

## A NOTE ON THE RELATIONS OF CERTAIN PARAMETERS FOLLOWING A LOGARITHMIC TRANSFORMATION

By Mary Bagenal

The Marine Station, Millport

(Text-fig. 1)

The logarithmic transformation has been used in the statistical analysis of certain marine biological data. Parameters calculated from the transformed distributions have been used in the description of the observations, but it appears that some of the mathematical procedures adopted have not been fully understood.

Winsor & Clarke (1940) have analysed data from plankton hauls. The problem considered by them was the determination of the variability in numbers of animals caught by repeated hauls through the same body of water. The raw data were characterized by a constant coefficient of variation, i.e. the variability in catch was proportional to the size of the catch. By transformation from the actual numbers caught to their logarithmic values, it was possible to equalize the variances and to apply the method of analysis of variance to estimate the variability due to the different sources. Finally, an estimate of the coefficient of variation was obtained from the logarithmic values. Their method of estimation was quoted by Snedecor (1946, p. 451) and was employed by Barnes & Bagenal (1951) in their study of repeated trawl hauls. Barnes & Bagenal also calculated confidence limits for comparison of observations following the method of Silliman (1946) in his work on pilchard eggs. These methods seem to be based on a misunderstanding of the nature of a transformation. The formal relation between the variance of such transformed data and the coefficient of variation of the untransformed data follows from the moments of the 'log-normal' distribution, given first by Wicksell (1917). From these moments it will be shown that the method given by Winsor & Clarke is mathematically unsound. However, the discrepancies between their values and those now obtained are comparatively small (ranging from about 2% for a coefficient of variation of 20% to about 15% for one of 60%). In all cases the method given in this paper results in a smaller estimate.

For comparison of catches Silliman gave confidence limits which were expressed as percentages of the mean catch. The mean used, although not explicitly stated, was the geometric mean of the catches.

## THE LOGARITHMIC TRANSFORMATION

For untransformed data with a positively skew frequency distribution a transformation from the values of  $x$ , to the logarithmic values  $y = \log_{10} x$ , may result in a distribution of the normal form. The relation between the two distributions is such that the proportion of observations falling within a given range of

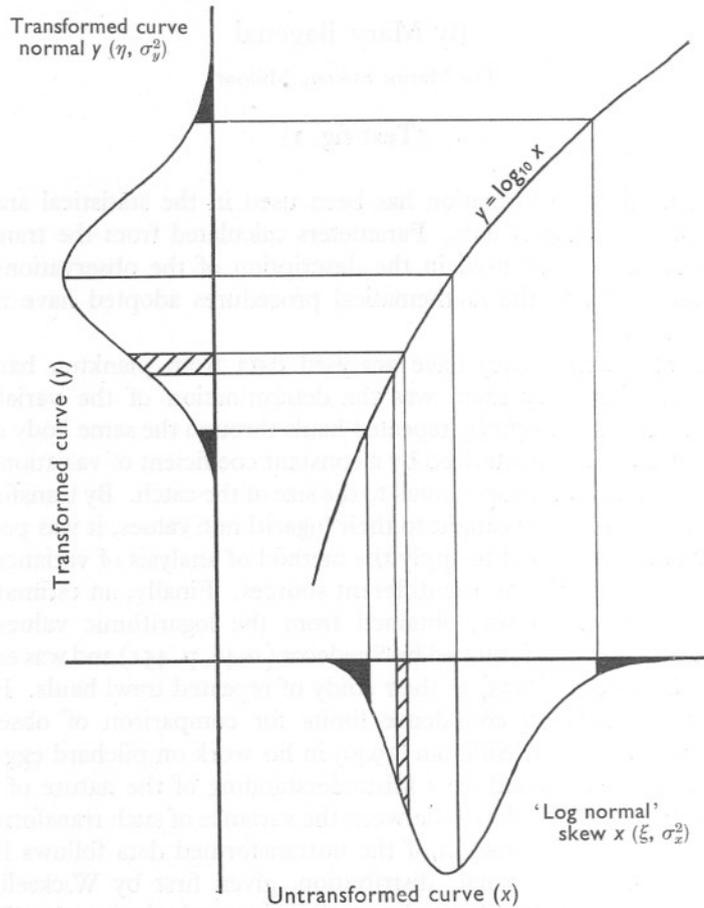


Fig. 1. Diagram to show the relation between the 'log-normal' and the normal distributions. For explanation see text.

the  $x$  values will equal the proportion of observations falling within the corresponding range of  $y$  values. This is so since there is a corresponding value of  $y$  to every value of  $x$ .

Such a transformation is shown in Fig. 1 (after N. L. Johnson, 1949).

An untransformed frequency distribution, which is noticeably skew, is represented on the lower margin. Let this be the distribution of  $x$  with mean

$\xi$  and variance  $\sigma_x^2$ . In the centre of the figure the relation  $y = \log_{10} x$  is represented graphically by the smooth curve. On the left-hand margin is the transformed frequency distribution, projected from the original distribution by means of the curve  $y = \log_{10} x$ . This distribution of  $y$ , with mean  $\eta$  and variance  $\sigma_y^2$ , is normal.

The cross-hatched columns representing frequencies over corresponding ranges of the two variates are equal in area. By consideration of this property of correspondence of areas certain relations between parameters of the two distributions become apparent. The mean, median and mode of the transformed distribution, which is normal, coincide and divide the area under the curve equally. The corresponding value in the untransformed distribution also divides the area under the curve equally and is the median, but because the curve is skew will not coincide with the mean and mode.

The normal curve has been thoroughly investigated and tabulated and use has been made of many of its properties. The 95% confidence limits are those values of the variate between which 95% of the total observations lie, and for the normal curve are approximately given by the mean plus and minus two standard deviations (or for means of small samples  $t$  standard deviations). In the figure these are at  $\eta \pm 2\sigma_y$  for the transformed distribution, and the corresponding values in the untransformed distribution will also represent 95% confidence limits since the shaded 'tail areas', are equal.

The mean and variance of a frequency distribution do not depend simply on area and therefore the values for one distribution cannot be obtained simply by transformation of the appropriate values for the transformed distribution; but must be obtained from the mathematical relations between them.

Wicksell (1917) gives for the transformation  $y = \log_e x$ , where  $y$  is normally distributed with mean  $\eta$  and variance  $\sigma_y^2$ , the relation:

$$r\text{th moment about origin of } x = e^{r\eta + \frac{1}{2}r^2\sigma_y^2}, \tag{1}$$

whence,  $\text{mean } x = e^{\eta + \frac{1}{2}\sigma_y^2}, \tag{2}$

and  $\text{variance } x = e^{2\eta + \sigma_y^2} (e^{\sigma_y^2} - 1). \tag{3}$

For the transformation  $y = \log_{10} x$ , rearrangement gives the relations

$$\text{mean } y = \eta = \log_{10} \left[ \frac{\xi}{(1 + \sigma_x^2/\xi^2)^{\frac{1}{2}}} \right], \tag{4}$$

$$\text{variance } y = \sigma_y^2 = \log_{10} e \log_{10} (1 + \sigma_x^2/\xi^2). \tag{5}$$

Now  $\sigma_x/\xi = V_x$  is the ratio of the standard deviation of  $x$  to its mean value and is the coefficient of variation of the untransformed data. Rewriting (5) we have

$$\sigma_y^2 = 0.43428 \log_{10} (1 + V_x^2),$$

from which it is seen that if the coefficient of variation of  $x$  is constant the variance of the transformed variate  $y$  will be constant and the methods of the 'analysis of variance' may be used. From equation (4) it is seen that the relation between the means of the transformed and untransformed distributions is not simple. For a sample the mean of the transformed values equals the logarithm of the *geometric mean* of the actual numbers.

Since

$$y = \log_{10} x,$$

and 
$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n \log_{10} x_i = \log_{10} \sqrt[n]{x_1 \times x_2 \times \dots \times x_n},$$

i.e. 
$$\text{mean } y = \log_{10} (\text{geometric mean } x).$$

The mean of a sample drawn from a normal population is itself distributed normally with a variance equal to that of the parent population divided by the number in the sample, and therefore appropriate confidence limits may be calculated; since the *geometric mean* of a sample of  $n$  observations is related to the arithmetic mean of the transformed values in the same way as a single observation of  $x$  is to the transformed value  $y = \log_{10} x$ , for all values of  $n$ . If the limits for use with the arithmetic mean of the untransformed data are required, estimates of the population mean will have to be made from the actual numbers observed. Finney (1941) has shown that for data of this type

the usual estimate of the sample mean, i.e.  $\frac{1}{n} \sum_{i=1}^n x_i$  decreases in efficiency as an estimator of the population mean as the coefficient of variation increases. He has given corrections which can be used, but it is probably better to perform all comparisons on the transformed values, and thus also to avoid confusion between the geometric and arithmetic means.

#### THE METHOD OF WINSOR & CLARKE

The data considered by Winsor & Clarke consisted of catches from plankton nets hauled vertically, obliquely and horizontally. In each case similar methods of analysis were used. The observations were given in the form of estimated total catches of each of given groups of animals (defined by species, age and sex) calculated from laboratory samples, which represented numbers caught by nets hauled consecutively, and therefore effectively through the same body of water. The subsequent analysis of variance divided the total variance observed into its component parts.

In order that such an analysis of variance may be performed the data should satisfy several conditions, one of which is that the variances within different groups and due to different factors must be estimated on comparable scales, or 'equalized', and the resultant residual variations should be normal with zero mean. A transformation from the raw data may result in variates which satisfy these conditions.

Winsor & Clarke applied the logarithmic transformation and subsequent analysis of variance, which they used to estimate the variances and appropriate coefficients of variation. To illustrate this latter step it is best to quote from their paper.

From the estimates 0.00781 and 0.00600 for the within haul and haul to haul variances, we obtain 0.0884 and 0.0775 as the estimated standard deviations. These are obtained from the logarithms of the catches. To interpret these figures in terms of variability of the actual catches we proceed as follows:

$$0.0884 = \log 1.226,$$

$$0.0775 = \log 1.195.$$

Now a deviation of 0.0884 in the logarithm of the catch means that the catch has been multiplied (or divided) by 1.226. Hence we may say that one standard deviation in the logarithm corresponds to a percentage standard deviation, a coefficient of variation, of 22.6% in the catch. Similarly, a logarithmic standard deviation of 0.0775 corresponds to a coefficient of variation of 19.5%.

This implies that the relation between  $\sigma_y$  and  $V_x$  is

$$\sigma_y = \log_{10} (1 + V_x), \quad (6)$$

but from equation (5) we see that this is not so. Equation (6) will always overestimate the value of  $V$  for a given  $\sigma_y$ , since it gives

$$V_x = 10^{\sigma_y} - 1, \quad (7)$$

whereas from (5)  $V_x = \sqrt{(10^{\sigma_y^2/0.43429} - 1)}$ . (8)

and the magnitude of the bias may be calculated. For the case quoted by Winsor & Clarke, calculation from equation (5) gives

$$\sigma_x^2 = 0.00781 = 0.43429 (1 + V_x^2),$$

$$1 + V_x^2 = \text{antilog}_{10} \frac{(0.00781)}{(0.43429)} = 1.0422,$$

and  $V_x = 0.205$  or 20.5%

for which Winsor & Clarke obtained 22.6%.

For  $\sigma_y^2 = 0.0400$  eqn. (5) gives 48.6% and the method of Winsor & Clarke 58.5%; and  $\sigma_y^2 = 0.0900$  eqn. (5) gives 78.2% and the method of Winsor & Clarke 99.5%.

The bias due to the method of Winsor & Clarke increases with increase in  $\sigma_y^2$ .

#### THE METHOD OF SILLIMAN

Silliman's data consisted of observations on the number of pilchard eggs in hauls made with two identical nets, each hauled once at a number of different stations. The data was not given as estimated catch, but as the numbers

counted in laboratory samples. Two laboratory samples were taken from each haul and both counts were given. This data, like that of Winsor & Clarke, has a constant coefficient of variation, and a logarithmic transformation was used. The analysis of variance showed a highly significant variation between stations; estimates of the 'within station between haul' and the 'between samples' variances were obtained.

Winsor & Clarke suggested the use of confidence limits in order to compare catches, and if these are calculated from the numbers obtained in several hauls from a given station, or at a given time, they provide a test of significance for a further observation from, say, a different station, or at a different time.

From equations (4) and (5) it is seen that the 95% limits for the untransformed distribution correspond to the antilogarithms of the values for the 95% limits for the transformed (normal) distribution; and are therefore  $\text{antilog}_{10}(\eta - t\sigma_y)$  and  $\text{antilog}_{10}(\eta + t\sigma_y)$ . If these are to be expressed as proportions of the mean of the untransformed distributions, we have

$$\frac{\text{antilog}_{10}(\eta - t\sigma_y)}{\xi} \quad \text{and} \quad \frac{\text{antilog}_{10}(\eta + t\sigma_y)}{\xi}$$

$$\text{or} \quad \log_{10} \frac{(\text{lower limit})}{\text{mean}} = -t\sqrt{[0.43429 \log_{10}(1 + V_x^2)] - \frac{1}{2} \log_{10}(1 + V_x^2)}, \quad (9)$$

$$\log_{10} \frac{(\text{upper limit})}{\text{mean}} = +t\sqrt{[0.43429 \log_{10}(1 + V_x^2)] - \frac{1}{2} \log_{10}(1 + V_x^2)}, \quad (10)$$

in which form they are expressed in terms of the coefficient of variation only. Silliman, recognizing this property, has argued as follows:

Solution of these equations gives  $\sigma_H^2 = 0.024$  as the variance of hauls and  $\sigma_S^2 = 0.007$  as the variance of samples. Therefore  $\sigma_H = \sqrt{0.024} = 0.155$  and  $\sigma_S = \sqrt{0.007} = 0.084$ . The 95% fiducial limits are of interest and may be calculated from these estimates of  $\sigma$ . For hauls  $2 \times 0.155 = 0.310$  ( $2\sigma$  limits include 95% of the distribution). Since these values are logarithmic, the antilogs are used to convert to ratios. The antilog of 0.310 is 2.04 giving fiducial limits of 49% ( $100 \times 1/2.04$ ) to 204% ( $100 \times 2.04$ ). Thus the egg number from one haul may not be considered significantly different from the egg number at another, unless it is less than one half, or more than double, that of the other. Similarly, the 95% fiducial limits for samples are 68% to 147%, and may be interpreted in a like manner.

This method implies that the ratios are

$$\frac{\text{antilog}_{10}(\eta - 2\sigma_y)}{\text{antilog}_{10} \eta} \quad \text{and} \quad \frac{\text{antilog}_{10}(\eta + 2\sigma_y)}{\text{antilog}_{10} \eta},$$

i.e. the mean employed is the antilogarithm of the mean of the transformed distribution, the geometric mean of the untransformed data. Similarly, estimates of the confidence limits for the mean of a sample are expressed as percentages of the geometric mean of the sample. Throughout Silliman has used the word mean, which is generally taken to apply only to the arithmetic mean.

If the arithmetic mean is used in this way, where the method is strictly only applicable to the geometric mean, anomalous results may be obtained. A numerical example taken from Silliman's data may illustrate the different results obtained with the use of the geometric and arithmetic means.

The laboratory counts of pilchard eggs for two samples drawn from each of two hauls taken at twenty-four different stations were given. Do the mean values of the first samples differ significantly between the two hauls? The means of the first samples are:

	Haul A	Haul B
Mean of actual numbers	70.5	85.6
Mean of logarithmic values	1.6499	1.7031
Geometric mean catch	44.66	50.48

The variance to be associated with a single haul, obtained from the analysis of variance, and based on 69 degrees of freedom, is 0.031.

The variance of the mean of 24 hauls is therefore  $0.031/24$   
 and the standard deviation 0.0359  
 $2 \times$  standard deviations\* 0.0718

and the confidence limits are  $\text{antilog } 1.6499 + 0.0718 = 52.69$   
 $\text{antilog } 1.6499 - 0.0718 = 37.85.$

From which it is seen that the geometric mean of the samples for the B hauls falls within the confidence limits derived from the A hauls. If the limits are calculated following Silliman and applied to the arithmetic mean of the untransformed observations, they are:

Lower limit 84.7% of 70.5 = 59.11,  
 Upper limit 117% of 70.5 = 83.19,

and the observed value of 85.6 for the B hauls appears to be significant as it falls outside the confidence limits derived from the A hauls. This result differs from that obtained above.

For confidence limits for a single observation the limits in actual numbers given by the two methods are equal as there is no difference between the estimate of the arithmetic mean and the geometric mean for a sample of one.

#### DISCUSSION

Although the estimates of the coefficient of variation as given by Winsor & Clarke have been shown to have been incorrectly obtained, the correct coefficient of variation does give a useful indication of the variability to be expected with any particular method and gear. The estimates given in this paper are lower than those given by them, but for their data are nevertheless high. For critical comparisons of catches, using the same gear and method confidence

\* In this case  $t=2$  for 95% probability level and 69 degrees of freedom.

limits are more useful. These may be calculated from the transformed data and expressed as actual numbers for a single observation, or as limits for the geometric mean of a number of catches calculated as percentages of the geometric mean of the observed data. This is equivalent to performing the comparisons on the logarithmic values.

It must be emphasized that the derivation of estimates, both of the coefficient of variation and of confidence limits, given in this paper, relies on the assumption that the logarithmic transformation employed has normalized the observed distribution. In practice this will not be completely realized, but the estimates given will be better than those obtained by the methods of Winsor & Clarke and Silliman, which take little account of the mathematical relations involved.

For some observed data, a logarithmic transformation of the form  $y = \log_{10}(x + a)$  is more appropriate and, from the equations given above, estimates of the coefficient of variation can be obtained. For such a transformation, the variance of the untransformed data remains  $\sigma_x^2$ , but the mean becomes  $(\xi + a)$ . From Fig. 1 it will be seen that confidence limits for a single observation depend only on area, and may therefore be obtained for any *normalizing* transformation, even when the mathematical relations between the parameters of the two distributions are not readily available. Comparisons of sample means may be carried out on the transformed values.

Thanks are due to Mr N. W. Plese for his helpful comments on this paper.

#### REFERENCES

- BARNES, H. & BAGENAL, T. B., 1951. A statistical study of variability in catch obtained by short repeated trawls taken over inshore ground. *J. Mar. biol. Ass. U.K.*, Vol. 29, pp. 649-60.
- FINNEY, D. J., 1941. On the distribution of a variate whose logarithm is normally distributed. *J. R. statist. Soc.*, Suppl. Vol. 7, pp. 155-61.
- JOHNSON, N. L., 1949. Systems of frequency curves generated by methods of translation. *Biometrika*, Vol. 36, pp. 149-76.
- SILLIMAN, R. P., 1946. A study of variability in plankton towner catches of Pacific pilchard *Sardinops caerulea* eggs. *J. Mar. Res.*, Vol. 66, 74-83.
- SNEDECOR, G. W., 1946. *Statistical Methods*. Fourth edition. Iowa State College Press.
- WICKSELL, S. D., 1917. On the genetic theory of frequency. *Ark. Mat. Astr. Fys.*, Vol. 12, No. 20. pp. 1-56.
- WINSOR, C. P. & CLARKE, G. L., 1940. A statistical study of variation in the catch of plankton nets. *J. Mar. Res.*, Vol. 3, pp. 16-34.