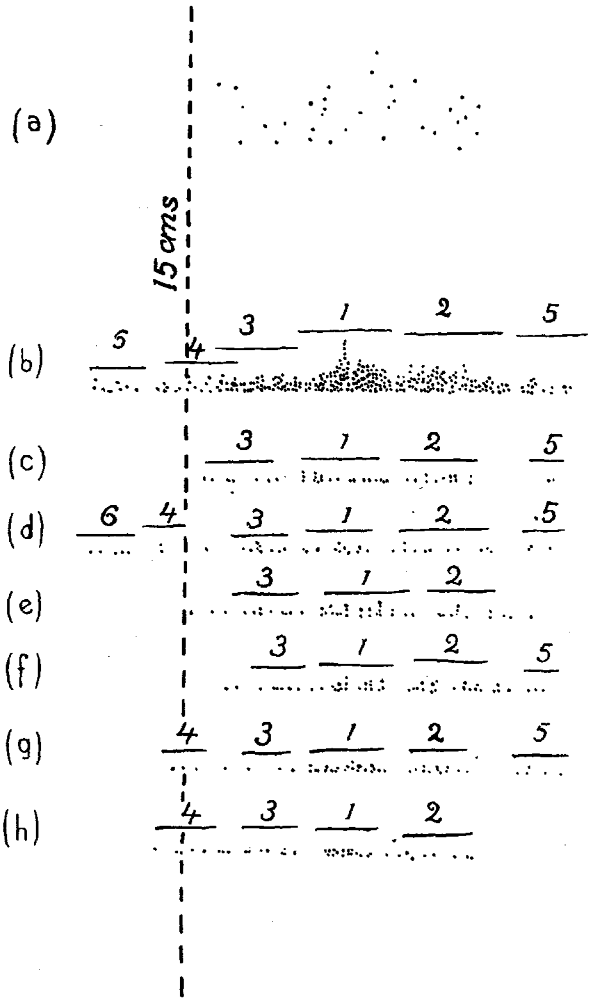

Fig. 1.

bution of pin-pricks after this projecting process, (*c*) to (*h*) giving the distribution for each day separately, (*b*) giving the total.

Before proceeding, it should be mentioned that only the measurements of the I-group[1]) are shown on these slips, those of the older fish being omitted. On a glance at Fig. 1 it is at once obvious that the measurements fall into subgroups, and cannot by any stretch of imagination be regarded as conforming to a simplex frequency distribution. The subgroups, which have been numbered 1 to 6 according to their number content, are indicated by horizontal lines. It is not, of course, possible to find the exact centre or the exact range of each subgroup, but the fact that a connecting line, varying only $^1/_4$ cm. from straightness can be drawn through the apparent centres of corresponding subgroups in each day's data seems sufficient proof of their reality[2]). Some of the subgroups seem to be again divisible into subgroups of a lower order, but there would appear to be no advantage to be gained by carrying analysis further than the main subgroups. It is necessary, however, to have regard for the main subgroups, for the following reasons. A subgroup is of such narrow spread that its position is given accurately enough even if it is represented by only a very few fish. Thus the positions of corresponding subgroups may be compared in different samples though all the subgroups may not be represented in each sample and may appear in quite different proportions in different samples. Consistency in position is the test of correspondence between subgroups. It has been suggested that the appearance of subgroups is due to errors of random sampling, but when there is consistency in their positions this cannot be the case. Errors due to random sampling certainly may account for the fact that the subgroups do not conform very nearly to a smooth frequency curve and may account in part for variations in their apparent representation. The main object of fitting a frequency curve to a group of measurements or drawing a smooth curve near the points corresponding to rough data is to smooth out irregularities due to random sampling. The curve is held to give a result more nearly in correspondence with that which would be obtained if the whole "sampled universe" were measured than that obtained from the rough points themselves. Given that the curve is well chosen and that the roughnesses of the data are really due to random sampling, this belief is in general justified by results. It is necessary, however, so to

[1]) As shown by otolith readings.

[2]) For a practical method for resolving such complex frequency distributions into their component sub-groups, see WOLLASTON & HODGSON, Journ. du Conseil, Vol. IV, No. 2, p. 207 *et seq.*

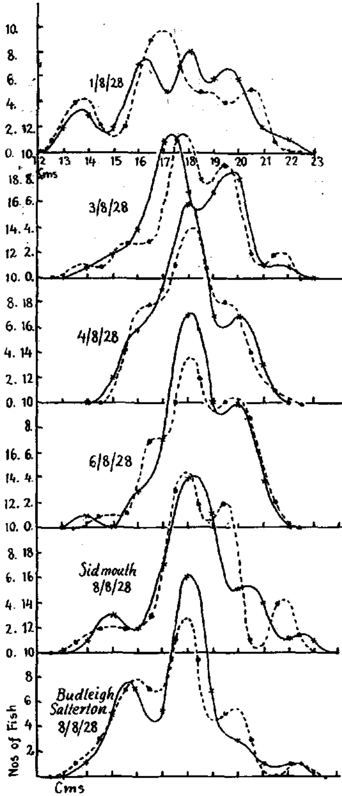

Fig. 2.

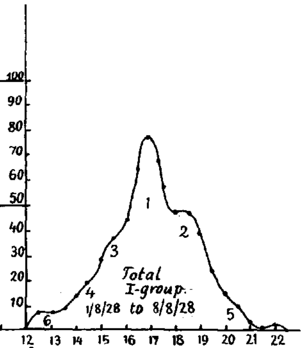

Fig. 2 a.

group the original measurements and to choose such a method of drawing a curve that consistent irregularities are not obscured.

If we have continuous and ungraduated original data as given on celluloid measuring slips we can graduate them in any desired manner. We may put them into half-centimetre, centimetre or other equal class-intervals or we may use *variable* intervals on the axis of abscissae. The curves in *continuous* line in Fig. 2 are drawn in as smooth a manner as possible through the points representing the data shown in Fig. 1 after grouping at each centimetre, that is to say all measurements between

14.5 and 15.5 cms. are grouped together and so on. We may change the limits of grouping, taking 15.5, 16.5 and so on as the centres. The resulting curves are shown in broken line in Fig. 2; the two sets of curves are fairly consistent on the whole as to the positions of the chief modes, though the proportionate representation of some of the subgroups is not in agreement in the two sets. The subgroups 1, 2, 3, 4 and 5 would be placed in the same order of magnitude whether we had one set of groupings or the other, and this is sufficient proof of their reality. The dots and crosses are really points on the same curve, but if we draw this curve we shall find that we begin to reproduce inconsistent irregularities in the data. Overlapping grouping carried to the limit will reproduce all the irregularities of the original data and this does not seem desirable unless large samples are being dealt with, in which case irregularities due to random sampling may be expected to disappear. Fig. 2*a* shows the curve for all the measurements shown in Fig. 1 (*c* to *h*). There are sufficient data here for the method of overlapping to be used, the points on the curve corresponding to the number of pin-pricks in centimetre intervals which overlap the adjacent intervals to the extent of half a centimetre. Since the original data are continuous, this is a true interpolation method and is not at all analogous to methods of smoothing which, in data of measurements originally grouped in centimetres, combine the data of several centimetre intervals into larger intervals overlapping each other. These latter methods are quite artificial, while the former is merely a means of using additional information inherent in the data themselves. It is useless to attempt interpolation by overlapping centimetre intervals in cases where the original data are grouped in centimetre intervals. The intermediate points on the resulting "curve" would of course lie on straight lines between the original points and not on a continuous curve at all. Fig. 2 (*a*) shows well the great advantage of having originally ungrouped data. It also indicates that, in cases where the data are originally grouped in cms., drawing a continuous curve through the centimetre points on the assumption that the group is complex and made up of subgroups fairly nearly symmetrical and of narrow range is likely to give a result far nearer the truth than either smoothing out the original "roughnesses" by drawing a continuous simplex curve near the points or joining them simply by straight lines. The last method is, of course, known to be wrong in every case. It often hides statistical facts, instead of showing them clearly which is supposed to be the main function of diagrams. It is undoubtedly one of the worst possible ways of graphing data and it is difficult to comprehend why it should be so persistently used in fishery research.

Before proceeding to discuss the question of grouping data in variable intervals, it is advisable to consider generally the question of growth and growth-rate. Assuming that a number of plaice-eggs are hatched simultaneously, the resulting fish, growing up under natural conditions, will not all have the same length after a year, but these lengths will form a frequency-curve. The curve will be unimodal if there are a large number of variable subjective and objective conditions which influence growth, but of which the variation is continuous and such that the fish may all be said to have lived "under the same conditions" during the period. To attempt to define more precisely this somewhat vague phrase might introduce mere mathematical abstractions without a counterpart in Nature. Growth of a group of fish during the first year is thus what we may term divergent. The lengths of the larvae on hatching would spread over only two or three millimetres at most. The lengths of the same fish after a year's growth would undoubtedly spread over at least two centimetres. Thus to obtain a frequency-curve for larvae we should have to group the measurements in millimetres or half-millimetres; to obtain a corresponding curve for the year-old fish, grouping would have to be in centimetres or half-centimetres. We may now make another definition, namely that of "parallel growth" of a group of fish. In "parallel growth" the result at the end of the period is the same as if every fish had grown the same amount in the given period, thus, in our hypothetical group, the measurements at the end of a year would still have to be grouped in millimetres or half-millimetres to form a frequency-curve. The third type of growth is the convergent or compensatory type. In this type, at the end of any period the group has less "spread" than at the beginning of the period. We are not referring at the moment to growth-type pertaining to a particular fish. In that case divergent growth may be defined as growth persistently greater or persistently less than the average. It is quite clear that, though every fish in a *group may indulge, during any subperiod, in any one of the three* types of growth, the growth of the group as a whole must be divergent. Anything else would be absurd in the period of free growth, before the onset of maturity introduces a point of discontinuity. It is impossible of course to trace the growth of an individual fish under natural conditions by observations on the fish itself. We may, however, trace the life-history of a plaice by calculation from the otolith, by a method corresponding to that used in calculating the growth of a herring from its scales. This use of otoliths has been tested in the case of about 50 fish which were marked in late summer and autumn in the Poole district of Dorset and the Beer district of Devon and which were recaptured at various times in the year after marking. At the time of year at which the fish were marked the edges of otoliths show, nearly always, dark growth. In every case the calculated position of the edge of the otolith on marking proved to be in the "dark growth" of the preceding year or just on the edge of the white growth, thus, exactly where one would expect it to have been in reality. This method of course does not assume that the size of

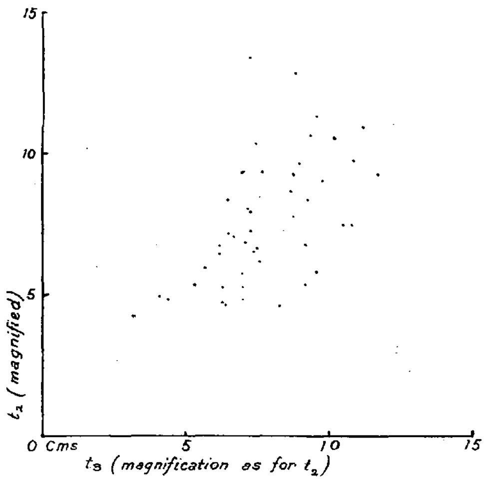

Fig. 3.

an otolith is proportional to that of the fish to which it belongs. This proportion varies quite widely owing to differences in shape of otoliths. The method assumes only that in each fish the otolith grows, in length as measured from the nucleus, proportionally to the fish. The average length of the otoliths has been found to be approximately a linear function of that of the fish, in spite of the variation in shape. It thus appears to be quite safe to assume that in most cases the otolith gives a picture of the past growth of a plaice, the picture being probably as reliable as that given by scale readings in the case of the herring. As compared to such scale-calculations, otolith calculations show the advantage that their reliability may, in any district, be tested on fish marked in autumn or winter. This question will be dealt with much more fully in a subsequent paper. It is introduced here merely because it provides a means for testing whether growth, in a given sample of fish, has been divergent, in the sense that growth in the individual fish has been on the whole divergent.

Calling the length-increment in the first year of life t_1, that in the second year t_2 and so on, as in the corresponding herring-scale work, we may compare these increments and examine whether a large "t" in one year is correlated with a large "t" in another year or the reverse. Thus we may find out whether the growth in the years dealt with has been divergent, parallel or convergent. It is of little use attempting to correlate t_1 with a subsequent increment, since the value of t_1 depends so very much on the time at which the corresponding fish was spawned. The spawning-season extends over four months, and this spread is quite sufficient to give a very wide "range" to t_1. Correlation between t_2 and t_3 should, however, give a useful picture. If positive, growth in the 3rd and 4th years of the life of the fish is divergent, if absent, growth is parallel, if negative, growth is convergent. Fig. (3) shows the correlation between t_2 and t_3 in the case of all the otoliths which were preserved after being taken from fish of 3 years old and upwards during the years 1924 to 1927 in the Poole district. Selection of these otoliths for preservation was quite haphazard since the idea of using them for growth calculations only arose last year (1929).

Fig. (3) shows that correlation between t_2 and t_3 is in this case quite definitely positive, and therefore that individual growth is divergent in general. Fig. (4) shows l_1, l_2, and l_3 for the same collection of plaice[1]). In this figure m represents the median, q_1 and q_2 the quartiles. In Fig. (5) the total range and the interquartile range of l_1, l_2 and l_3 are given compared to the median lengths. It will be seen that the interquartile range is proportional to the median length. The total range in l_3 may well be affected by the onset of maturity in the fourth year in the case of fish with an abnormally large l_2. Therefore the variation of the interquartile range is probably a more reliable index of increasing spread of a group than the total range of l_1 and l_2.

In cases such as the above, *the sizes of the class-intervals should be proportional to the size of the variates.* If a one-centimetre interval is used for fish of 15 cms., a 2-cm. interval should be used for those of 30 cms., a half-centimetre interval for fish of 7 to 8 cms. and so on. If this is not done, it cannot be expected that only modes or subgroups of the same order will be shown in curves of different year

[1]) l_1, l_2, l_3 = calculated length of fish at first appearance of 1st, 2nd and 3rd white ring on otolith.

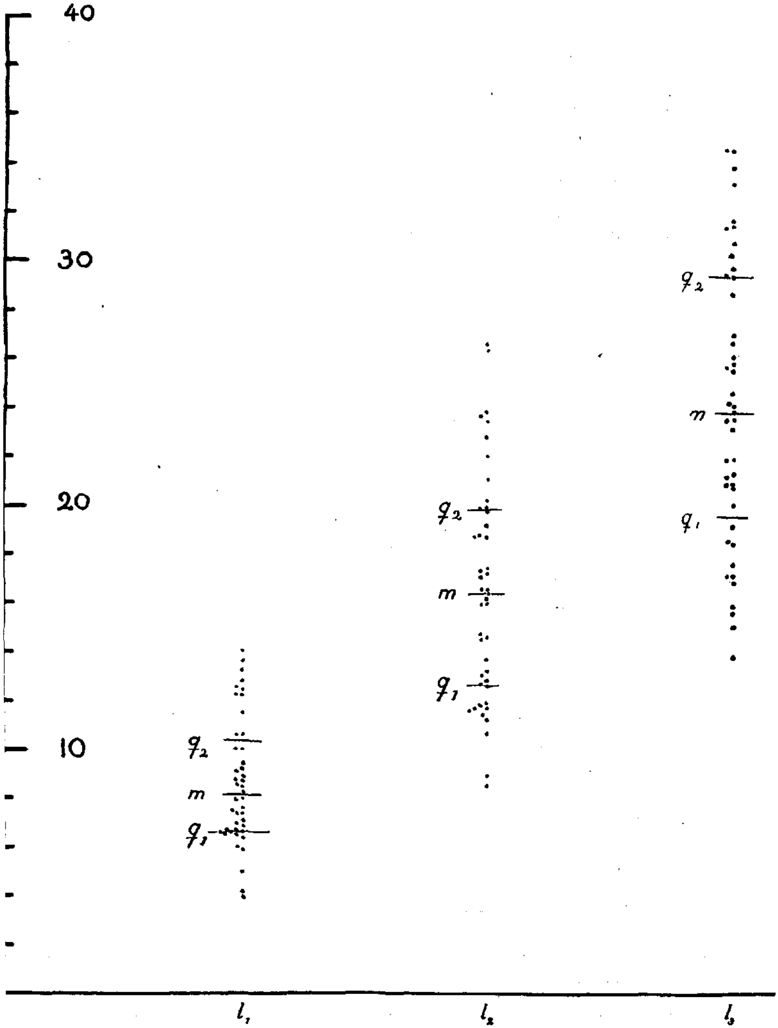

Fig. 4.

classes to be compared. If the modes in the I-group are considered significant for comparison, there will be too few modes in the 0-group, too many in the II-group, and any law there may be connecting the modes in the groups will be obscured.

Table (1) gives the end-points and midpoints of class-interval scalculated on the hypothesis that a 1-cm. interval is correct for fish of 15 cms. the other class intervals being proportional to the length of the variates. The ordinates below the 15-cm. point are smaller than those for equal class-intervals, and above that point are larger. Therefore, if

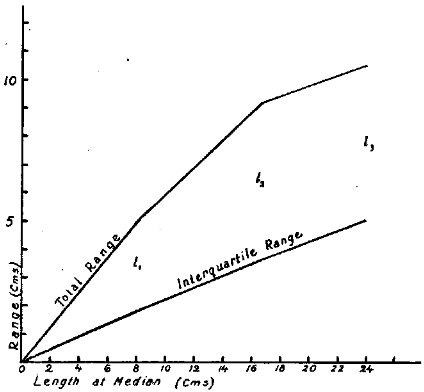

Fig. 5.

it is desired to compare a curve drawn from proportionally grouped data with a curve drawn from equally-grouped data, each class-frequency in the former must be multiplied by the reciprocal of the length of the class interval measured in centimetres. If this is not done the two curves will have different areas. The reciprocals are given in Table (1). The end points are calculated as follows. The end-points of the class-interval 15 are 14.5 and 15.5. Thus any end-point, say "$n+1$" is given by multiplying the end-point "n" by $\frac{15.5}{14.5}$ or 1.069. Any end-point "$n-1$" is obtained by multiplying "n" by $\frac{14.5}{15.5}$ or .9354[1]). If the original data

[1]) The continuous use of a decimal to four places only gives rise to an error which becomes serious in the upper intervals. Table 1 was calculated, with a mechanical calculating-machine, using the greatest number possible of significant figures.

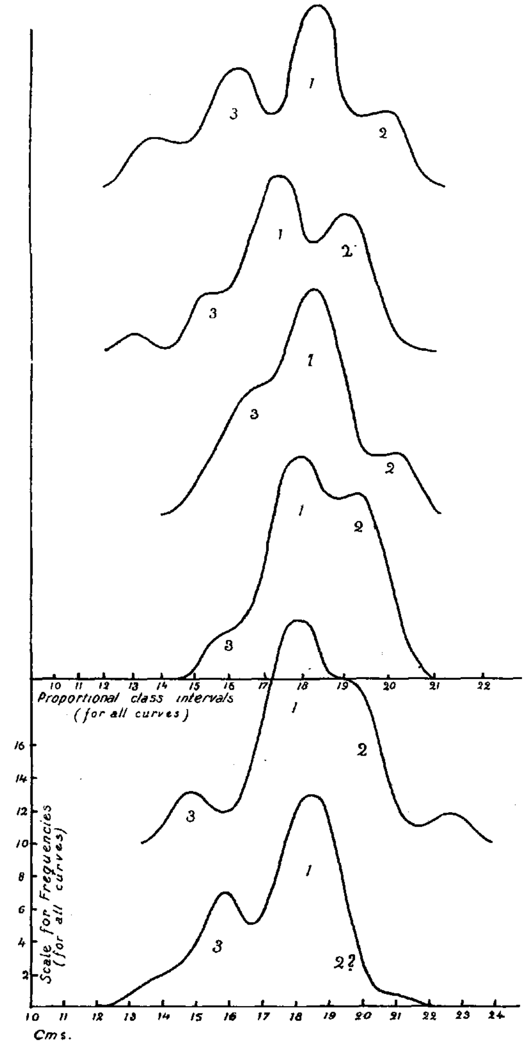
1
2
3
10
11
12
13
14
15
16
17
18
19
20
21
22
Proportional class intervals
(for all curves)
Scale for Frequencies
(for all curves)
2?
23
24
Cms.

Fig. 6.

are ungrouped we may use the overlapping method of grouping, as was done in the case of centimetre-grouping, by taking, as intermediate ordinates, the data falling between adjacent mid-points in the proportional series. By continued overlapping, every character of the original collection of data will be reproduced in the curves.

Fig. (6) shows curves drawn from the data given in Fig. 1, grouped

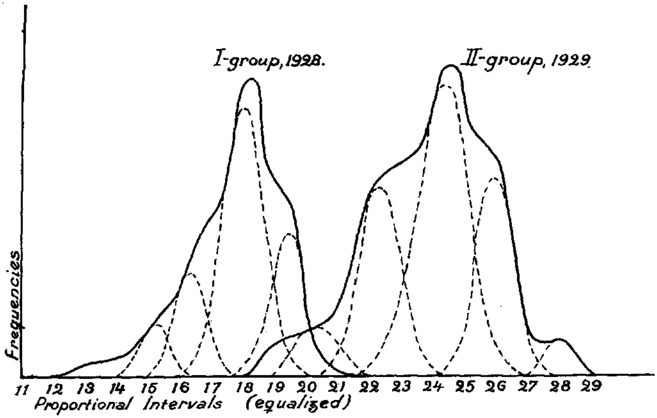

Fig. 7.

in proportional intervals. The positions and relative richness of the subgroups 1, 2, and 3 are well shown in the curves, but the smaller subgroups 4, 5, and 6 are apt to become included in the ends of the larger groups. Whether this matters or not depends on the purpose for which the curves are drawn. If this purpose is the comparison of the n year old fish of one year with the $(n + 1)$ year old fish of the next year with the object of deducing growth between the two years, it is probably sufficient to limit ourselves to the richer subgroups and to neglect the "tail-groups". These in any case form but a very small proportion of the whole collection.

In comparing the I-group with the II-group, using proportional class-intervals, it is assumed, *ipso facto*, that the range of each group is proportional to the size of the variates. If that assumption is correct, for the purpose of comparison the class-intervals should be regarded all as equal, and the ordinates should not be multiplied by the reciprocals. Fig. (7) shows such a curve for all the fish of the I-group measured in the

Beer district in August 1928, compared with that for all the fish of the II-group measured there in August 1929. There were not nearly so many fish of the II-group as of the I-group, therefore the curves have been drawn on different scales, chosen so as to make the highest ordinate of each curve about equal. In drawing the I-group curve we may be guided by our knowledge of the component subgroups obtained from examination of the original data or of the "overlapping centimetre" curve Fig. (2*a*). The II-group curve cannot reasonably be drawn except as shown, having regard to the fact that all frequency-groups so far examined, pertaining to a year-class of plaice from either the Poole or the Beer district, resolve into component subgroups approximately symmetrical and of narrow range. It will be seen then that the subgroups of the I-group 1928 are all represented in the II-group 1929, and their comparative representation, their order of importance, is the same in the two curves. The spread of the II-group curve is, however, still greater than that of the I-group curve. If the same population is being sampled in both cases, we are forced to the conclusion that the increase of spread with size is even greater than directly proportional. It may be, however, that the net did not take all the fish on the lower side of the I-group in 1928. When they enter the II-group these fish will of course be properly sampled. This method gives what would appear to be a reliable method of growth-determination. The main mode in the I-group is at 18.3 divisions (= 18.645 cms.), and that in the II-group at 24.6 divisions (= 28.344 cms.) the growth between August 1928 and August 1929 is thus 9.7 cms. This method of estimating growth assumes, of course, that the same population is being dealt with in both cases, but the similarity in shape of the two curves appears to indicate that at least the main mode in one curve corresponds to that of the other. The modes are much better suited for growth-rate calculations than are the means. The central mode would be very little changed as to position by the addition of a large number of fish of subgroup 3, while the mean would be much reduced by the same addition. If it is required to deduce the growth rate of the whole of the I—II-group represented by the samples, the curves may be resolved into components as shown in broken line in Fig. (7). This diagram is only intended as an indication of the way in which the problem may be attacked. Strictly speaking, if these component curves are symmetrical when the variates are grouped in equal class intervals, they should be positively skew[1]) when the variates are grouped in proportional intervals "equalized". To resolve the whole, however, into the most

[1]) If the mean is greater than the mode, skewness is termed positive and *vice versâ*.

probable skew components is an elusive problem at best. It may be preferable to carry out the resolution on curves drawn on the original "unequalized" proportional intervals, and to assume that they are symmetrical in the absence of other indications. It should be mentioned, though, that in cases where the individual variates themselves show proportionally divergent growth, positive skewness in the frequency curve is produced. It is quite possible that such positive skewness is what may be expected in frequency-curves of fish measurements. The solution of such a problem, however, requires much wider research than has been undertaken up to the present time. In a case such as the present,

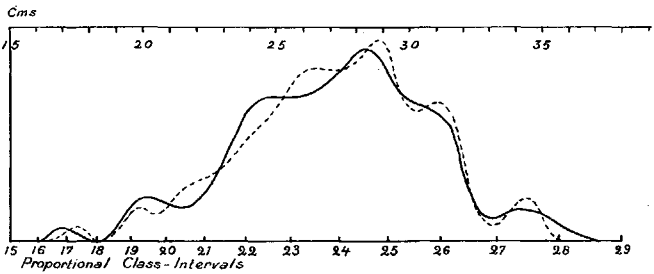

Fig. 8.

the subgroups are of such small range that failure to assign the correct skewness would have little effect on the indicated positions of the modes. These are all we require for growth-calculations on the basis of comparison of frequency-curves.

In such growth-calculations it is usually necessary to use proportional class-intervals for the reasons already stated. This is illustrated in Fig. (8). This figure shows, in unbroken line, the II-group of 1929, in proportional class-intervals unequalized, the ordinates being multiplied by the corresponding reciprocals; the broken line is the frequency-curve for the same data in centimetre class-intervals. The latter curve shows four closely placed modes on the left-hand side.

The majority of fish measurement data are originally grouped in equal class-intervals. The rigid method of proportional grouping cannot be used with such data, but it may be modified to meet such cases. We may assume that each frequency is equally distributed throughout its class-interval. This is straining Nature no more, indeed much less, than the usual assumption that it is concentrated at the centre of the class-interval. We may now allot the appropriate proportions of each frequency

Table 1. Proportionate Intervals (15 = 15 cms.).
End-points, Centre-points, Reciprocals of Lengths.

Interval	Ranges Cms.	Reciprocal of Length of Interval in cms.	Midpoint Cms.
0	5.33— 5.70	2.70270	5.52
1	5.70— 6.09	2.56410	5.90
2	6.09— 6.51	2.38095	6.30
3	6.51— 6.96	2.22222	6.74
4	6.96— 7.44	2.08333	7.20
5	7.44— 7.96	1.92308	7.70
6	7.96— 8.50	1.85185	8.23
7	8.50— 9.09	1.69492	8.80
8	9.09— 9.72	1.58730	9.41
9	9.72— 10.39	1.49254	10.06
10	10.39— 11.10	1.40845	10.75
11	11.10— 11.87	1.29870	11.49
12	11.87— 12.69	1.21951	12.28
13	12.69— 13.56	1.14943	13.13
14	13.56— 14.50	1.06383	14.03
15	14.50— 15.50	1.00000	15.00
16	15.50— 16.57	0.93458	16.04
17	16.57— 17.71	0.87719	17.14
18	17.71— 18.93	0.81967	18.32
19	18.93— 20.24	0.76336	19.59
20	20.24— 21.63	0.71942	20.94
21	21.63— 23.13	0.66667	22.38
22	23.13— 24.72	0.62893	23.93
23	24.72— 26.43	0.58480	25.58
24	26.43— 28.25	0.54945	27.34
25	28.25— 30.20	0.51282	29.23
26	30.20— 32.28	0.48077	31.24
27	32.28— 34.51	0.44843	33.40
28	34.51— 36.89	0.42017	35.70
29	36.89— 39.43	0.39370	38.16
30	39.43— 42.15	0.36765	40.79
31	42.15— 45.06	0.34364	43.61
32	45.06— 48.16	0.32258	46.61
33	48.16— 51.48	0.30120	49.82
34	51.48— 55.04	0.28090	53.26
35	55.04— 58.83	0.26385	56.94
36	58.83— 62.89	0.24631	60.86
37	62.89— 67.23	0.23041	65.06
38	67.23— 71.86	0.21598	69.55
39	71.86— 76.82	0.20161	74.34
40	76.82— 82.12	0.18870	79.47
41	82.12— 87.78	0.17668	84.95
42	87.78— 93.83	0.16529	90.81
43	93.83—100.30	0.15456	97.07
44	100.30—107.22	0.14451	103.76
45	107.22—114.62	0.13514	110.92

Table 2. Length-Measurements of Plaice of the II-group 1929, originally in cm. groups, showing Allotment to Proportional Intervals.

Cms.	No.	Prop. Ints.	Proportions of cm. inst.	No to nearest .5	Nos. of originally ungrouped data
18	0	18	$.39 \times 17 + .93 \times 18$	0	0
19	2	19	$.07 \times 18 + 19 + .24 \times 20$	$2 + .5 = 2.5$	3
20	2	20	$.76 \times 20 + .63 \times 21$	$1.5 + 2.5 = 4.0$	4
21	4	21	$.27 \times 21 + 22 + .13 \times 23$	$1 + 5 + 1 = 7.0$	6
22	5	22	$.77 \times 23 + .72 \times 24$	$5.5 + 6.5 = 12.0$	15
23	7	23	$.28 \times 24 + 25 + .43 \times 26$	$2.5 + 12 + 5.5 = 20.0$	18
24	9	24	$.57 \times 26 + 27 + .25 \times 28$	$7.5 + 13 + 4 = 24.5$	23
25	12	25	$.75 \times 28 + 29 + .20 \times 30$	$11 + 11 + 2 = 24.0$	24
26	13	26	$.8 \times 30 + 31 + .28 \times 32$	$8 + 9 + .5 = 17.5$	19
27	13	27	$.72 \times 32 + 33 + .51 \times 34$	$1.5 + 2 + 1.5 = 5.0$	4
28	15	28	$.49 \times 34 + 35 + .89 \times 36$	$1.5 + 0 + 0 = 1.5$	3
29	11	29		0	0
30	10			118	119
31	9				
32	2				
33	2				
34	3				
35	0				
36	119				

to the corresponding proportional class-interval. Thus the proportional interval "15" contains half the frequency of the 14 cm. class-interval, half that of the 15 cm. class-interval, the proportional interval 29 contains .11 of the frequency of the 36 cm. class-interval plus the whole frequencies of 37 and 38 cms. + .43 of the 39 cm. interval.

The result cannot be expected to be so near to reality as that obtained from ungrouped original data. By its use, however, we do obtain the benefit of allowance of increasing spread of a group with increase in size of the variates.

It will be interesting to test this modified method on the figures illustrated by Fig. (8). Table (2) shows the working, and compares the result with the figures in the original groups, from which Fig. (8) was constructed, but before multiplication by reciprocals: it will be seen that the figures given by the modified method agree fairly well with the original figures. The position of the three chief subgroups is indicated fairly well if the draughtsman remembers to draw the curve on the assumption that it is made up of components nearly symmetrical and of narrow spread. The curve is certainly superior, for the purposes of comparison with other curves, to the centimetre curve. It shows, however, that our original assumption that a frequency is equally distributed in its class-interval has an unnaturally "smoothing" effect and serves to emphasize the great superiority of ungrouped original data over originally grouped data.