INSTRUCTIONS FOR THE USE OF PROBABILITY PAPER IN ANALYSING THE FREQUENCY DISTRIBUTION OF SAMPLES IN BIOLOGICAL PROJECTS

HELMUT F. VAN EMDEN
School of Biological Sciences, The University of Reading
h.f.vamemden@reading.ac.uk

Probability graph paper enables the statistical analysis of frequency data to be accomplished graphically in a way which gives students an easily comprehended visual representation of normal and non-normal distribution statistics. Approximations to a t-test are possible, as is the separation of two overlapping normal distributions. Procedural instructions are provided with numerical examples and the relevant probability paper graphs.

INTRODUCTION

The use of probability graph paper (PP)[1] in the study and comparison of distributions in ecology was described by Lewis and Taylor (1967). Harding (1949) had earlier addressed the separation of polymodal distributions using PP. Although included by Southwood (1966) in his seminal work *Ecological Methods*, most ecologists have probably never explored the uses of PP in their discipline, and it is timely to raise awareness of the uses of this form of graph paper.

Although mathematicians would probably argue that the use of PP with non-normal distributions is not entirely justifiable, comparison of the technique with more elaborate computational methods suggests that PP methods are unlikely to lead to misleading conclusions.

There is practical value in being able to use a simple graphical method with theoretically difficult distributions as are often encountered in the field. Additionally, PP methods provide a graphical presentation of statistical concepts in a form that is easily understood by beginners, including at school level. I have found PP methods useful in introducing statistical concepts in the field courses I have run for the Field Studies Council at four of its centres and in field courses for zoology departments at Reading as well as at two other universities. Moreover, the number of participants in field courses means that good frequency data can readily be collected.

Only the technique will be described here. The parameters being estimated are likely to be familiar and are explained in any textbook on biological statistics.

THE FREQUENCY DISTRIBUTION TABLE

Tabulate the data in order of magnitude of the observations. The example used here (Table 1, columns x and y) concerns the number of lichen patches found in 74 randomly placed square metre quadrats on a disused hard tennis court.

With many different numbers in samples, it may be sensible to group the observations into classes, e.g. 1-10, 11-20, 21-30, 31-40 etc. and use the mid-point of each class (e.g. 5, 15, 25 etc.) when plotting x values on the graph paper.

We then calculate the percentage of the samples each y value represents (Table 1. '% of total y' column) and accumulate these percentages so that the figures (Table 1 'Accumulated. % of y' column) show the proportion of the data which is included by each class plus all lower classes.

[1] Sheets of probability graph paper do not appear to be available for purchase any longer but free downloads can be found by searching the internet for "normal probability graph paper"; the download www.weibull.com/pubs/paper_normal.pdf is an excellent example.

TABLE 1. Number of lichen patches in 74 square metre quadrats.

Observations (x) = no. of patches	Frequency (y) = no. of quadrats	% of total y	Cumulative % of y
6	1	1.35	1.35
7	4	5.40	6.75
8	8	10.81	17.56
9	15	20.27	37.83
10	17	22.98	60.81
11	15	20.27	81.08
12	9	12.16	93.24
13	3	4.06	97.30
14	2	2.70	100.00
	Total=74		

PLOTTING THE DATA ON PP

Make the vertical scale on the graph paper a scale of the x values. The horizontal scale is already one of accumulated percentages (Cum. % y).

Now plot each x value at the intercept with its accumulated % y, i.e. 6 and 1.35, 7 and 6.75, 8 and 17.56 etc. Omit the 100% point; you will not have included the largest possible value in your sampling and you will see that if you place this point at the extreme right of the graph paper (99.99%) its position bears no relation to the other points.

View the graph paper from the vertical axis with the paper held horizontally at eye level. This is the best way of checking what kind of line (e.g. straight, curved, sigmoid) best fits the points.

IF THE POINTS APPROXIMATE TO A STRAIGHT LINE:

The distribution is 'normal' or close to 'normal'. The frequency distribution is symmetrical around the 50% point (median), which is also the arithmetic average (mean) of the observations in a sample. This is the maximum per cent of total y (see table), and this percentage decreases symmetrically in both upward and downward directions. Draw the straight line (Figure 1).

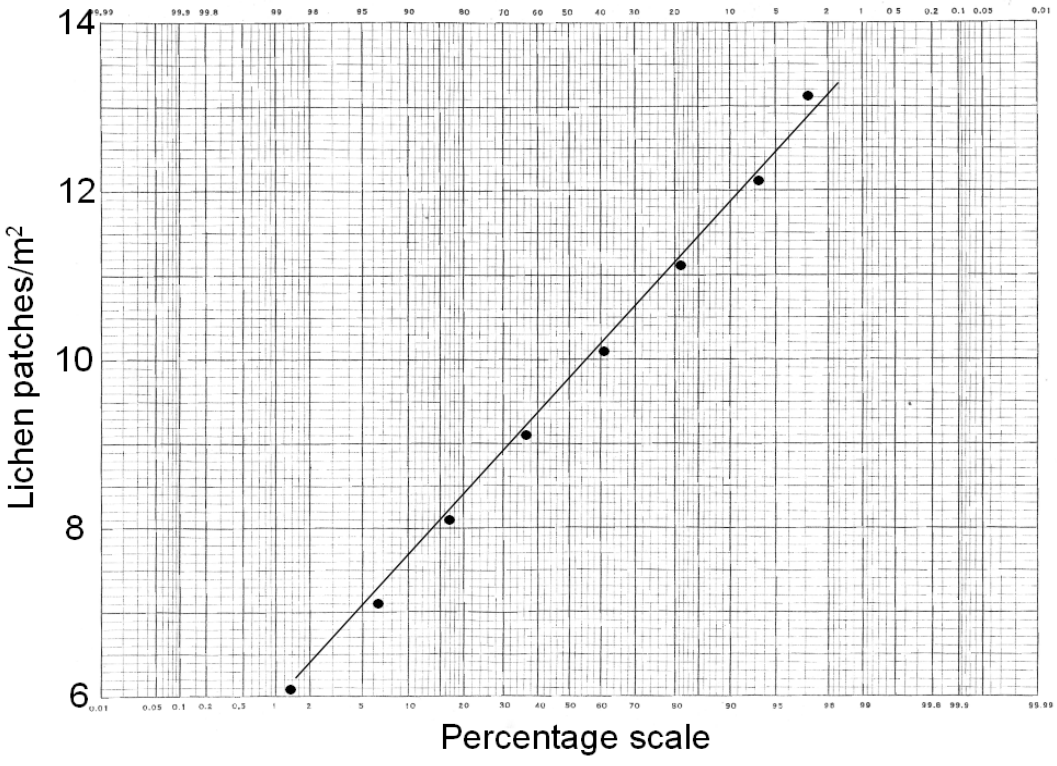

FIGURE 1. Probability plot of the normal distribution of lichen patches on a tennis court (data of Table 1).

To determine the median of the distribution (also the mean with a straight line):

Project vertically from 50% on the horizontal scale to the straight line (Figure 2). The intercept with the line is the mean (the same as the median in a normal distribution) and can be read off on the vertical scale (9.78 in the example).

To determine the standard deviation of the distribution:

This gives the limits within which the middle 68% of the population falls. Therefore, project vertically to the line from the 16% and 84% points on the horizontal percentage scale. Project sideways to the vertical scale (Figure 2) and use a millimetre ruler to read off the two values (their distances from the mean should be equal). In the example, the two values are 8.18 and 11.38. The frequency distribution can therefore be defined as 9.78 ± 1.60.

It is inevitable that fitting a straight line to points which deviate slightly from the line will give answers a little different from the standard statistical calculation, which gives the very similar result of 10 ± 1.69.

To determine the standard error of the mean of the distribution:

This is simply the standard deviation divided by the square root of the number of observations, i.e. in the example it is 1.60 ÷ √74 = ± 0.19. These limits can be added to Figure 2.

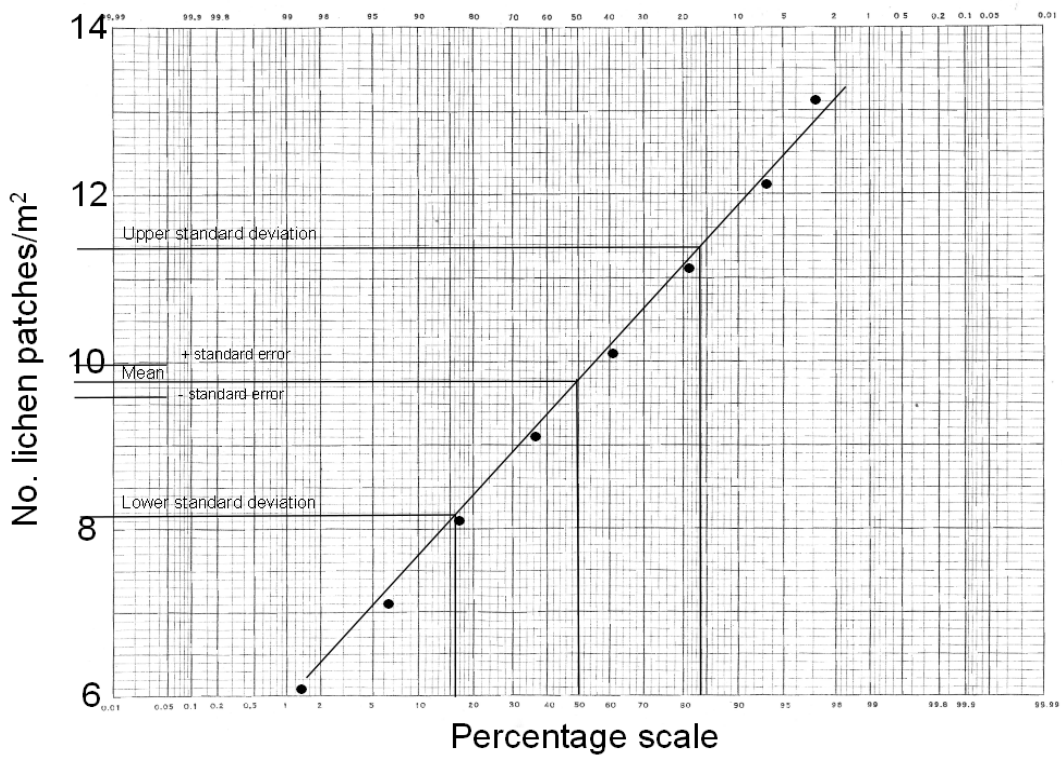

FIGURE 2. The statistics of mean, standard deviation and standard error calculated from the probability plot of Figure 1.

IF THE POINTS APPROXIMATE TO A SIMPLE CURVE:

A curve shows that the frequency distribution is not normal, and therefore not symmetrical about the median. This occurs, for example, when the observations are clumped, so that the arithmetic mean lies either well below or well above the range of most frequent observations. The median (i.e. the most frequent observation) replaces the mean in the calculations.

Draw the curve:

Example data are given in Table 2 and Figure 3. They concern the number of grasshoppers in 50 randomly placed square metre quadrats on a large area of grazed grassland. Figure 3 shows the fitted curve. On an even sward, grasshoppers can be expected to show a random (not the same as 'even') distribution, and this type of distribution will plot as a curve, but variation in vegetation may lead to additional non-random clumping.

TABLE 2. Number of grasshoppers in 50 randomly placed square metre quadrats (the calculations on small italics have been added in relation to the later calculation of the type of distribution these data represent).

Observations (x)	Frequency (y)	*Number of grasshoppers*	% of total y	Cumulative % of y
1	1	*1*	2	2
2	9	*18*	18	20
3	10	*30*	20	40
4	7	*28*	14	54
5	7	*35*	14	68
6	4	*24*	8	76
7	3	*21*	6	82
8	2	*16*	4	86
9	1	*9*	2	88
10	2	*20*	4	92
11	1	*11*	2	94
12	1	*12*	2	96
13	0	*0*		
14	1	*14*	2	98
15	1	*15*	2	100
	Total=50	***Total =254***		

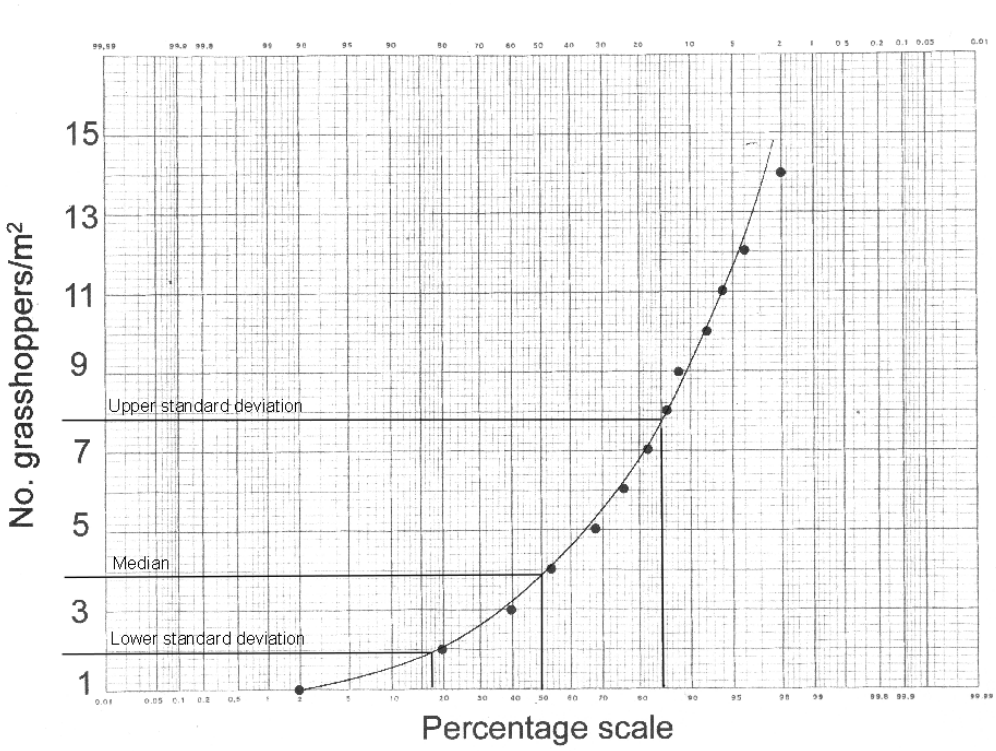

FIGURE 3. Probability plot of the slightly clumped distribution of grasshoppers (data of Table 2) with median and non-symmetrical standard deviations added.

To determine the median and standard deviation of the distribution:

Proceed as for a straight line relationship; i.e., project vertically to the curve from 50%, 16% and 84% on the horizontal scale to the curve, and read off the median (3.90 grasshoppers/quadrat in the example) and upper and lower standard deviations (1.78 and 7.80 respectively) on the vertical scale (Figure 3). The median replaces the mean as the most frequent sample size in a non-normal distribution (the arithmetic mean in this example is actually larger at 5.08). The frequency distribution can therefore be appropriately defined asymmetrically as having a median of 3.90 (-2.12, + 3.90).

To determine the standard error of the median of the distribution:

With asymmetric standard deviations, we first need to transform our distribution to a normal one (i.e. a straight line). This is done very simply by drawing a straight line between the standard deviation percentages on the

curve (this line is shown dotted in Figure 4). This exactly transforms the curve to normality without involving any particular known transformation function such as log, square root or arcsin.

Projecting to the original straight line from 50% gives us a mean (transformed median) for this imaginary normal distribution, with the original standard deviation values now symmetrical around it. The numbered sequence of steps that now follows is illustrated by the encircled numbers in the Figure:

1. Extend the transformed median and upper and lower standard errors to the right-hand margin of the graph paper. Measure the distance between the two standard deviation values (in millimetres on a ruler is often more convenient than using the units on the vertical scale). This is 5.8 grasshoppers in the example. Halve it, you have the standard deviation for the mean on the transformed scale. In Figure 4 this is 5.8/2 = 2.9 grasshoppers.
2. The standard error for samples of 50 data is 2.9/√50 = 0.41 Plot this symmetrically around the extended transformed median line.
3. From where each of these standard errors of the transformed median intersect with the dotted line joining 16% and 84%, project upward or downward to the original curve.
4. Project these intersections as well as the line for the original median to the right, and the median will now have, as is appropriate, asymmetrical standard errors at 3.9 (-0.38, + 0.45).

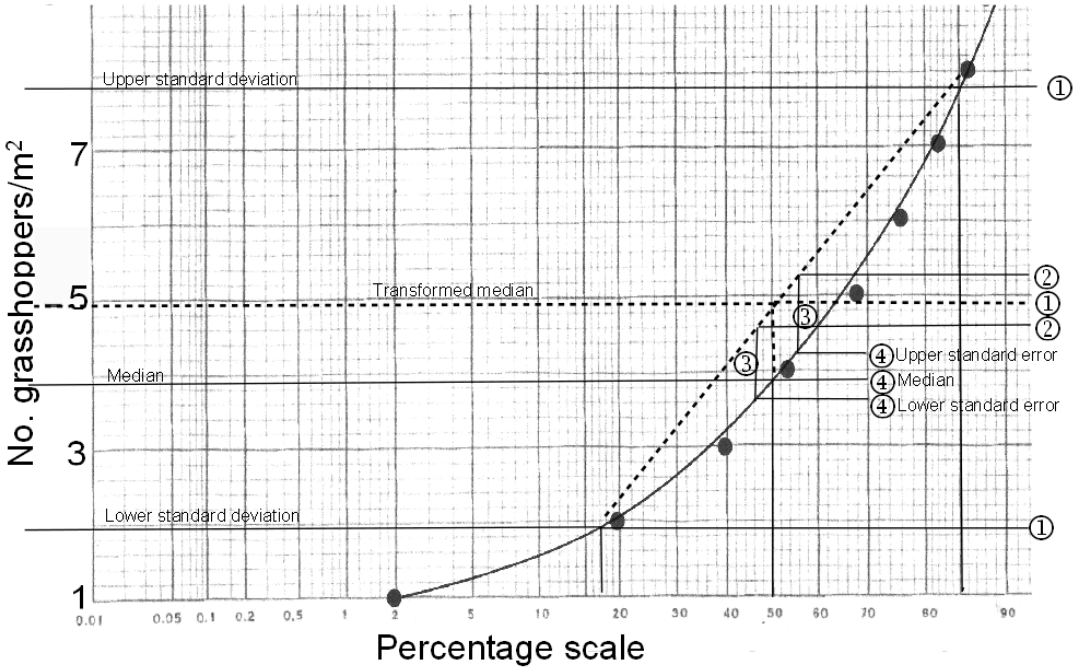

FIGURE 4. Calculating standard errors for the probability plot of Figure 3. Circled numbers show the sequence of graphical operations (see text).

USING PP PAPER TO CHOOSE THE BEST TRANSFORMATION FOR NON-NORMAL DATA

The graph paper is also useful for obtaining guidance as to the best transformation for a set of frequency data. If the non-normal distribution in question is well defined by the available data points, there are two approaches to identifying an appropriate transformation.

1. If one of the standard transformations such as logarithmic or square root looks likely, this can quickly be checked by seeing if the appropriately transformed values plotted on the vertical scale are fitted by a straight line.
2. Alternatively a straight line may be drawn between the standard deviation points as in Figure 4. Then project several points equally spaced on the vertical axis back to the straight line, and from there upwards or downwards to the plotted line for the original data. Project these intersections back to the vertical scale, and the relative distances between these last lines may suggest an appropriate transformation.

The type of non-normal distribution

To identify the type of distribution, we need to calculate the arithmetic mean (not the median) and the variance. We will use the data in Table 2 and Figure 3 for an example, where the arithmetic mean is 5.1 (i.e. 254/50 from Table 2).

Begin with the difference between the upper and lower standard deviations (i.e. projecting to the vertical scale from the curve for the 16 and 84 percentage points on the horizontal scale (= 5.8).

Halve this value and square the result (= 8.4). This gives a rough estimate of 'variance'. Divide this by the mean (= 8.4/5.1 = 1.7).

If variance/mean approximates to 1, a random distribution is indicated. If variance/mean is less than 1, the distribution is more regular than random and if more than 1 it is more clumped.

Our example, with a variance/mean ratio of 1.7, is somewhat clumped. Incidentally, doing the calculations directly from the data for comparison also gives a similarly clumped variance/mean ratio (2.0).

TESTING THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN THE MEDIANS OF TWO DISTRIBUTIONS:

This is a graphical analogy of the t-test. For the two distributions to compare we will use the lichen patch data in Table 1 and Table 3 (from a tennis court with a different surface).

Draw separately the lines for the two distributions as described earlier (Figure 5). In this example the distributions of lichen patches on both tennis courts are normal. However, perhaps controversially, the significance test described here might also be applied to two non-normal distributions or even one normal and one non-normal one!

We already have the relevant calculations for the first tennis court (closed circles on Figure 5), and only need to transfer the median (= the mean with a straight line) of 9.78 and its ± standard error limits of 0.19 to the vertical scale (solid lines).

For tennis court 2 (open circles), project vertically to the lines the median and – and + standard deviations (i.e. respectively the 16, 50 (median) and 84% points on the probability scale). Project these intersections to the vertical scale (dotted lines on Figure 6), giving a mean of 7.18 ± 1.35 (standard statistical computation gives 7.29 ± 1.38). Divide this standard deviation by the square root of the number of observations, i.e. $1.35 \div \sqrt{158} = \pm 0.11$ to obtain the standard error of the mean for tennis court 2 for insertion on the vertical scale on Figure 5.

The *standard error of difference* between the two means is the sum of the two standard errors of the means, i.e. 0.19 + 0.11 = 0.30. By comparison, the means are 2.6 units apart, 8.7 times the standard error of difference. Frequency distributions involve sufficient observations to use the rule of thumb that this figure only needs to be 2 for us to accept that the means are significantly different in statistical terms i.e. taking many repeat samples of the same number of observations, the chances of getting a mean from either distribution outside twice the standard error of difference is less than 5%

If either line is not straight (i.e. the distribution is not normal), simply obtain the standard error of difference by summing the two standard errors on side of the medians towards the other line.

TABLE 3. Data for lichen patches in 158 quadrats in tennis court 2.

Observations (x) = no. of patches	Frequency (y) = no. of quadrats	% of total y	Cumulative % of y
4	1	0.78	0.78
5	7	5.47	6.25
6	49	14.84	21.09
7	33	25.81	46.90
8	36	28.12	75.02
9	22	17.18	92.20
10	8	6.24	98.44
11	2	1.56	100.00
	Total=158		

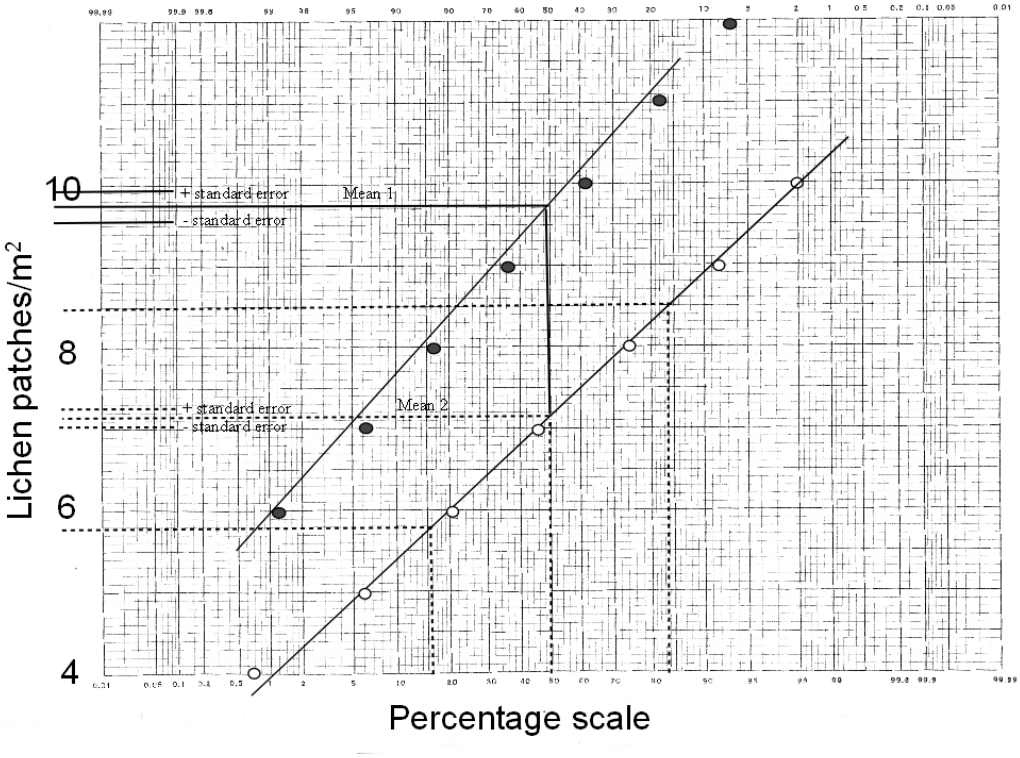

FIGURE 5. Probability plot of the normal distribution of lichen patches on the second tennis court (data of Table 3) added to Figure 2.

SEPARATING TWO OVERLAPPING NORMAL DISTRIBUTIONS

The plot of data against cumulative % on a probability scale sometimes results in a sigmoid curve (as in Figure 6). Such a curve is indicative of two overlapping normal distributions, and PP can be used to separate them.

The example in Figure 6 is a plot of the width of the head between the eyes (in mm) of 200 specimens of adult females of a leaf mining fly found in water traps. That the curve is sigmoidal suggests that two populations are represented; perhaps there are actually two very similar species masquerading under the same Latin name.

The first stage is to tabulate (Table 4) the data as far as 'cumulative % of y' as in previous Tables and plot the data on PP (Figure 6). Note that we use the mid-point for each 2 mm head width range. We now identify the point of inflexion of the curve, the point where it changes from concave to convex. In Figure 6 this is at the cumulative % of 58, and we mark this in the Table with emboldened typeface. The cumulative % values for the concave part of the curve (shown as A in Figure 7) remain unchanged and refer to the percentage scale at the bottom of the PP, but for the higher concave part (B) we start at the right hand end of the curve and use the cumulative % scale at the top of the PP which is reversed and increases from right to left. This is the same as subtracting the figure in the cumulative % column from 100 and produces the first extra column in Table 4.

Why do we do this? The overlap between the populations is greatest between the right end of distribution A and the left end of distribution B. So, we can best identify the straight lines to plot for the two normal distributions from the non-overlapping ends.

The inflexion point of 58 % suggests that the two distributions represent respectively 58% and 42% of the 200 flies measured. To find the lines to plot as two distinct normal distributions we treat the 1-58 cumulative % values as population A and the 64-100 as population B. For A we convert the data by multiplying them by 100/58 (italicised numbers in the end column of Table 4) and for B by multiplying the values in the extra column in the Table by 100/42 (the underlined numbers in the end column).

The straight line plots for the two now separated populations A and B can now be drawn through the points for cumulative % so obtained, and the statistics of mean and standard deviation can be evaluated separately for the two distributions (respectively the 50% and 16% cumulative percentages – see earlier). This gives (Figure 7) a mean of 2.39 ± 0.12 for A and 2.95 ± 0.16 for B.

TABLE 4. Head widths (mm) of females of a species of leaf-mining fly.

Head capsule width(x)	Mid-point of class	Frequency (y)	% of total y	Cumulative % of y	100 - cumulative % for points on the B part of the curve	Points to plat for the two populations
2.1-2.2	2.15	2	1	1		*1.7*
2.2-2.3	2.25	18	9	10		*17.2*
2.3-2.4	2.35	32	16	26		*44.8*
2.4-2.5	2.45	32	16	42		*72.4*
2.5-2.6	2.55	16	8	50		*86.2*
2.6-2.7	2.65	16	8	*58*		100.0
2.7-2.8	2.75	12	6	64	36	85.7
2.8-2.9	2.85	12	6	70	30	71.4
2.9-3.0	2.95	20	10	80	20	47.6
3.0-3.1	3.05	14	7	87	13	31.0
3.1-3.2	3.15	18	9	96	4	9.5
3.2-3.3	3.25	6	3	99	1	2.4
3.3-3.4	3.35	2	1	100		
		Total = 200				

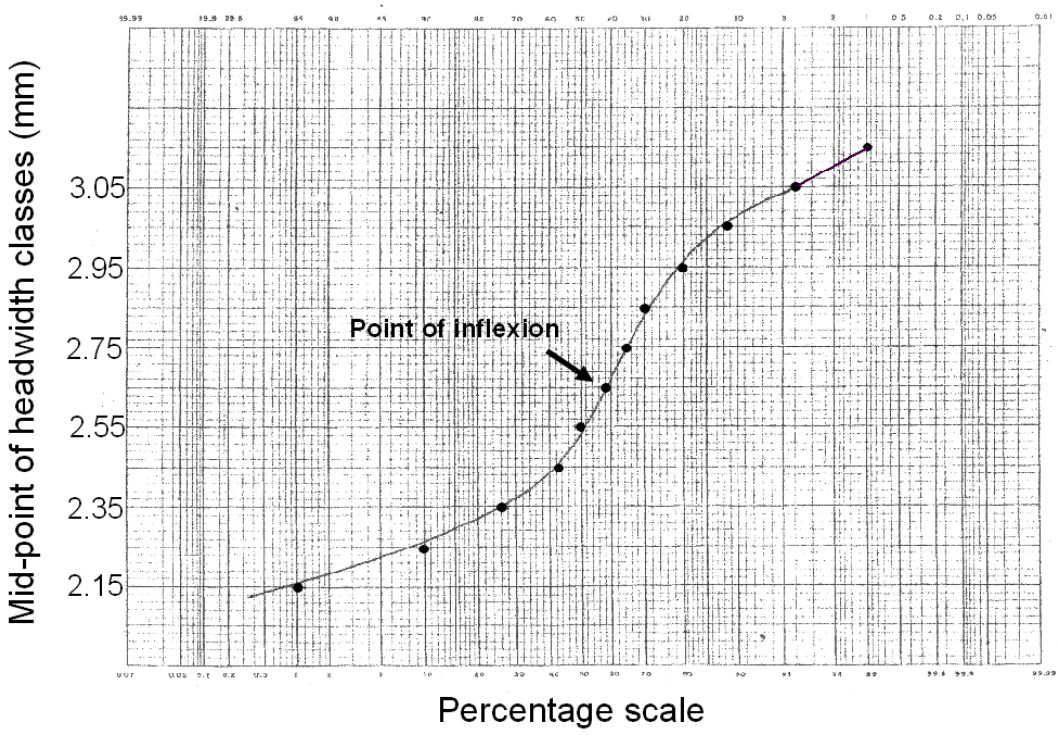

FIGURE 6. Probability plot of the head width of females of a species of leaf mining fly suggestive of two overlapping frequency distribution. The point of inflexion on the percentage scale is identified.

As before, the standard error of the mean for each distribution can be calculated by dividing the standard deviation by the square root of the relevant number of observations (respectively 58% and 42% of 200 = 116 for A and 84 for B). Not unexpectedly, given the close relationship of the two groups of flies, the standard error of both means is very similar, 0.01 for A and 0.02 for B. However, even taking the larger value of 0.02, the means are over 30 standard errors apart; this really rules out the possibility that the flies belong to the same population. It therefore appears there may be two what are called 'cryptic species' in the water trap samples.

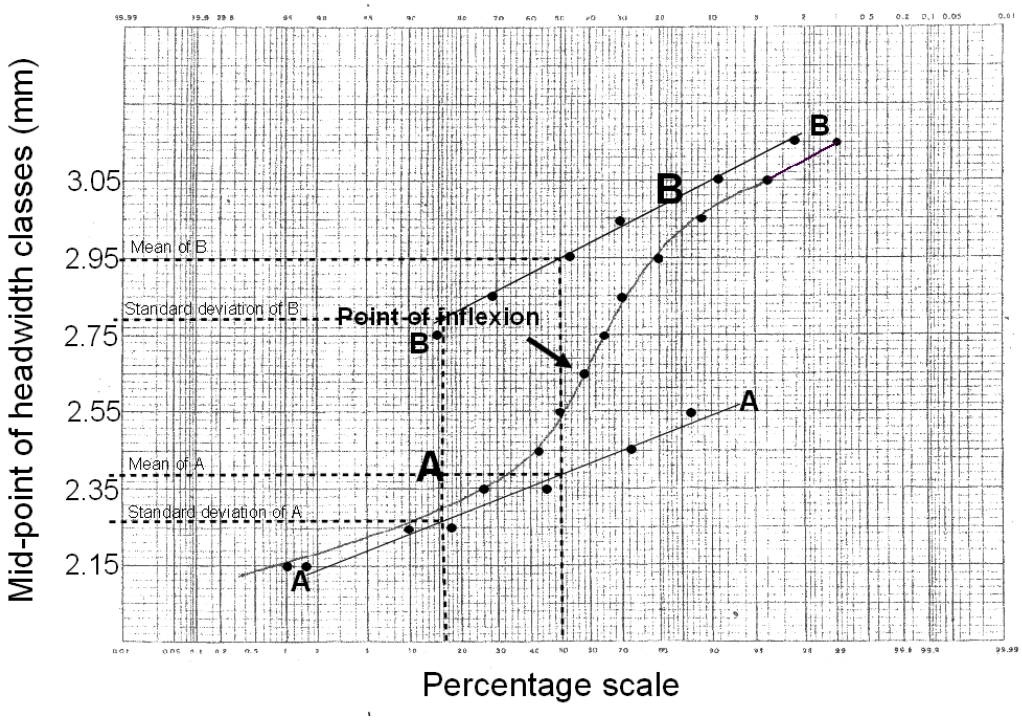

FIGURE 7. The separation of the probability plot of Figure 5 into the two normal distributions A and B (see text).

CONCLUSIONS

The number of participants on Field Courses leads to the results of projects often being in the form of frequency distributions. Probability graph paper can be a quicker way of analysing the data than entering a lot of data into a computer programme, and has the added advantage of teaching statistics in a way that is easily understood, even by those with no previous acquaintance with the subject. Drawing a "best fit" line by eye through the points on the graph paper has some subjective element, and the estimates obtained for the statistics may well vary slightly from those obtained by the traditional algorithms, but they should be close enough that ecological interpretations will be valid.

REFERENCES

Harding, J.P. (1949). The use of probability paper for the graphical analysis of polymodal frequency distributions. *Journal of the Marine Biology Organisation UK*, **28**, 141-153.

Lewis, T, and Taylor, L.R. (1967). *Introduction to experimental ecology*. Academic Press, London, 401pp.

Southwood, T.R.E. (1966). *Ecological methods, with particular reference to the study of insect populations*. Methuen, London, 391pp.

