LOGNORMAL DISTRIBUTIONS

By Prof. J. H. GADDUM, F.R.S.
University of Edinburgh

A LARGE part of statistical theory is based on the assumption that measurements are distributed in normal probability curves and that the variance is constant. The normal curve was discovered by de Moivre in 1753 and developed as a useful mathematical tool by Laplace, Gauss, Maxwell and others. It has a number of interesting properties, and various attempts were once made to establish a law of Nature known as the 'normal law of errors', which implied that measurements ought always to be distributed in this way. These attempts were undeservedly successful. The use of the adjective 'normal' in connexion with this law implies more than is justified by the facts, whereas in fact the adjective is established by custom and provides a convenient label for distributions of this type.

Referring to the normal law in the introduction to his "Thermodynamique", H. Poincaré (1892)[1] quotes the remark that "Everybody firmly believes in it because the mathematicians imagine it is a fact of observation, and observers that it is a theory of mathematics". Since that time the mathematical conditions for normality have been firmly established[2], but the best evidence that these conditions are fulfilled in any particular case is still the observation that the distribution actually is normal. It is therefore important, either to show that observations are normally distributed, or to convert them into a form which is normally distributed, or at least has the best possible chance of being normally distributed.

In some cases the normal curve gives a very close approximation to the observed facts. These cases are the exception rather than the rule; but it is usually possible to transform the distribution by means of some function of the actual observations which is normally distributed, so that if this function is used instead of the observations themselves the subsequent calculations are simplified and their scope is extended. This device is well known to statisticians, but not used so widely as it should be. It is obviously justified in simple cases. Suppose, for example, that the original observations were measurements of the diameters of drops in an emulsion (d) and that it was found that the distribution of d was not normal, but that the distribution of d^3 was normal. This would mean that the volume of the drops was normally distributed, and would show that it would be more convenient to consider the distribution of volumes than the distribution of diameters. The size of the drops can, in fact, be measured in a number of different ways. There is no particular reason why the original observations should represent the most convenient way, and no particular reason why the most convenient way should not be used even if it is less obviously reasonable than in the example given above.

"The linear scale, since it was first cut on the wall of an Egyptian temple, has come to be accepted by man almost as if it were the one unique scale with which Nature builds and works. Whereas it is nothing of the sort" (Bagnold[3]).

Even if one method of measurement gives normal distributions, it necessarily follows that most others will not; and since it is unlikely that the appropriate method will often be chosen first, it is only to be expected that most distributions will not be normal unless care is taken to select the appropriate method of measurement[4].

The second condition for the unfettered use of statistics is that the variance of the distribution shall be constant, or at least shall vary in some predictable way. *If these two conditions are fulfilled, it is easy* to calculate the significance of differences, regression lines, correlation coefficients, the analysis of variance and so on; but, if not, the possibilities are very restricted. For example, if the distribution is not normal, it is unjustifiable even to assume that the arithmetic mean is the best estimate of a quantity that can be derived from a set of measurements of it. Recent developments in statistical technique have thus greatly increased the importance of methods for normalizing distributions by altering the method of measurement, or by transformation of each observation by some suitable device such as taking its logarithm. This is particularly important when the standard deviation is large compared with the mean. When it is small, all ordinary transformations of this kind have less effect, and in the limit when it is very small no such transformation has any effect at all.

The theoretical justification for using normal curves depends on the mathematical fact, which is true within certain defined limits[2], that the sum, or arithmetic mean, of a large number of variables tends to be normally distributed even when the individual variables are not. If an observation (x) depends on the sum of the effects of a large number of small causes acting at random, and if each effect is independent of x, the distribution is likely to be normal. If it is not normal, we assume that this is because x is not measured in the right way. We find that $f(x)$ is normally distributed and we therefore transform our observations of x by converting them into observations of $f(x)$. The normal shape of the distribution of $f(x)$ is presumably due to the random combination of numerous causes each of which has a small effect on $f(x)$, and this effect is independent of $f(x)$. The effect of each of these causes on x will be proportional to $dx/d\ f(x)$, and Kapteyn[5] calls this quantity the 'reaction'. For example, if $f(x) = \log x$, the reaction is x. The effect of each small cause on $\log x$ is independent of $\log x$ (and of x); its effect on x is therefore proportional to x. If x is an astronomical observation involving a reading on a dial, the absolute value of x is usually irrelevant; the reaction is likely to be constant and the distribution is likely to be normal. In most scientific observations, however, the absolute value of x is not irrelevant. A fine day which adds a given weight to a large animal is not likely to add the same weight to a smaller animal. The gains in weight are more likely to be proportional to the size of the animal. In other words, the effect of the fine day on the logarithm of the weight is more likely to be constant than its effect on the weight, so that the logarithm of the weight is likely to be normally distributed. The assumption that this will be so is not likely to be exactly true; but it is likely to give a closer approximation to the facts than the assumption that the effect is independent of the weight.

The problem has been tackled in a rather different way by Curtiss[6], who discusses methods of calculating the appropriate transformation from various assumptions regarding the relation between measurements and their standard errors, and comes to the conclusion that when the error is proportional to the measurement the use of logarithms is likely to produce normal curves.

Testing Transformations

It is usually necessary to discover the best method of normalization by trial and error. Standard statistical methods such as the calculation of goodness of fit and moments give the ultimate criterion of the success of any particular method of normalization, but a simple graphical method indicates at once not only which transformations are successful, but also in what ways other transformations fail.

This graphical method consists in first applying the appropriate transformation to the observations and then plotting the results as abscissæ either on probability paper, or against probits as ordinates. The probit is best defined in terms of the normal equivalent deviation (N.E.D.). The N.E.D. corresponding to any given percentage is calculated from the shape of a normal curve the standard deviation of which is one ; it is the deviation (from the mean) equivalent to the given percentage of the area of the curve. The probit is equal to the N.E.D. + 5. This is now a standard technique[7] and need not be discussed in detail. It has been widely used in experiments on toxicity, when the data are obtained as a relation between the dose of a drug and the percentage mortality which it causes[8,9]. Each mortality is converted, by means of tables, to a probit. If the animals vary in such a way that the dose (or transformed dose) is normally distributed, the probit should bear a linear relation to the dose (or transformed dose). It is usually obvious at once whether the plotted points are distributed at random about a straight line ; and, if they are, it is possible with practice to draw a straight line which fits the results well enough for most purposes. Regression lines may be fitted if necessary, and it is possible to obtain the solution giving maximum likelihood by a suitable technique using successive approximations, which was discovered after less satisfactory methods had been used[10]. Any regular tendency for the points to diverge from a straight line shows that the distribution is not normal, and how it differs from the normal curve. The reciprocal of the slope gives an estimate of the standard deviation when the line is straight, and is proportional to the 'reaction' at each point when the line is not straight.

The primary use of this graphical method is where the data are limited to the percentages corresponding to different values of the variable (for example, dose) ; but the same method is sometimes helpful even when fuller information is available. If n observations have been made of a given variable, one method is to arrange them in order of size and then allot to the smallest observation a percentage of $100/2n$ and to succeeding observations percentages of $300/2n$, $500/2n$ . . . $(2n-1)100/2n$. These percentages are then converted to probits and each individual observation is plotted. If the data are numerous they are grouped and added cumulatively, and the probits are plotted against the points separating the groups. When the number in a group is very small it is best to plot individual readings, or to assume that the observations are evenly distributed in the range covered by the group. If straight lines are obtained the distributions are normal, and the mean and standard deviation can be estimated fairly accurately from the graph. When results are treated in this way, the usual technique for calculating a regression line is inapplicable. The best estimates of the mean and standard deviation will be obtained by applying the ordinary methods directly to the transformed observations, but in some cases they can also be estimated from the moments of the original distribution. This is likely to be the most convenient method when the original observations are grouped, but its scope is limited. Examples are given below.

The use of probits also provides a general graphical method of normalizing distributions without the use of any formulæ. This may be useful when the scale on which the results are measured is a purely arbitrary one. If a random sample of observations is plotted against probits and a smooth curve is drawn through the results, this curve may be used for converting subsequent observations to a scale of probits, which are, of course, normally distributed. It is important in this case to be certain that the shape of the original curve and the variance of the transformed curve are stable. This principle has been applied by Ferguson[11,12] to the results of mental tests.

The Transformation $X = \log x$

The theoretical justification for using this transformation for most scientific observations is probably better than that for using no transformation at all. The general arguments in favour of it were discussed by Galton[13], who also directed attention to the fact that the normal law predicts negative observations. The existence of men of more than double the average weight implies the existence of other men with negative weight. When logarithms are used this difficulty does not arise.

When measurements are made of the size of a number of small objects of the same shape, it is often possible to measure either their diameters or their volumes. If the distribution of the diameters is normal, then the distribution of the volumes will necessarily be asymmetrical, and vice versa. The normal law cannot be true in both cases. The use of logarithms removes this difficulty[5]. If the logarithms of the diameters are distributed normally with standard deviation λ, the logarithms of the volumes will be distributed normally with standard deviation $3\,\lambda$.

These theoretical arguments thus lead to the conclusion that this transformation should facilitate the interpretation of results when the variations are large ; when they are small it can at the worst do very little harm, since it has very little effect on the shape of the curve. The extra labour involved in converting the observations to logarithms to base 10 is small, though perhaps sufficient to deter those who deal with small variations. Mathematicians prefer natural logarithms, but experimenters usually prefer ordinary logarithms to base 10, and these give equally good results. The symbol λ is used[8] to denote the standard deviation of the logarithms to base 10.

One logical consequence of the adoption of this method would be that the mean of the logarithms, or the geometric mean of the observations, would be taken as the most likely value, instead of the arithmetic mean. Williams[14,15] obtained direct evidence of the value of this in experiments on the number of insects caught in a light trap. The use of logarithms had the double advantage of making the results more consistent and of preventing the result from depending almost entirely on one aberrant large catch. It is likely that this device would increase the precision of most experiments in which the variation is large.

If the original observations have been grouped on an arithmetic scale, the direct computation of the constants of the transformed distribution may be laborious. These constants can be calculated from the mean ($\bar{x}$) and standard deviation (σ) of the original distribution, but the estimates so obtained are only reasonably efficient when λ is less than 0·14[16]. In this region λ can be estimated within 3 per cent by dividing the coefficient of variation by 231. The general formulæ are

$$\bar{X} = \log_{10} (\bar{x}/(1 + \sigma^2/\bar{x}^2)^{1/2})$$
$$\text{and } \lambda^2 = 0{\cdot}4343 \log_{10} (1 + \sigma^2/\bar{x}^2).$$

Lognormal Distributions

It is proposed to call the distribution of x 'lognormal' when the distribution of log x is normal. Lognormal distributions have been discovered in many fields of work. Kapteyn[4] quotes the example of some data by Heymans on the threshold of sensation as measured by the smallest weight which was just perceptible on the skin. Wightman, Trivelli and Sheppard[17] found that the size of the particles of silver in a photographic emulsion were lognormally distributed. The diameters and areas of projection were measured, and it appeared at first that the form of the distribution of both measurements varied in different experiments. When the logarithmic transformation was used, normal curves fitted all the measurements of both diameter and area.

The distribution of sensitivities to drugs among individual animals of the same species, as measured by the dose required to cause some definite effect, has been widely studied. The sensitivity may vary over a tenfold range or more, but in practically every case the logarithm of the sensitivity is normally distributed or nearly so[8,18]. This fact, which is now generally accepted, has proved very convenient in the design of toxicity tests and in the calculation of their errors. The values of λ range from 0·014 to 0·91.

Similar methods have been applied to the action of disinfectants[19] and it has been found that the relation between time and the death-rate of bacteria can be explained, in some cases at least, on the theory that the logarithm of the survival time is normally distributed.

Hemmingsen[20] studied the distribution of the average size of the different species in each of various phylogenetic groups and found that they were distributed approximately lognormally (λ = 0·083–0·673).

Other examples of lognormal distributions have been found in estimates of the numbers of plankton caught in different hauls with a net[21], and in the amounts of electricity used in medium-class homes in the United States[22].

Bacteriologists have to deal with wide ranges of variation, as is illustrated by the data of Brew[23], who calculated coefficients of variation up to 640 per cent in bacterial counts. Such data can only be effectively treated by some such device as that discussed here, and bacteriologists often use a logarithmic scale.

Wechsler[24] collected a large number of data relating to measurements of human beings, and had trouble with their interpretation. The distributions were sometimes normal, but when the variation was large, the curves were skew with the mode less than the mean. He found that it was best to express the range of variation in terms of a ratio, and used for this purpose the ratio of the measurement on the 999th individual out of 1,000 to the measurement on the second individual. When there were less than 1,000 individuals, he was at a loss for a method of extrapolation and took the total range actually covered by the observations. The curves obtained are just the kind which are improved by taking logarithms. A number of the distributions studied by Wechsler are, as a matter of fact, lognormal, or approximately so, and there can be little doubt that their interpretation would have been facilitated by taking logarithms. The values for λ in some of these data have been estimated as follows: height, 0·015, 0·0164, 0·0172, 0·017; blood sugar, 0·029; weight, 0·045, 0·055; blood pressure, 0·049; pulse-rate, 0·061, 0·067. It will be seen that λ for weight is about three times λ for height, as is to be expected theoretically.

Cooper[25] made a number of measurements of the size of drops in emulsions. Some of his distributions were apparently complex, but in the only example for which details are given the diameters (and volumes) are distributed lognormally to a close approximation (λ = 0·17 for diameter).

Cochran[26] discusses various transformations and comes to the conclusion that the logarithmic one is particularly effective, but that it makes no significant difference when the coefficient of variation is less than 12 per cent. He gives evidence that the standard deviation was proportional to the mean in measurements of the reaction time of human beings in a word test, and concludes that the distribution is likely to be lognormal.

Williams[15,27] has applied the same transformation successfully to data in which the variable was not an actual physical measurement but a count of the number of insects caught in a light trap, or the number of words in a sentence by G. B. Shaw.

Some of the effects of drugs on enzymes, and of oxygen on hæmoglobin, can be explained on the theory that the molecules of a protein in a solution show continuous variations among themselves, and that certain of their properties are lognormally distributed[28].

The Transformation $X = \log (x + x_0)$

If the probit is plotted against log x and the points diverge from a straight line, it is sometimes possible to find an explanation in such factors as a failure to allow for the mortality occurring when no drug is given, or a failure to count drops too small to be seen. In some cases it may be desirable to find a transformation giving a closer fit to the data than the simple logarithm. One convenient formula is

$$X = \log (x + x_0).$$

This is useful when the curve shows a more or less constant curvature, or with flat curves which diverge most at their lower ends. If the curve is convex upwards with probits as ordinates, x_0 is negative; if it is concave upwards, x_0 is positive. If a good fit is obtained, it suggests that the variations are proportional to the amount that the variable exceeds the value $- x_0$. If full data are available the constants of the curve can be estimated from the mean and the second and third moments of the original distribution[29,30]. Kapteyn[4] gives another method. A rough estimate of x_0 may also be obtained from a graph in the following way. Take three equidistant points on the scale of probits, such as 4, 5 and 6. Estimate from the curve the corresponding values of log x and thus of x. If these values are x_1, x_2 and x_3, then we wish to determine x_0 so that $(x_2 + x_0)$ shall be the geometric mean of $(x_1 + x_0)$ and $(x_3 + x_0)$, so that $(x_2 + x_0)^2 = (x_1 + x_0)\ (x_3 + x_0)$, or $x_0 = (x_2{}^2 - x_1\ x_3)/(x_1 + x_3 - 2x_2)$. The success of this correction can be tested by plotting log $(x + x_0)$ against probits. For extrapolation at one end of the curve it may be convenient to choose x_1, x_2 and x_3 near that end.

This transformation has been fitted to observations of the value of house property and of the size of the foreheads of crabs[4], the ages of employees of a commercial firm and the number of petals on a buttercup[29] and the weights of female students[30]. The curves obtained in this way fitted the observations better than formulæ obtained by the methods described by Pearson.

Curves of this type can be fitted to measurements recorded by Nagelschmidt[31] of the diameters of particles of airborne dust in coal mines. In some cases the distributions appear lognormal ($\lambda = 0{\cdot}2$–$0{\cdot}4$), but in others the curves obtained when probits were plotted as ordinates against log diameter were concave upwards. When the transformation $X = \log (x + x_0)$ was used, the calculation of χ^2 showed that the fit was good.

Williams[15] used this transformation with $x_0 = 1$ in order to avoid complications in calculating the geometric mean when $x = 0$. This only affects the shape of the curve for small values of x.

Conclusion

If it were the normal custom, when scientific observations show uncontrolled variations large compared with the observations themselves, to convert them to logarithms before estimating their mean or variance, the usual result would be an increase in the accuracy and scope of the conclusions drawn from them.

[1] Poincaré, H. (1892), quoted from Mellor, J. W., "Higher Mathematics for Students of Chemistry and Physics" (Longmans, Green and Co., 1922).

[2] Cramér, H., "Random Variables and Probability Distributions", Cambridge Tract. Math., 36 (1937).

[3] Bagnold, R. A., "The Physics of Blown Sand and Desert Dunes" (London: Methuen, 1941).

[4] Kapteyn, J. C., "Skew Frequency Curves in Biology and Statistics" (Groningen: Noordhoff, 1903).

[5] Kapteyn, J. C., *Rec. Trav. Bot. Neerland*, **13**, 105 (1916).

[6] Curtiss, J. H., *Ann. Math. Statist.*, **14**, 107 (1943).

[7] Fisher, R. A., and Yates, F., "Statistical Tables" (Edinburgh: Oliver and Boyd, 1938).

[8] Gaddum, J. H., Med. Res. Coun. Spec. Rep., 183 (1933).

[9] Irwin, J. O., *J. Roy. Stat. Soc.*, Suppl. **4**, 1 (1937).

[10] Bliss, C. I., *Quart. J. Pharm.*, **11**, 192 (1938).

[11] Ferguson, G. A., *Psychometrika*, **7**, 19 (1942).

[12] Finney, D. J., *Psychometrika*, **9**, 31 (1944).

[13] Galton, F., *Proc. Roy. Soc.*, **29**, 365 (1879).

[14] Williams, C. B., *Ann. Appl. Biol.*, **24**, 404 (1937).

[15] Williams, C. B., *Trans. Roy. Entom. Soc. Lond.*, **90**, 227 (1940).

[16] Finney, D. J., *J. Roy. Stat. Soc.*, Suppl. **7**, 155 (1941).

[17] Wightman, E. P., Trivelli, E. P. H., and Sheppard, S. E., *J. Phys. Chem.*, **28**, 529 (1924).

[18] Bliss, C. I., and Cattell, M., *Ann. Rev. Physiol.*, **5**, 479 (1943).

[19] Withell, E. R., *J. Hyg.*, **42**, 124 (1942).

[20] Hemmingsen, A. M., *Vidensk Medd. fra Dansk naturh. Foren.*, **98**, 125 (1933).

[21] Snedecor, G. W., "Statistical Methods" (Iowa State College Press, 1940).

[22] Croxton, F. E., and Cowden, D. T., "Applied General Statistics" (New York: Prentice-Hall, 1939).

[23] Brew, J. D., *J. Dairy Sci.*, **12**, 304 (1929).

[24] Wechsler, D., "The Range of Human Capacities" (Baillière, Tindall and Cox, 1935).

[25] Cooper, F. A., *J. Soc. Chem. Ind.*, **56**, 447 T (1937).

[26] Cochran, W. G., *Empire J. Exp. Agric.*, **6**, 157 (1938).

[27] Williams, C. B., *Biometrika*, **31**, 356 (1940).

[28] Gaddum, J. H., *Proc. Roy. Soc.*, B, **121**, 598 (1937).

[29] Fisher, Arne, "The Mathematical Theory of Probabilities" (The Macmillan Company, 1922).

[30] Yuan, P. T., *Ann. Math. Statist.*, **4**, 30 (1933).

[31] Nagelschmidt, G., Med. Res. Coun. Spec. Rep. 244 (1943).

NEW VIEWS OF THE ORIGIN OF THE SOLAR SYSTEM

By Prof. W. H. McCREA
Royal Holloway College, University of London

THE field is still open for the formulation *de novo* of theories of the origin of the solar system. No observational test has ever been discovered which would impose a definitive restriction upon the basic possibilities, as, for example, to show whether or not we must look to some catastrophic event to produce the necessary initial conditions.

Interest in the problem has recently been revived by attempts to approach it afresh along three totally independent lines. Von Weizsäcker[1] proposes what amounts to a re-instatement of the nebular hypothesis, with the incorporation of several novel features; Alfvén[2] makes the radical suggestion that it was the magnetic field of the sun which governed the manner in which it acquired the material of the planets from interstellar space; Hoyle invokes a nova outburst as a catastrophic origin of this material. None of these theories can as yet claim completely to demonstrate that a planetary system must inevitably have resulted from its postulated initial conditions. But they all offer apparent explanations of some or other of the general characteristics of the actual planetary system. In spite of their diverse approaches, it is not impossible that further progress may be made by combining ideas drawn from more than one of them*.

At the same time, thanks to suggestions by Jeffreys and others, the idea is gaining ground that the basic agency in the formation of the solid substance of planets may be no more than the most common form of condensation from a vapour to a solid. This may prove to be a simplifying factor in the whole problem.

Weizsäcker's Theory

Weizsäcker's theory belongs to the class of those which seek the origin of a planetary system in a rotating nebulous envelope around an already existing sun. His fundamental hypothesis is that the chemical composition of the envelope was initially the same as that of the sun itself. Since the chemical composition of the present planets is very different (comprising a bigger proportion of the heavier elements) they can then represent only a fraction of the original mass of the envelope. Weizsäcker takes this fraction to be as low as one per cent. For he accepts a recent estimate by Biermann[3] for the composition of the sun, which gives it a much higher content of the lighter elements than is currently assumed. In this manner he proposes to overcome the difficulty, encountered by older theories of this class, that the initial density of the envelope would have been too small to lead by the ordinarily recognized gravitational processes to condensation into planetary bodies. At the same time, through the dissipation of the surplus 99 per cent composed of light elements, he proposes to account for the disappearance of an embarrassing amount of angular momentum.

Weizsäcker envisages several stages in the evolution of his system, each of which he investigates in detail. But it has to be said that the way in which it

* The meeting of the Royal Astronomical Society on July 13, 1945, was devoted to a discussion of these theories, which were outlined respectively by the present writer, Dr. A. Hunter, and Mr. F. Hoyle. Subsequent speakers were Prof. H. H. Plaskett, Prof. S. Chapman, Dr. H. Jeffreys, Dr. G. P. Kuiper and others. This article is a digest of what was said, with some slight additional commentary; it does not treat other recently proposed theories not included in the discussion, such as that of J. B. S. Haldane, *Nature*, **155**, 133 (1945).