THE USE OF PROBABILITY PAPER FOR THE GRAPHICAL ANALYSIS OF POLYMODAL FREQUENCY DISTRIBUTIONS

By J. P. Harding, Ph.D.
British Museum, Natural History

(Text-figs. 1–6)

The mathematical analysis of bimodal distributions is very complex. Karl Pearson (1894) investigated the problem and developed equations for the purpose; but found them unsolvable as the 'majority [of the relations] lead to exponential equations the solution of which seems more beyond the wit of man than that of a numerical equation even of the ninth order'. He did indeed evolve an equation of this order and used it to analyse a few bimodal distributions, but the arithmetic involved was very laborious. Later he (Pearson, 1914) gives a table for 'Constants of normal curve from moments of tail about stump' which, as he describes in the introduction, occasionally permits a rough analysis of a distribution which is known to be bimodal. This method is much more rapid than the solution of the nonic equation, but 'owing to the paucity of material in tails and corresponding irregularity there will be large probable errors'. Gottschalk (1948) discusses the question and shows that in the special case where the bimodal distribution is symmetrical comparatively simple solutions can be found.

The purpose of this paper is to describe by a series of examples a straightforward graphical method which enables one to analyse not only bimodal distributions, both symmetrical and asymmetrical; but also more complex distributions. The last example will show how a distribution comprising three unequal and overlapping populations can be analysed, an estimate being obtained of each separate mean, each separate standard distribution, and the relative proportions in which the three populations are mixed.

The method makes use of probability graph paper[1] devised by Hazen (1913) for the analysis of the flow of water in rivers. The use of this and other probability papers for various engineering and industrial purposes has been described by Rissik (1941), Doust & Josephs (1941–42), Levi (1946) and others; but I know of no description of its use for the analysis of bimodal or polymodal distributions. Perhaps the methods I am describing are so obvious to mathematicians familiar with probability graph paper, that for them description is superfluous. To the biologist, however, the simplicity of this

[1] Obtainable from Messrs Wightman Mountain, London, as 'Data sheet No. 37. Arithmetic Probability'. A similar paper is also stocked by H.M. Stationery Office.

powerful tool is perhaps its greatest attraction. He has long been aware that the mean and standard deviation of populations he is confronted with are often of little biological significance; because these populations are compounded of individuals belonging to the two sexes, to different species, or to different age-groups, and are therefore necessarily bimodal or polymodal in character.

Hazen's probability graph paper (Fig. 1) has along the bottom a scale of percentages reading from 0·01 % on the left to 99·99 % on the right (Fig. 4 shows the complete ruling). Along the top of the paper the same scale reads from right to left. The scales are not evenly spaced; but are much more crowded in the middle than at the sides, being specially arranged so that when any normally distributed population is plotted, in a manner which will shortly be described, the points all fall on a straight line. The position of this line is determined by the mean and its slope by the standard deviation, and these statistics may be estimated without laborious calculations. If a bimodal or polymodal distribution is compounded of distributions which are themselves normally distributed, and with biological data this is generally true enough for practical purposes, it will give a curve when plotted which is the resultant of two or more straight lines.[1] These lines are usually not difficult to find and the irrespective positions and slopes give the means and standard deviations of the component populations.

Fig. 1 shows the analysis of 1572 cuckoos' eggs each measured to the nearest half millimetre. The histogram to the left of the figure gives the size distribution; there are for example 156 eggs between 20·75 and 21·25 mm. in length. The histogram is not part of the method, but is included to facilitate description.

The data are plotted as cumulative percentages. Take, for example, the point P indicated by a small circle: here 4·7 % (scale along the bottom) of the sample is less than 20·75 mm. in length, for there are $1+3+30+39=73$ eggs or 4·7 % of 1572 eggs shorter than this. Alternatively, 95·3 % (scale along the top) of the sample of eggs exceed 20·75 mm. in length. When all the points are plotted it is found that they lie approximately in a straight line AB. This line represents a normal distribution whose mean length is the length corresponding to the point M at which the line cuts the vertical for 50 %. i.e. 22·35 mm. The standard deviation is estimated from the points S_1 and S_2 where the line AB cuts the verticals for 15·87 and 84·13 % respectively, S_1 corresponds to a length of 21·35 mm. and S_2 to 23·4 mm. and the standard deviation is half the difference $(23{\cdot}4-21{\cdot}35)/2=1{\cdot}025$ mm. This is because 15·87 % of any normally distributed population is less than the mean by an amount equal to the standard deviation or more, and another 15·87 % exceeds the mean by this amount. The vertical lines for 15·87 and 84·13 % are not

[1] Apart from 'Normal' distributions the ones most frequently encountered with biological data are 'log-normal' and Poisson distributions. Log-normal distributions may be handled either by plotting the logarithm of the measurement instead of the actual measurement, or by the use of logarithmic probability paper, and Doust & Josephs (1941–42) describe a probability paper which is specially designed for Poisson distributions.

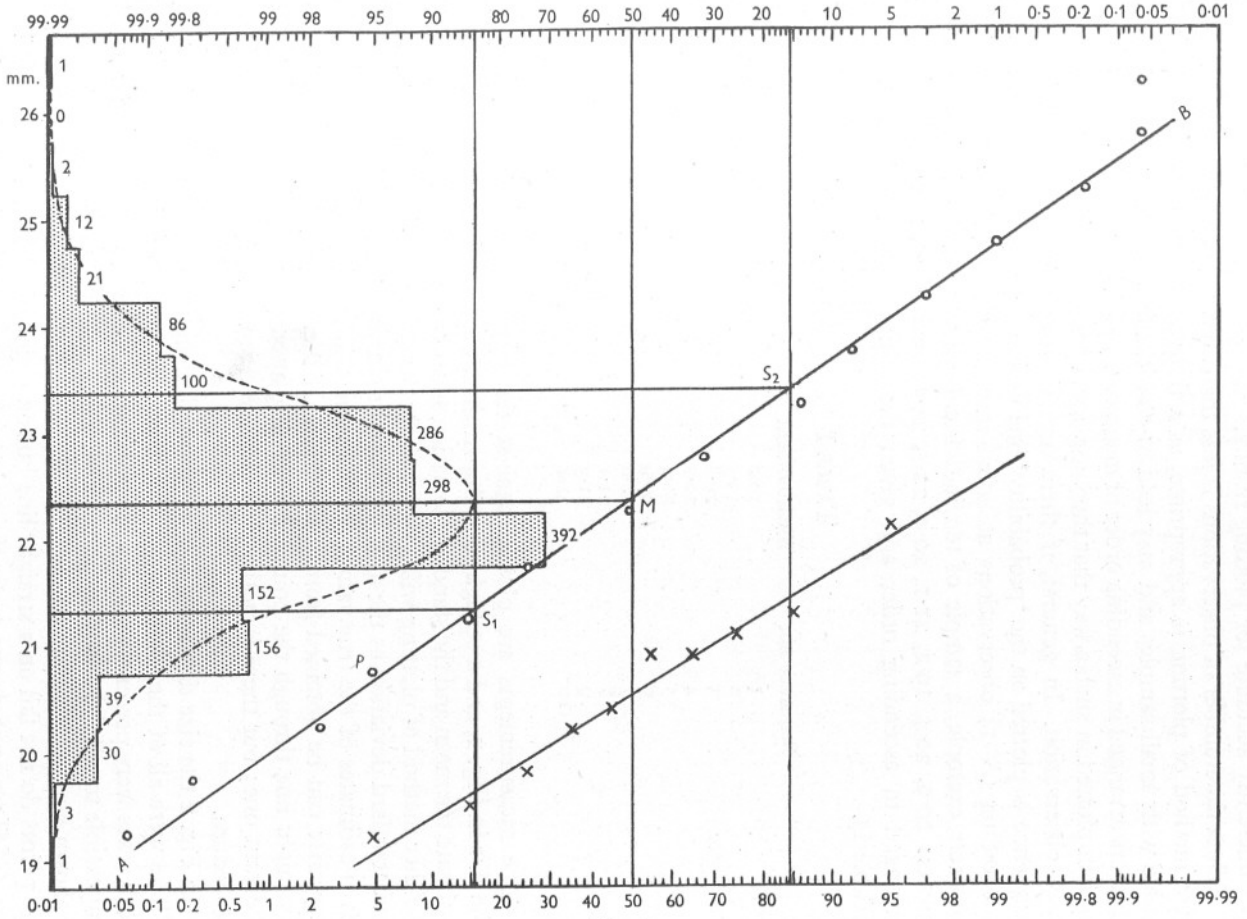
99·99 99·9 99·8 99 98 95 90 80 70 60 50 40 30 20 10 5 2 1 0·5 0·2 0·1 0·05 0·01
mm.
26
25
24
23
22
21
20
19
1
0
2
12
21
86
100
286
298
392
152
156
39
30
3
1
S2
S1
M
P
A
B
0·01 0·05 0·1 0·2 0·5 1 2 5 10 20 30 40 50 60 70 80 90 95 98 99 99·8 99·9 99·99

Fig. 1.

marked on the paper as supplied; but are easily put in. A 10 in. slide-rule is sufficiently accurate for plotting results.

When the number of observations is less than about twenty-five an alternative method of plotting is appropriate, as a frequency tabulation is of little value with small samples and may lead to considerable grouping errors. The data are arranged in ascending order of magnitude, each individual observation is then plotted in such a way that there are equal percentage intervals between each observation. In general, if there are N observations, the first in the sequence is plotted on the 'probability' line whose value is $50/N$ %, and the succeeding $(N-1)$ observations at equal intervals of $100/N$ %. To give a concrete example, a sample of ten individuals gave the following measurements: 20·9, 20·4, 19·2, 21·1, 20·9, 22·1, 21·3, 19·8, 19·5, 20·2. These are arranged in ascending order and given the appropriate percentage value (Table I).

TABLE I

Sequence no.	Measurement	Percentage
1	19·2	5
2	19·5	15
3	19·8	25
4	20·2	35
5	20·4	45
6	20·9	55
7	20·9	65
8	21·1	75
9	21·3	85
10	22·1	95

The measurements are plotted against the percentages as indicated by the X's in Fig. 1, and a straight line drawn through the points enables one to estimate the mean and the standard deviation as 20·5 and 0·9 mm. respectively. Neither method of plotting will give precisely the same estimates of the mean and standard deviation as the arithmetical method; but the graphical solution gives estimates of the true values for the population which are as reliable as any that can be obtained from the sample. If there is difficulty in placing a straight line through the points, it is at once apparent that the sample is an aberrant one; but there is nothing to indicate this with the usual arithmetic procedure.

In Fig. 2 the size distribution of forty-one immature copepods is analysed. These were all of the same species and all in the same developmental stage; both sexes were present, probably in approximately equal numbers, but it was impossible to tell the sex of any individual. The size distribution of the sample is shown by the histogram on the left. When plotted on the probability paper the points do not fall on a straight line but on a sigmoidal curve. The dotted line *GLH* was not drawn to fit the points plotted from the data, but is the resultant of the two straight lines *CD* and *EF*. The lines *CD* and *EF* were found by assuming that the distribution was a bimodal one due to the

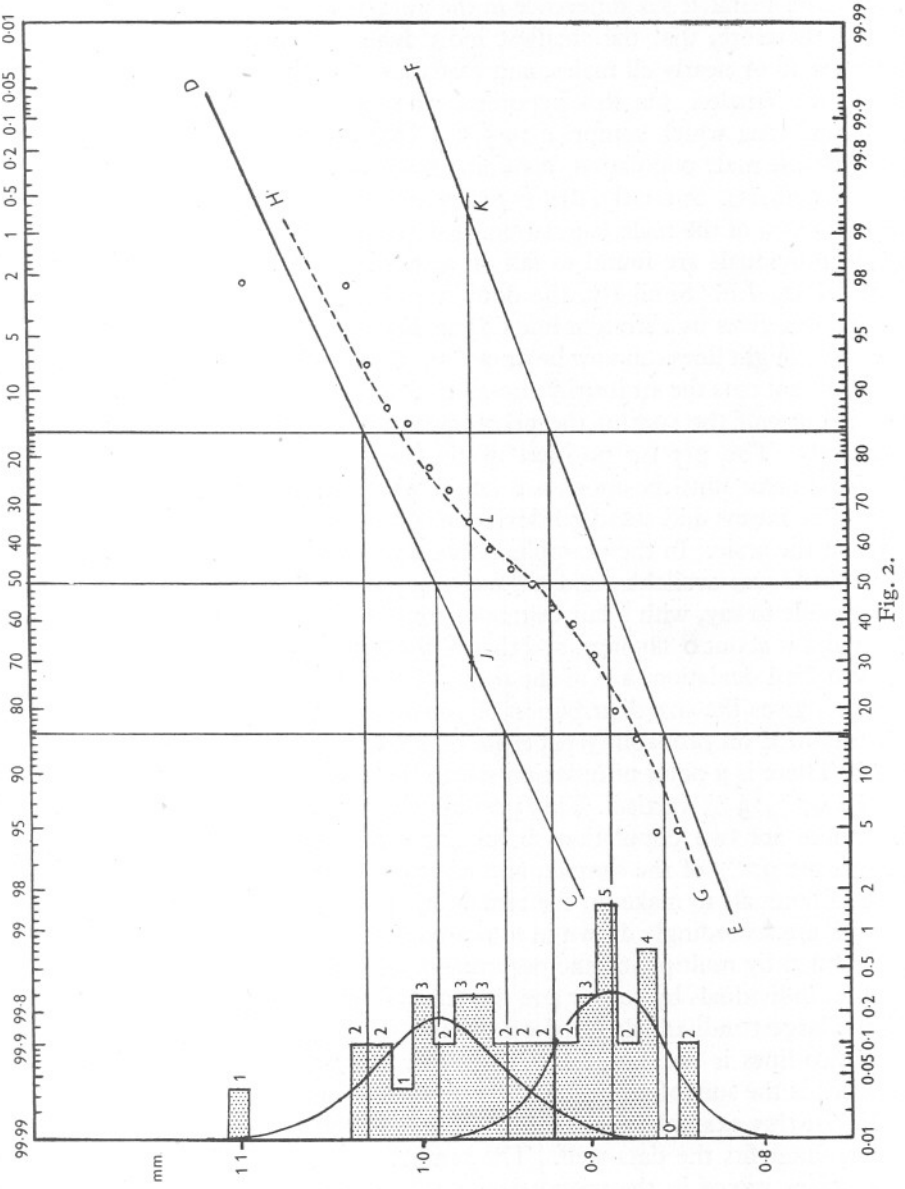

Fig. 2.

difference in the average lengths of the males and females. It is known that the adult females are distinctly larger than the adult males and it is reasonable to assume a similar sex difference in the immature specimens measured. It is likely, therefore, that the smallest individuals, say those less than 0·9 mm. long are all or nearly all males, and that those longer than about 1·0 mm. are nearly all females. On this hypothesis the eight male individuals less than 0·88 mm. long which comprise 19·5 % of the total population will comprise 39 % of the male population (assuming that males and females are present in equal numbers). Similarly, the six males less than 0·72 mm. long are estimated to be 29·3 % of the male population. A few points plotted in this way for the small individuals are found to fall on a straight line which is approximately that drawn, *EF*. Similarly, the doubling of the percentages for the largest size groups gives us a straight line *CD* or thereabouts. The resultant, *GLH*, of the two straight lines can now be drawn in. For example, the position *L*, where the resultant cuts the arbitrarily chosen horizontal line *JK*, is placed at 64·7 %; this being half the sum of the percentages for *J* and *K*, 30 and 99·4 % respectively. The precise position of the two lines should be adjusted by trial and error until positions are found whose resultant curve best fits the data. The means and standard deviations of the two populations can then be read off the scale. In the example, although only a small sample of forty-one individuals was available, and although no single individual could be sexed we are able to say, with a fair degree of confidence, that the average length of the males is about 0·885 mm. and that of the females about 0·99 mm. and that the standard deviations are of the order of 0·375 mm.

Fig. 3 gives the size distribution of 360 adult female copepods. The points when plotted on probability paper lie on an asymmetrically placed sigmoidal curve. There is a point of inflexion where the direction of curvature changes, on the 97 %:3 % vertical. This position for the point of inflexion suggests that there are two populations involved: a population of small individuals comprising 97 % of the sample, and mixed with them a small population of large individuals to make up the remaining 3 % of the sample. The lines *ST* and *QR* are accordingly drawn in to represent these two populations. The line *ST* is fitted by multiplying the percentages (read on the bottom scale) for the smallest individuals by 100/97 and the line *QR* by multiplying the percentages for the largest individuals (read on the top scale) by 100/3. The resultant of these two lines is the dotted line *UV*. The percentage at *Z* on this line, for example, is the sum of 3 % of the percentage at *X* and 97 % of the percentage at *Y*. In this example adjustment of the position of *QR* is unnecessary as the resultant fits the data well. The sample is analysed as representing two populations mixed in the proportions 97:3, with mean lengths of 1·223 and 1·477 mm. respectively, the standard deviations of both populations being about 50 μ.

Before leaving Fig. 3 I should like to draw attention to the abnormally

small-size group for 1·26 mm. where there are only nine individuals although there are twenty-four in the groups on either side. This is probably a sampling error due to the groupings being too small, there may have been some unconscious bias against the number for this group, in the units used for measurement. This irregularity is not apparent when the data are plotted on the probability paper.

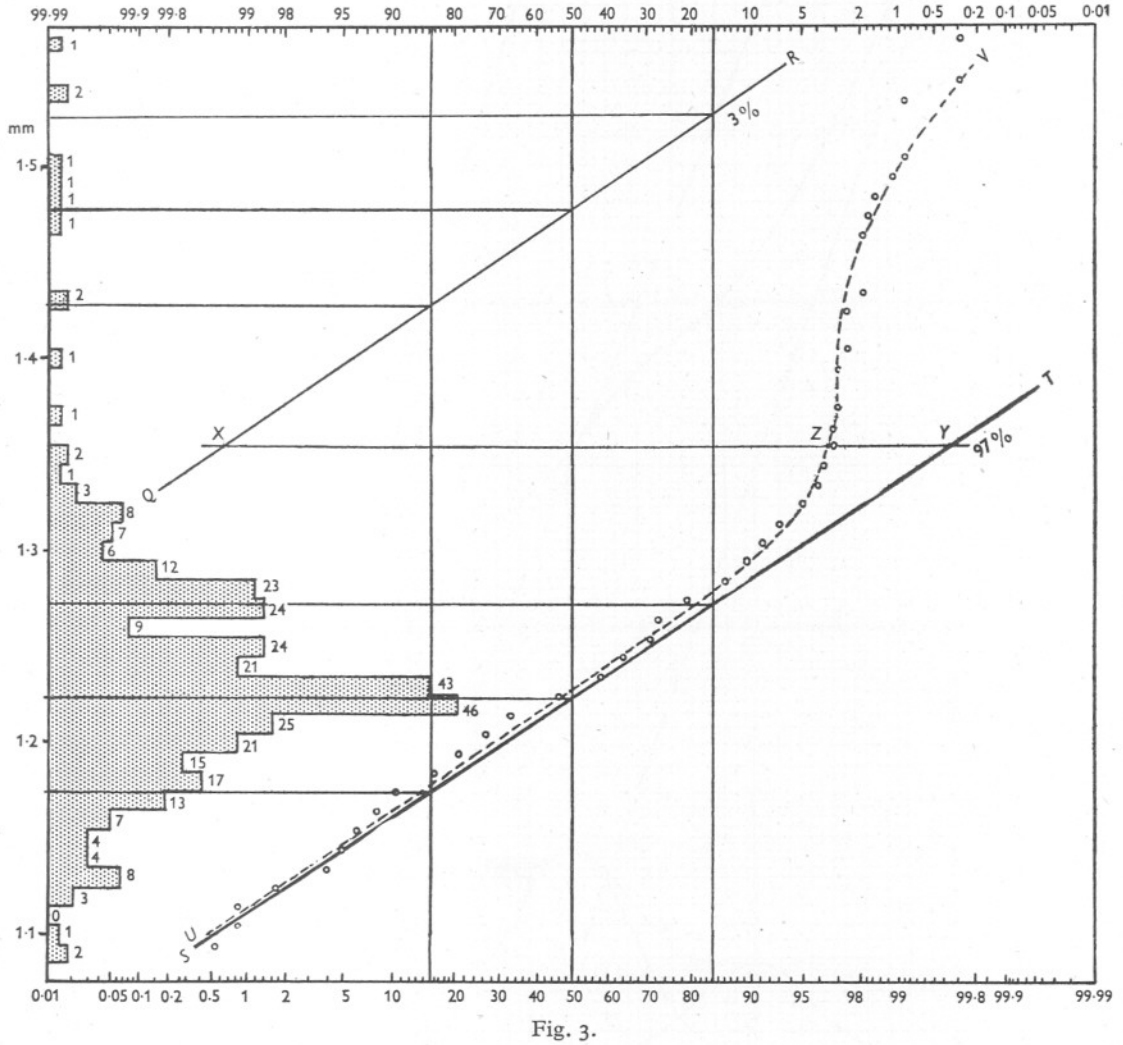

Fig. 3.

Polymodal curves are not difficult to analyse if the sample is large enough and the component populations do not overlap each other too much. Fig. 4 is drawn to show the effect of mixing three hypothetical populations in the

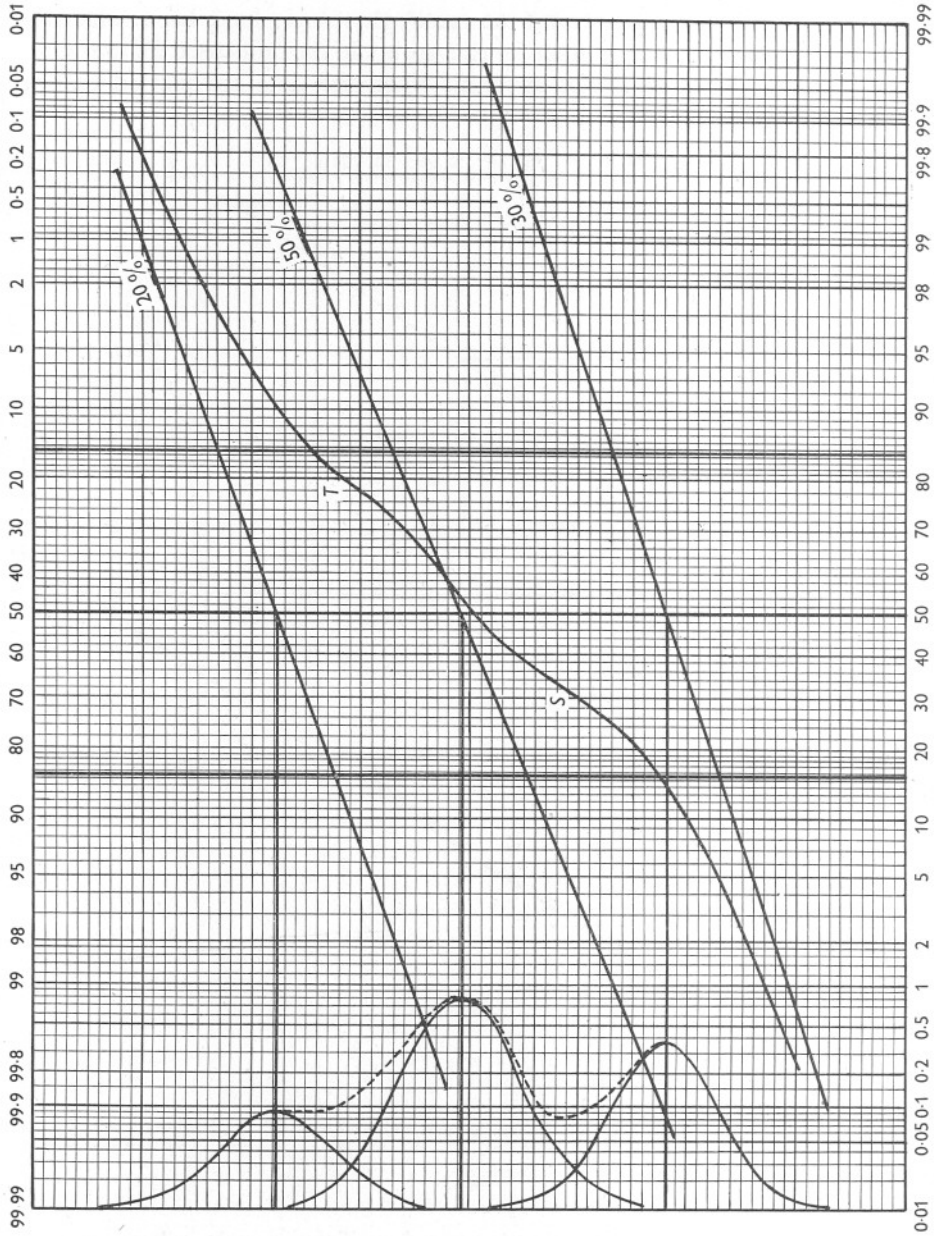
20%
50%
30%
T
S
0·01 0·05 0·1 0·2 0·5 1 2 5 10 20 30 40 50 60 70 80 90 95 98 99 99·8 99·9 99·99

Fig. 4.

proportions 2:5:3. There are two points of inflexion at S and T, and these give the approximate proportions in which the populations are mixed, S occurring on the 30 % and T on the 80 % vertical, the difference between the two being 50 %. It is sometimes possible to analyse a polymodal curve even though there is considerable overlapping of the populations and there are no points of inflexion to guide one. Figs. 5 and 6 show two possible ways of analysing such a sample. The data analysed are the lengths of 1122 fish calculated by Ford (1928, p. 294), from the measurements of scales. Hodgson (Buchanan-Wollaston & Hodgson, 1929) considered that this population of fish 'certainly contains two modes at 12 cm. and about 15 cm.' He had reason to believe from experience with other samples that there was a third mode, he thought concealed between the two. We shall see that analysis by probability paper indicates a third, small population with a smaller mean length than either of the main populations. The sigmoidal shape of the distribution when plotted on probability paper is very similar in appearance to that of Fig. 1 and the data can be analysed as a symmetrical bimodal curve in the same way as Fig. 1 was. The resultant of the two lines shown fits the lengths of the fish above 13 cm. long very well and the fish below this length not quite so well. A better fit is shown in Fig. 6 where the upper tail of the distribution is fitted by a 20 % line. This leaves 30 % of the sample of fish to be fitted in between. By trial and error one finds that the only way of placing this line is at an angle to the other two as shown in Fig. 6. The resultant of these three lines is shown by the small x's, while the circles are plotted from Ford's data. The agreement between the two is remarkable. The summation of the three distributions, shown as a dotted line superimposed on the histogram, also shows much better agreement with the sample than does the cruder analysis of Fig. 5. Table II gives the expected frequencies calculated from the results of Figs. 5 and 6 with the help of either Sheppard's tables or of tables of probits.

A χ^2 test for goodness of fit applied to these shows that the analysis of Fig. 5 is very unlikely to be true; but that the data are quite consistent with the analysis of Fig. 6, the probability of a χ^2 of 4·65 being about 0·2. In the calculations of χ^2 the classes for the 6·5 and 18·5 cm. have been bracketed together. A degree of freedom has to be subtracted for each mean, for each standard deviation, for all but one of the component populations and for the total. It is not claimed that the analysis of Fig. 5 gives the only possible, or even the best solution, indeed an attractive solution is the following:

50 % of the population with a mean 14·5 cm. and S.D. 1·3
40 % of the population with a mean 11·85 cm. and S.D. 1·1
10 % of the population with a mean 9·15 cm. and S.D. 0·9

The χ^2 summation for this solution is 7·95 which gives a value for $p = 0{\cdot}048$. The goodness of fit is therefore not quite so good; but the ratio of the standard deviation to the mean for each population remains constant and this might be

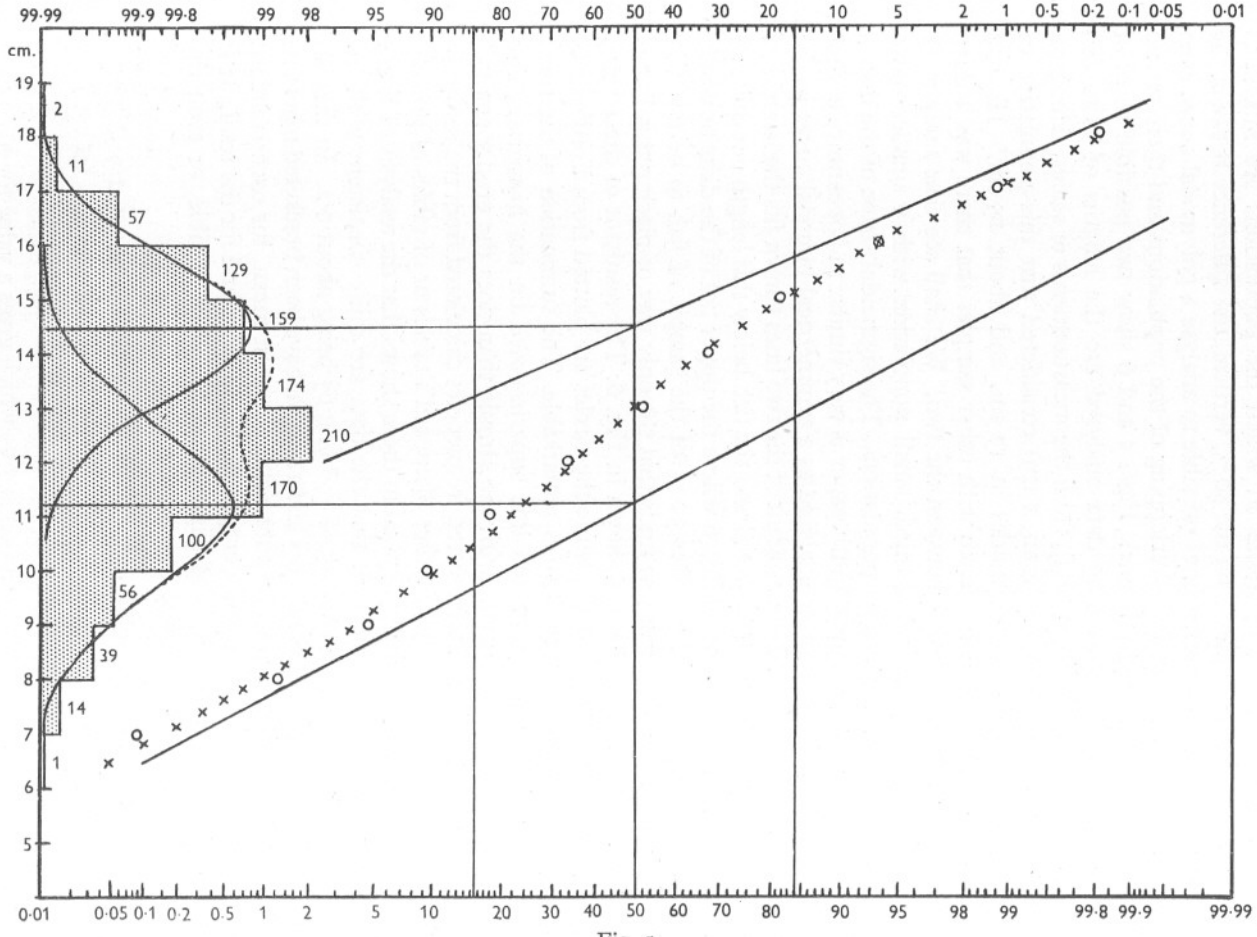
99·99 99·9 99·8 99 98 95 90 80 70 60 50 40 30 20 10 5 2 1 0·5 0·2 0·1 0·05 0·01
cm.
19 18 17 16 15 14 13 12 11 10 9 8 7 6 5
2
11
57
129
159
174
210
170
100
56
39
14
1
0·01 0·05 0·1 0·2 0·5 1 2 5 10 20 30 40 50 60 70 80 90 95 98 99 99·8 99·9 99·99

Fig. 5.

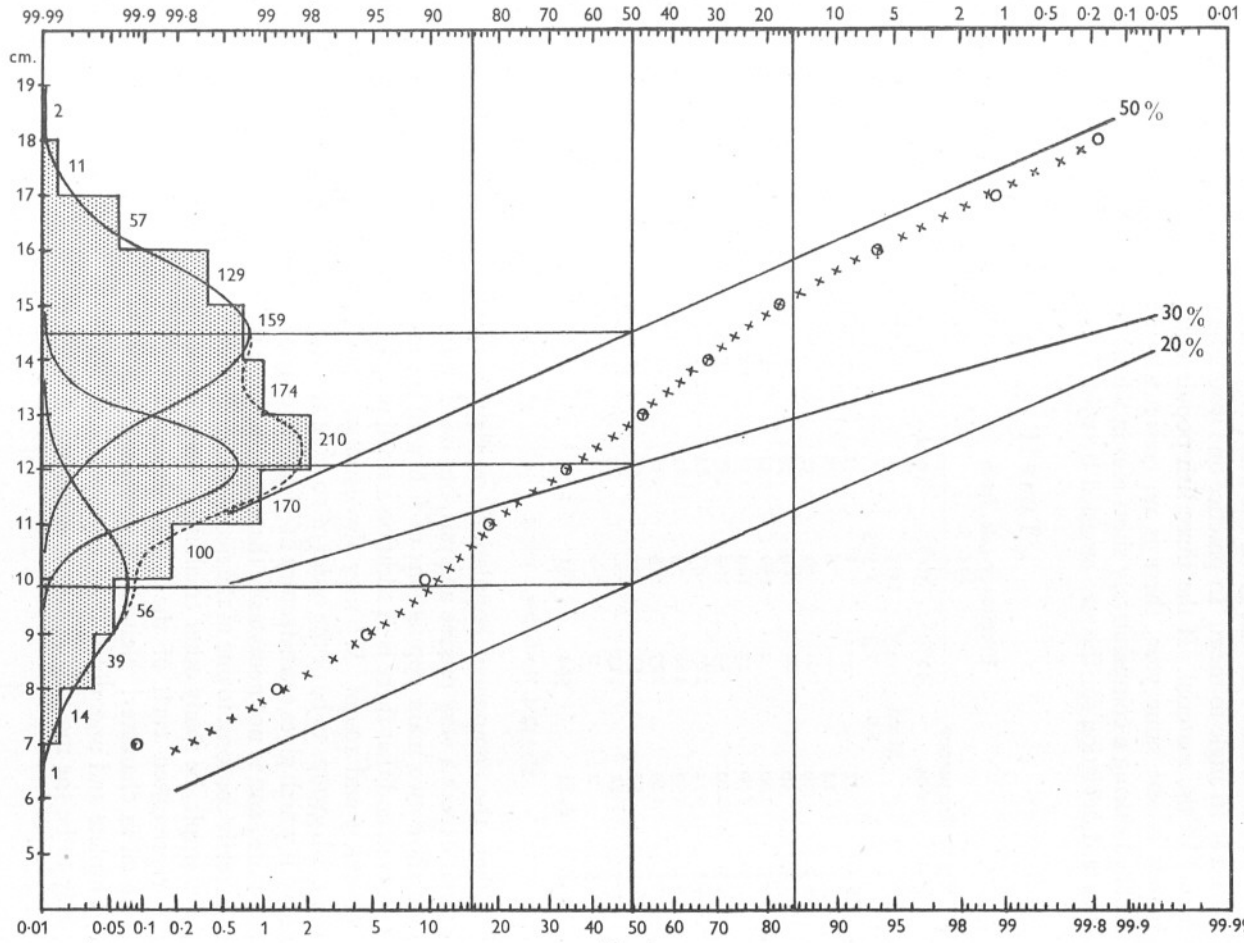
99·99 99·9 99·8 99 98 95 90 80 70 60 50 40 30 20 10 5 2 1 0·5 0·2 0·1 0·05 0·01
cm.
19 18 17 16 15 14 13 12 11 10 9 8 7 6 5
2
11
57
129
159
174
210
170
100
56
39
14
1
50 %
30 %
20 %
0·01 0·05 0·1 0·2 0·5 1 2 5 10 20 30 40 50 60 70 80 90 95 98 99 99·8 99·9 99·99

Fig. 6.

in its favour. I have chosen the solution of Fig. 5 for illustration because it is the more instructive in that given the other two lines the 30 % line has to be placed at an angle to them. In practice one would be guided by the biological nature of the material. If the three fish populations, for example, are three broods of the same year, there is no reason why one brood may not have hatched during a comparatively short season which might account for a small standard deviation for the size attained by winter.

TABLE II

Class (cm.)	Frequency obs.	Frequency calc. from Fig. 5			Frequency calc. from Fig. 6			
		50 %	50 %	Total	50 %	30 %	20 %	Total
	Mean ...	14·5	11·2		14·5	12·1	9·9	
	S.D. ...	1·3	1·505		1·3	0·8	1·3	
6·5	1	—	1	1	—	—	3	3
7·5	14	—	8	8	—	—	13	13
8·5	39	—	31	31	—	—	39	39
9·5	56	—	79	79	—	1	64	65
10·5	100	2	132	134	2	27	61	90
11·5	170	13	144	157	13	123	32	168
12·5	210	55	102	157	55	141	10	206
13·5	174	127	47	174	127	41	2	170
14·5	159	168	14	182	168	3	—	171
15·5	129	127	3	130	127	—	—	127
16·5	57	55	—	55	55	—	—	55
17·5	11	13	—	13	13	—	—	13
18·5	2	2	—	2	2	—	—	2
Total	1122	562	561	1123	562	336	224	1122
		$\chi^2=43{\cdot}4$, $n=6$, $p=<0{\cdot}001$			$\chi^2=4{\cdot}65$, $n=3$, $p=0{\cdot}2$			

When the component populations overlap as much as they do here one cannot expect a very precise analysis, but one is justified in saying that Ford's data show two main populations of fish with mean lengths of about 12 and 14·5 cm., and that there is in addition a small population whose mean length is between 9 and 10 cm. We may also estimate the standard deviations of the three fish groups to be of the order of 10 % of the respective mean lengths. An analysis which gives a satisfactory fit (Fig. 6) is only one of the simplest possible solutions, and is not necessarily the most complete picture of the facts, which may really conform to one of the many possible complex solutions. There may, for example, be many other small fish groups in the sample, and if both sexes are represented each of the three main groups are probably themselves bimodal in character. Neither the graphical method nor any other will give a complete and unequivocal solution; but fortunately the simplest solution is likely to be the most significant biologically, as well as statistically.

ACKNOWLEDGEMENTS

I am indebted to Dr G. A. Barnard, University of London, for bringing probability paper and its possibilities to my notice, and also to Mr W. F. Adams, Ministry of Transport, and Mr H. J. Joseph, Post Office Engineering Research Station, for their interest and help with the literature quoted.

REFERENCES

BUCHANAN-WOLLASTON, H. J. & HODGSON, W. C., 1929. A new method of treating frequency curves in fishery statistics, with some results. *Journ. Conseil Int. Explor. Mer*, Copenhagen, Vol. 4, pp. 207–25.

DOUST, J. F. & JOSEPHS, H. J., 1941–42. A simple introduction to the use of statistics in telecommunications engineering. *Post Office Eng. Journ.*, Vol 34, pp. 36–41, 79–84, 139–44 and 173–8.

FORD, E., 1928. Herring investigations at Plymouth. III. The Plymouth winter fishery during the seasons 1924–25, 1925–6 and 1926–7. *Journ. Mar. Biol. Assoc.*, Vol. 15, pp. 279–304.

GOTTSCHALK, V. H., 1948. Symmetrical bi-modal frequency curves. *Journ. Franklin Inst. Philadelphia*, Vol. 245, pp. 245–52.

HAZEN, A., 1913. Storage to be provided in impounding reservoirs for municipal water supply. *Proc. Amer. Soc. Civil Eng.*, Vol. 39, pp. 1943–2044.

LEVI, F., 1946. Graphical solutions for statistical problems. *The Engineer, London*, Vol. 182, pp. 338–40 and 362–4.

PEARSON, K., 1894. Contribution to the mathematical theory of evolution. *Phil. Trans. Roy. Soc.*, A, Vol. 185, pp. 71–110.

—— 1914. *Tables for statisticians and biometricians*, Part I. Cambridge.

RISSIK, H., 1941. Probability graph paper and its engineering applications. *The Engineer, London*, Vol. 172, pp. 276–82 and 296–8.