# On Transforming Biological Data to Gaussian Form

J. Holt, J. H. Lumsden and K. Mullen*

## ABSTRACT

Much of the statistical analysis of biological data depends on the assumption that the data are Gaussian (or normal). Some well-known procedures which use this assumption are (i) t-tests (ii) analysis of variance (iii) regression estimation and their attendant tests. If the data are not Gaussian, one can use nonparametric statistical techniques, if they exist, but they often require larger amounts of data to obtain equally precise results (see for example Lumsden and Mullen (7) for a discussion of this with regard to reference value estimation). If the data are not Gaussian a fruitful approach to their analyses lies in trying to find a transformation which will render them Gaussian. The data thus transformed to a Gaussian form, can be analyzed validly using standard statistical techniques. The process of finding a good transformation of the data has often been an arbitrary and ad hoc one. The purpose of this article is to look at a particular technique for attempting to render nonGaussian data Gaussian, and to illustrate its applicability and breadth of use.

## RÉSUMÉ

La majeure partie de l'analyse statistique des données biologiques dépend de la supposition que ces données présentent une distribution équivalente à celle de Gauss, i.e. normale. Voici quelques procédés bien connus qui utilisent cette supposition: 1. les tests-t, 2. l'analyse des écarts et 3. l'appréciation de la régression, ainsi que les tests qui les accompagnent. Si les données obtenues ne possèdent pas une distribution équivalente à celle de Gauss, on peut les analyser à l'aide de techniques statistiques non paramétriques, pourvu que de telles techniques existent; il faut cependant plus de données, de façon à obtenir des résultats aussi précis. Consultez par exemple l'article de Lumsden et Mullen (7), pour une discussion sur l'évaluation des valeurs de référence. Lorsque la distribution des données ne s'avère pas équivalente à celle de Gauss, une façon fructueuse d'aborder leur analyse consiste à tenter de trouver une transformation qui les convertisse en données dont la distribution correspondra à celle de Gauss. On peut ensuite les analyser validement à l'aide de techniques statistiques standards. La façon de trouver une bonne transformation des données s'avère souvent arbitraire et ad hoc. Le but de cet article consiste à jeter un regard sur une technique particulière qui vise à transformer en données dont la distribution ne correspond pas à celle de Gauss, des données qui correspondent déjà à cette distribution; il veut aussi illustrer l'applicabilité et les possibilités d'utilisation de cette technique particulière.

## INTRODUCTION

This article is concerned with the problem of finding a transformation for rendering nonGaussian data Gaussian. Examples of the use and necessity of such transformations are discussed in the establishing of veterinary reference values by Lumsden and Mullen (7) and applications to canine and bovine hematology and biochemistry data are discussed in Lumsden *et al* (8,9). The problem considered in the above references

*College of Physical Science (Holt and Mullen) and Ontario Veterinary College (Lumsden), University of Guelph, Guelph, Ontario N1G 2W1.

Submitted March 23, 1979.

is as follows. For any characteristic, its reference values or reference interval is a pair of numbers, calculated from a randomly selected sample of healthy animals, within which 95% of the population values of healthy animals can be expected to lie. The calculation of the reference interval is made simpler and more efficient if the data are assumed to have a Gaussian distribution. If this is not so, then one can attempt to transform the data to the Gaussian form. If the attempt is successful, then Gaussian methods can be used on the transformed data; if it is unsuccessful, one alternative is to use nonparametric techniques for estimating the reference values, but these, although better than wrongly assuming that the data are Gaussian, do not permit one to calculate the confidence that can be placed in them for small sample sizes.

In Lumsden *et al* (8,9) if the data were not Gaussian, then they were subjected to four transformations, the square root transformation $\sqrt{x}$, the inverse transformation 1/x, the logarithmic transformation $\ell n\ (x + 0.5)$ and the arcsine transformation. Each set of transformed data was subjected to a test for normality (usually a chi-square goodness of fit test or the Kolmogorov-Smirnov test, both discussed in Lumsden and Mullen (7) and the most appropriately transformed set (if any) was processed using the Gaussian technique. The use of the aforenamed transformations is arbitrary and *ad hoc* and more recently we have tried a more unified and theoretically justified approach, due to Box and Cox (1), which will be discussed and applied here.

## STATISTICAL DISCUSSION

Standard statistical techniques require that the response (measurement) variable follow a Gaussian distribution, have a variance independent of the experimental treatments and be acted upon additively by the treatments. If these conditions are not met it is convenient to transform the response variable x to a new variable y which satisfies the above conditions. The alternative would be the utilization of sophisticated, intricate and possibly inefficient statistical methods. A detailed discussion of various methods of transforming data is given by Thöni (13) and less extensive but excellent discussions can be found in the intermediate level statistics books by Steel and Torrie (12) and Snedecor and Cochran (11). We propose to illustrate the use of a unified and broadly applicable approach suggested by Box and Cox (1) on some data which was previously analyzed by standard transformations applied in a somewhat *ad hoc* fashion.

Box and Cox suggest transforming the original response variable x to a new variable y means of the relation

$$y = \frac{(x + A)^B - 1}{B} \quad \text{if } B \neq 0$$

$$\text{or } y = \log (x + A) \quad \text{if } B = 0.$$

Varying A and B generates a rich class of transformations which includes the square root transformation (B = 0.5), the reciprocal transformation (B = −1). and the log transformation (B = 0). These latter transformations were considered by Lumsden and Mullen (7) and the present approach allows one to decide which, if any, is appropriate. It is assumed one member of the Box-Cox family of transformations leads to Gaussian data. The statistical problems are to determine the best choice for A and B and then to decide if the corresponding response variable does in fact conform to a Gaussian distribution. Fortunately, transforming to Gaussian form often (but not always) stabilizes the variance and achieves additivity.

One efficient criterion for judging which transformation is the best is the intuitively appealing one of choosing those values of A and B which make the given data set the one most likely to be observed. The term likelihood function is given to the expression to be maximized and the values of A, B and other parameters (mean $\mu$, variance $\sigma^2$) in the model which maximize this expression are called maximum likelihood estimates.

This method of maximum likelihood estimation was first suggested by Sir R.A. Fisher (4) and its optimal properties are discussed in detail by Cox and Hinkley (2). Numerically it is equivalent and often easier to maximize the loga-

TABLE I. Data on Healthy Cattle in the Age Group One to 14 Days

| Serum Iron (mg/dL) | Monocytes (x $10^3$/mL) | Neutrophil Bands (%) |
|---|---|---|
| 37 | 3 | 10 |
| 177 | 4 | 0 |
| 77 | 5 | 4 |
| 95 | 15 | 1 |
| 82 | 9 | 0 |
| 83 | 2 | 0 |
| 200 | 0 | 2 |
| 50 | 1 | 0 |
| 45 | 6 | 3 |
| 102 | 3 | 0 |
| 27 | 5 | 3 |
| 35 | 2 | 3 |
| 45 | 3 | 13 |
| 100 | 3 | 1 |
| 50 | 3 | 3 |
| 53 | 2 | 0 |
| 156 | 2 | 4 |
| 65 | 10 | 0 |
| 63 | 4 | 5 |
| 106 | 5 | 0 |
| 28 | 3 | 2 |
| 142 | 6 | 0 |
| 64 | 3 | 0 |
| 121 | 1 | 3 |
| 164 | 7 | 0 |
| 283 | 9 | 1 |
| 67 | 6 | 0 |
| 136 | 9 | 0 |
| 45 | 5 | 2 |
| 127 | 8 | 2 |
| 45 | 7 | 0 |
| 38 | 4 | 1 |
| 193 | 2 | 1 |
| 69 | 5 | 0 |
| 52 | 4 | 0 |
| 75 | 5 | 0 |
| 95 | 1 | 0 |
| 135 | 1 | 0 |
| 47 | 10 | 1 |
| 102 | 6 | 0 |
| 224 | 6 | 2 |
| 78 | | |
| 161 | | |

TABLE IIa. Values of the Log Likelihood Function, L, for the Serum Iron (mg/dL) Data

| | B | | | |
|---|---|---|---|---|
| A | − 0.5 | 0 | 0.5 | 1 |
| 0 | − 167.1 | − 165.9 | − 168.5 | − 174.6 |
| 1 | − 167.0 | − 166.0 | − 168.5 | − 174.6[a] |
| 2 | − 167.0 | − 166.0 | − 168.5 | − 174.6 |

[a]The values of L for the original untransformed data

rithm of the likelihood function. After replacing the mean and variance by their maximum likelihood estimates, the log likelihood function L, is proportional to

$$\frac{-n}{2} \log (s^2) + (B - 1) \sum_{i=1}^{n} \log (x_i + A)$$

where

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

and

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

are the sample variance and mean, respectively. The use of L to find optimal values of A and B is ideally suited for a conversational APL terminal and an APL program is available from the authors on request. Basically one searches by trial and error to find A and B which give the largest value of L. Once A and B have been found, and the x's have been transformed to the y's, the resulting transformed data should be tested, using for instance the Kolmogorov-Smirnov test, to see if the transformation is adequate. Examples of this will be given in the next section. If the transformation adequately renders the data Gaussian, then Gaussian techniques can be used to complete the analysis.

## Some Examples

In order to illustrate this technique we take three sets of data which were used in Lumsden *et al* (9). The three sets, on healthy cattle in the age group one to 14 days are (i) serum iron (mg/dL), (ii) monocytes ($10^3$/mL) and (iii) neutrophil bands (%). The data are presented in Table I.

The values of L, the log likelihood function, are shown in Tables IIa, b and c for a choice of A and B for each of the three sets of data.

An informal graphical check on whether a variable follows a Gaussian distrib-

TABLE IIb. Values of the Log Likelihood Function, L, for the Monocyte Data

| | B | | | | |
|---|---|---|---|---|---|
| A | − 0.5 | 0 | 0.5 | 1 | 1.5 |
| 1 | − 47.7 | − 42.5 | − 42.0 | − 45.8[a] | − 53.2 |
| 2 | − 44.0 | − 41.9 | − 42.5 | − 45.8 | − 51.7 |
| 3 | − 42.8 | − 41.8 | − 42.8 | − 45.8 | − 50.8 |
| 4 | − 42.3 | − 41.9 | − 43.1 | − 45.8 | − 49.6 |

[a]The value of L for the original untransformed data

TABLE IIc. Values of the Log Likelihood Function, L, for the Neutrophil Band Data

| | B | | | |
|---|---|---|---|---|
| A | − 1 | − 0.5 | 0 | 0.5 |
| 1 | − 10.8 | − 10.5 | − 14.2 | − 23.5 |
| 2 | − 13.3 | − 15.1 | − 19.5 | − 27.6 |
| 3 | − 15.5 | − 18.0 | − 22.6 | − 29.7 |

ution is supplied by a normal probability plot of the data. The ith smallest observation is plotted versus the standardized normal variable corresponding to a cumulative percentage of $\frac{i}{n}$ 100%. This latter standardized normal value is called a normal equivalent deviate, or more simply, a normal quantile or percentile. An approximately straight line plot is evidence for a Gaussian distribution. Figures 1a,b,c provides normal probability plots for the three data sets in Table I. It can be noted that all three distributions are decidedly nonGaussian. Also the cumulative percentages are usually modified slightly to $\frac{i-\frac{1}{2}}{n}$ 100% so as to preserve symmetry between the smallest and largest observations and to avoid the problem of plotting the largest observation at $+\infty$. The standardized observations $\frac{y-\bar{y}}{s}$ were plotted in

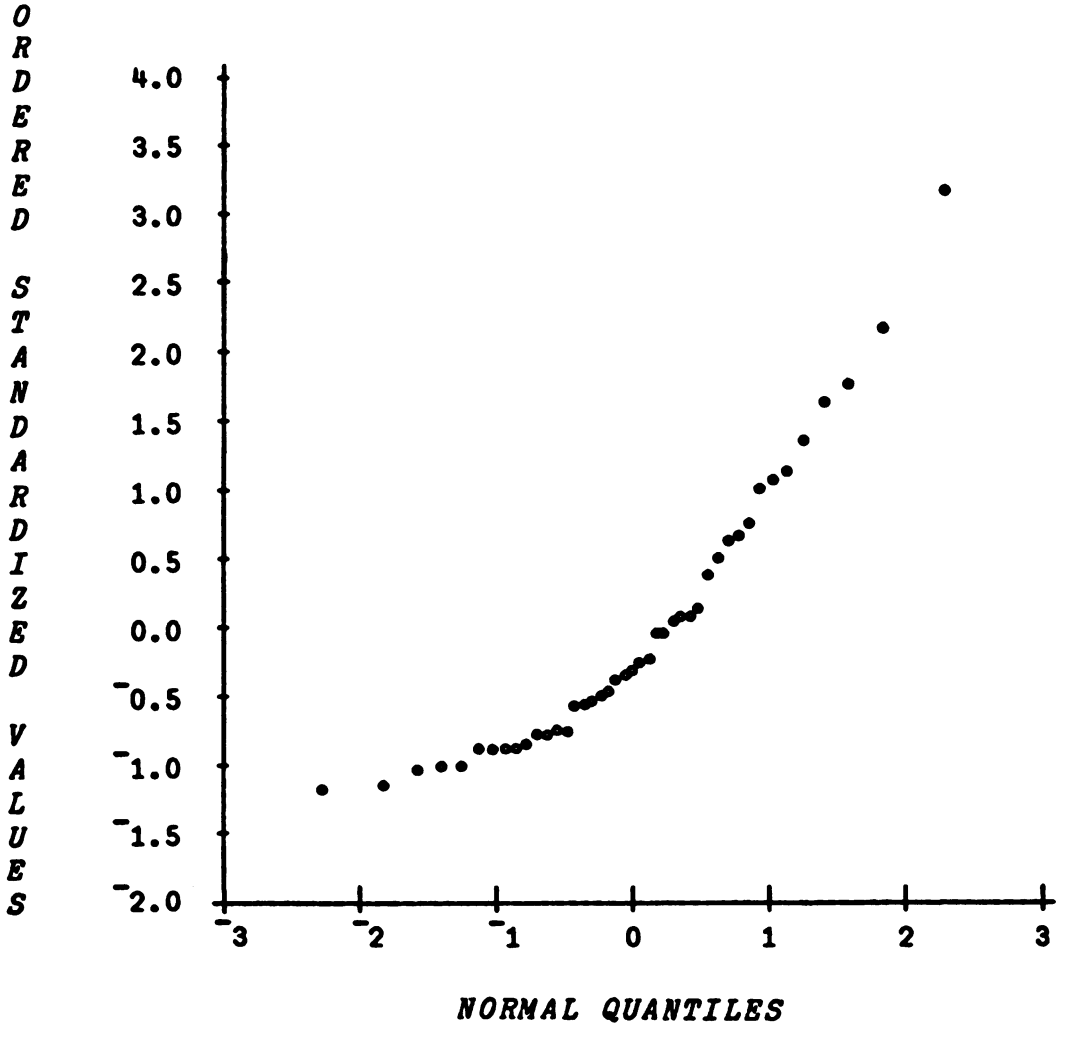


Fig. 1a. Normal probability plot for serum iron data.

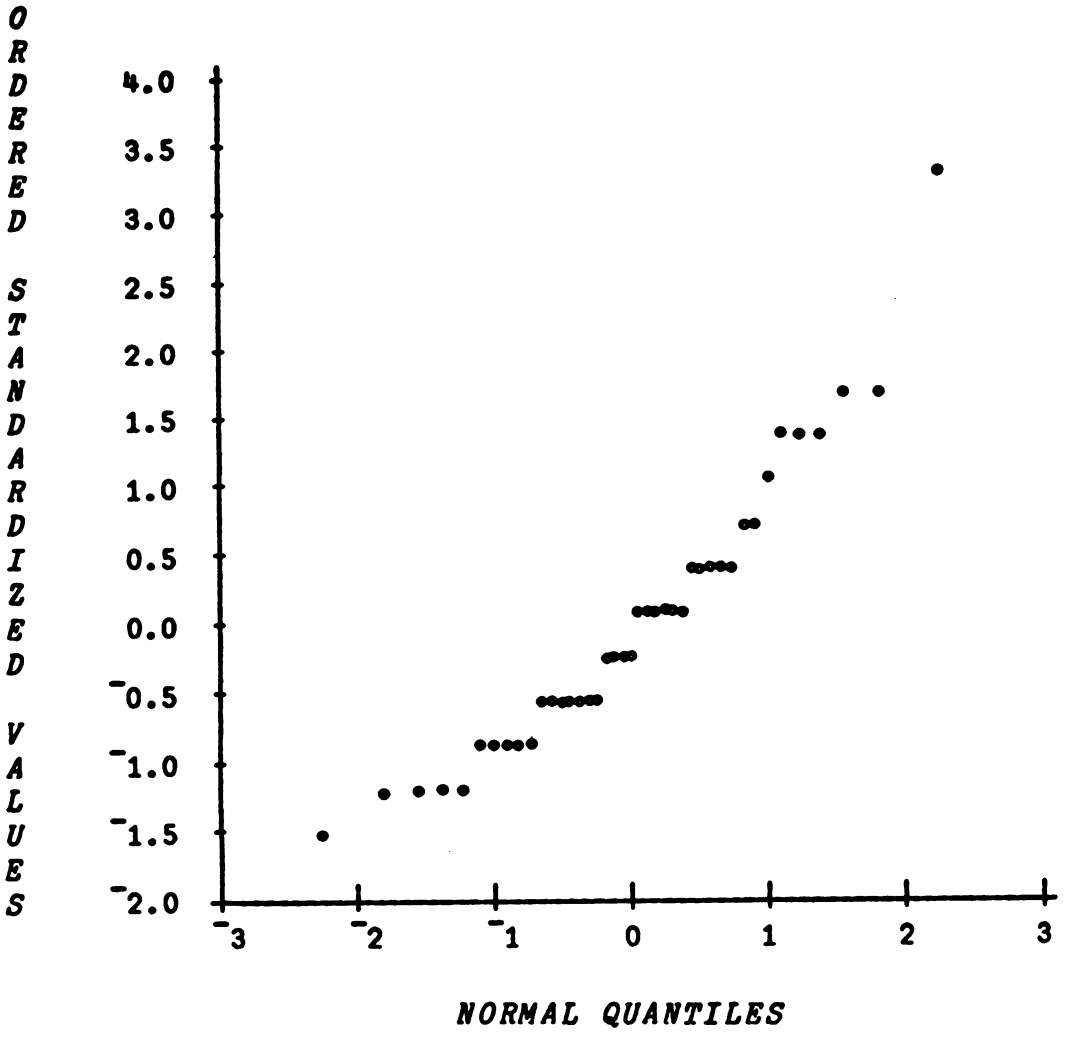


**Fig. 1b. Normal probability plot for monocyte data.**

order that the data can be compared with the 45° line. If a computer program is not available the use of specially prepared normal probability paper makes it easy to plot the data by hand.

The Kolmogorov-Smirnov test determines if the deviations between the points and the 45° line are larger than can be attributed to sampling fluctuation. For each data set the observed significance level (the probability of obtaining deviations as large as those observed by chance above) was less than 1% and this formal test confirms our conclusion that the three data sets are not Gaussian. Figures 2a,b show that the serum iron data and the monocyte data can be successfully transformed to normality. The method of determining these transformations will now be discussed.

For the serum iron data, we see from Table IIa that the value of A seems to have little effect on the value of L, hence A may be set equal to zero. We see also that for B = 1 the value of L is −174.6 and that L increases as B gets closer to zero and then starts to decrease again, implying that B = 0 is the optimum value. Thus the serum iron data were transformed to

$$y = \log_e x$$

which, when tested with the Kolmogorov-Smirnov test were concluded to have a Gaussian distribution. For the monocyte

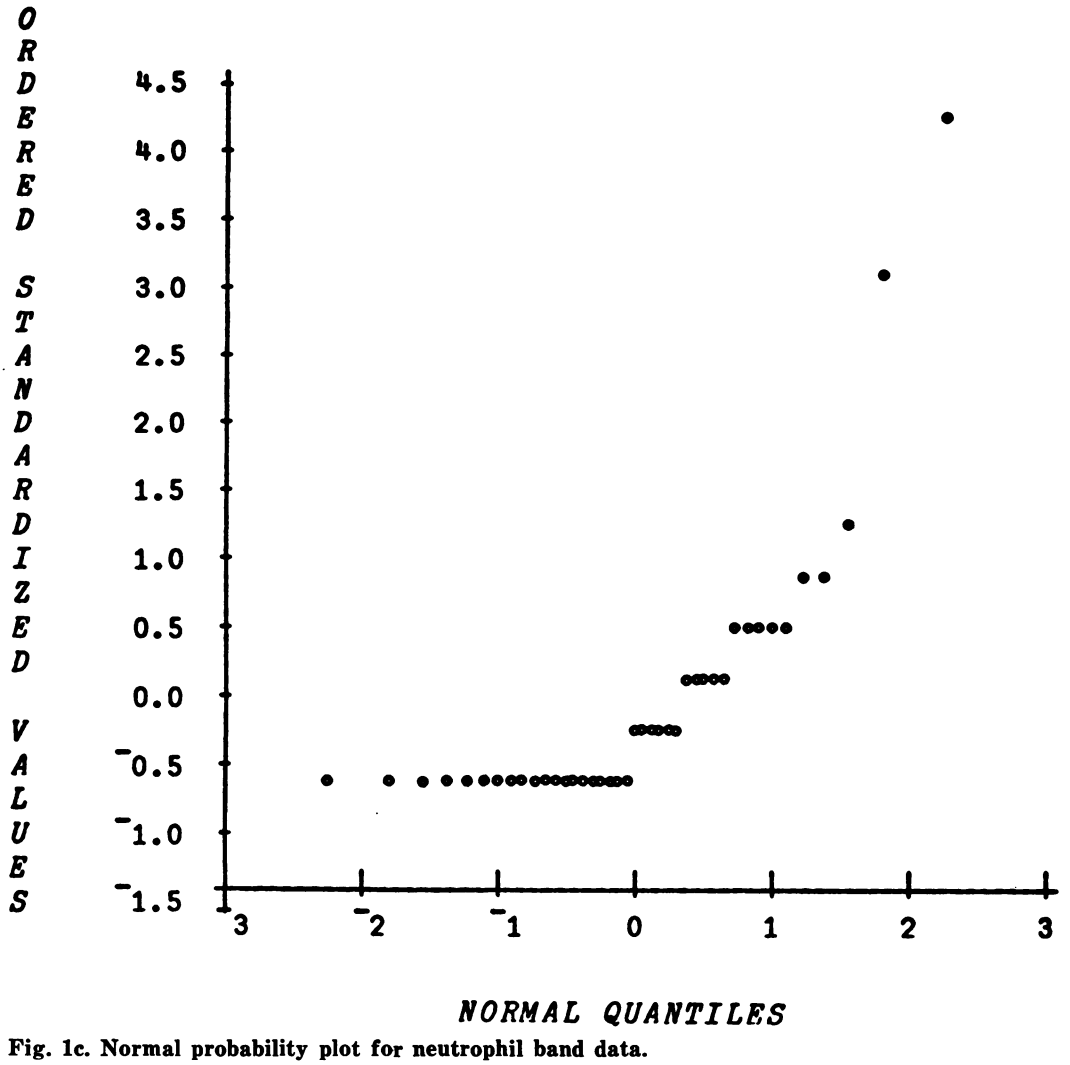


**Fig. 1c. Normal probability plot for neutrophil band data.**

data, the value of A again seemed to be unimportant and hence was set at a convenient value of 1 (A cannot be set equal to zero, because the transformations would necessitate taking the square roots of negative numbers). We see that L increases as B approaches zero. There appears to be little difference in the values of L for B = 0 (the logarithmic transformation) and for B = 0.5 (the square root transformation); in fact both transformations gave results which we concluded to be Gaussian, but for further discussion here, we used the square root transformation so that

$$y = 2\sqrt{x+1} - 2.$$

For the neutrophil band data, the largest value of L is obtained for A = 1, B = 0.5, for which the transformation is

$$y = 2 - \frac{2}{\sqrt{x+1}}.$$

However, upon making this transformation, and applying the test for Gaussianness, the data still did not satisfy the conditions for being Gaussian. This conclusion was to be anticipated for the neutrophil band data since there is a mode at x = 0 where half the observations were found. Clearly, no monotone transformation can symmetrize this data. Therefore no transformation was appropriate for them and further analyses

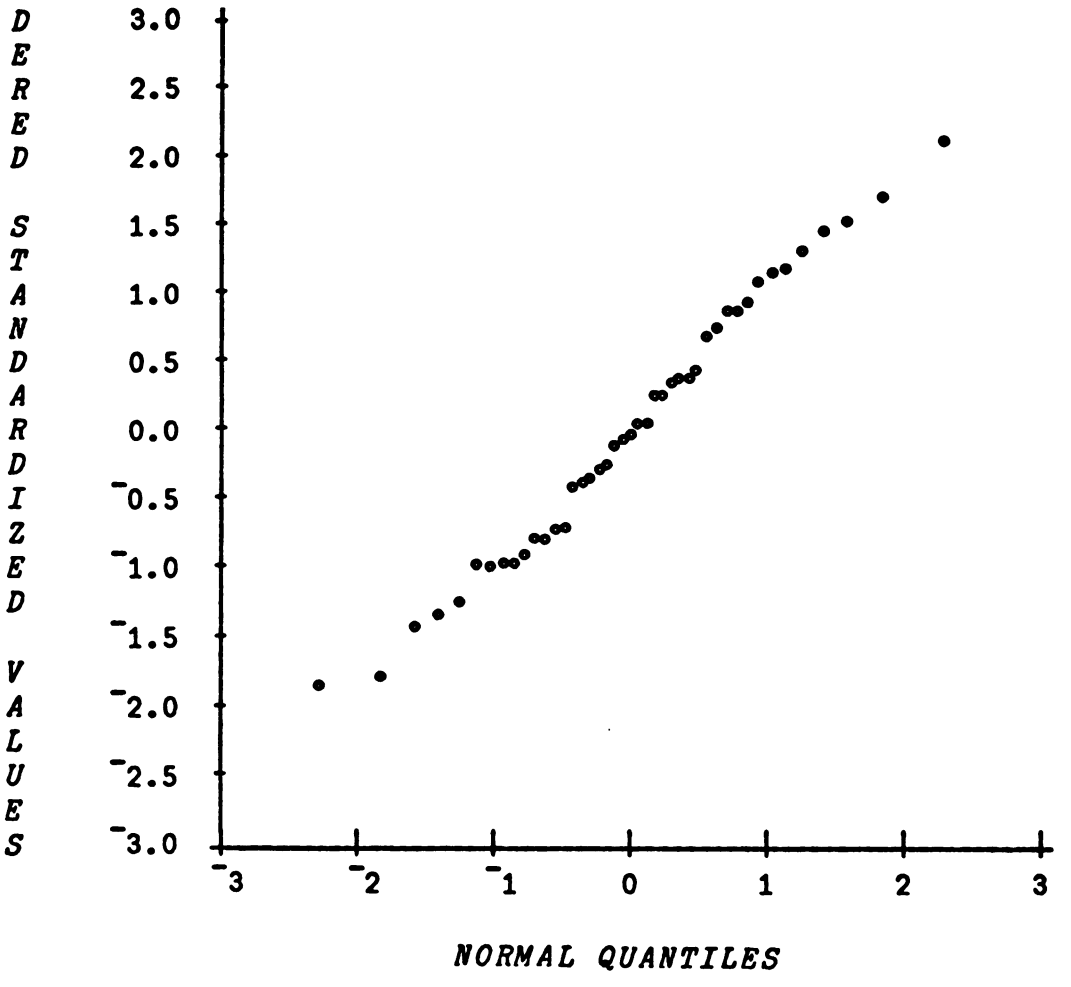


**Fig. 2a. Normal probability plot for log serum iron data.**

could be done using nonparametric techniques.

In conclusion, of the three sets of data, the serum iron data were rendered Gaussian by a logarithmic transformation, the monocyte data were rendered Gaussian by a square root-type transformation and the neutrophil data could not be rendered Gaussian.

## DISCUSSION AND OTHER APPLICATIONS

It has been assumed that for some (A,B) value, the transformed variable is normally distributed. The primary function of A will usually be to ensure that the numbers x+A are positive. Any values of (A,B) close to the maximum likelihood estimates will be reasonable ones. In fact, for fixed A, values of B whose log likelihood differs from the maximized log likelihood by no more than 0.65 (½ times the logarithm to the base e, of the 95th percentile of a chi-square variable on 1 degree of freedom) make up a 95% confidence interval for B. For example, for the monocyte data if it is decided to take A = 1 then the maximum likelihood estimate of B is 0.5 and any values of B in the range −0.2 to 0.7 are reasonably plausible values. A value for B which is biologically meaningful is to be preferred over the exact maximum likelihood estimate. For example, if the variable x is survival time (number of animals or cells in a certain

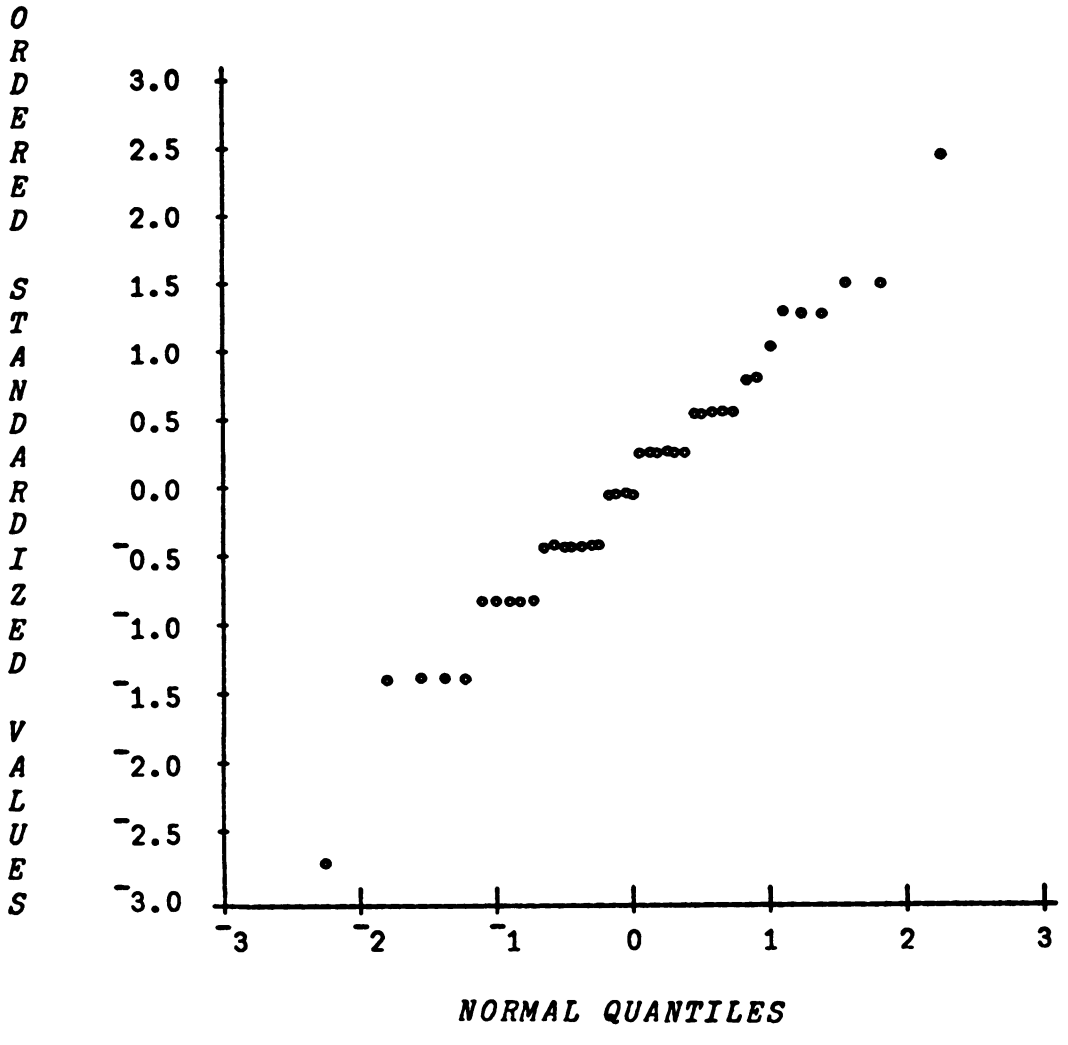


**Fig. 2b. Normal probability plot for $\sqrt{\text{monocyte}}$ data.**

region) then $\frac{1}{x}$ is interpretable as the "rate of dying" ("density") which, if B = −1 is a plausible value, would suggest that the reciprocal transformation should be used.

It is often important to draw conclusions in terms of the original metric rather than in the transformed one. Since the ordering of the observations is preserved by the family of transformations, one merely takes the inverse transformation of percentile estimates (and their confidence limits) computed in the transformed metric to obtain those in terms of the original one. For the serum iron data the median of the transformed data is 4.36, the 22nd smallest observation, and hence the estimated median in the original metric is $e^{4.36} = 78.3$ (mg/dL). If one wishes to estimate the 95% tolerance limits of the log serum iron measurements one uses $\bar{y} \pm ks$ where k = 2.3465 (7).

Taking antilogs yield $e^{\bar{y}+ks} = (20.7, 319.3)$. If one ignored the nonGaussian nature of the original measurements and worked directly with them one would have obtained (−39.6, 232.1) where the lower limit is clearly inadmissable. This example illustrates that using a properly chosen transformation can preclude nonsensical results. See Lumsden and Mullen (7) for details on tolerance and percentile estimation for veterinary data.

If one decided to use $y = 2\sqrt{2x+1} - 2$ for the monocyte data, one could equally well use the somewhat simpler transformation $y = \sqrt{x+1}$. This procedure of changing the scale 2, and location –2, is valid unless one is carrying out a regression analysis without a constant term (10).

The problem of estimating means of the original data is somewhat more involved since the inverse transformation of the mean in the transformed metric is a biased estimator of the mean of the original measurements. Details as to how to correct for this bias are given in Land (6).

It is becoming more common to analyze veterinary data arising from complex experiments using regression and analysis of variance techniques. One way of proceeding is to work with the "standardized variable"

$$z = \frac{(x + A)^B - 1}{BG}$$

where $G = e^{Avge(x+A)}$ is the geometric mean of the (x+A)'s (1). One then determines the value of (A,B) which leads to the smallest residual or error sum of squares.

The reader is encouraged to read Draper and Hunter (3) for more examples which are analyzed using a slightly different approach. Hinkley (5) provides a quick method of choosing a transformation that is suitable for hand calculation and is not sensitive to outlying observations.

## REFERENCES


1. **BOX, G.E.P. and D.R. COX.** An analysis of transformations. J.R. Statist. Soc. B 26: 211-252. 1964.
2. **COX, D.R. and D. L. HINKLEY.** Theoretical Statistics. pp. 283-310. London: Chapman and Hall. 1974.
3. **DRAPER, N.R. and W.G. HUNTER.** Transformation: Some examples revisited. Technometrics 11: 23-40. 1969.
4. **FISHER, R.A.** On the mathematical foundations of theoretical statistics. Phil. Trans. A 222: 309-368. 1921.
5. **HINKLEY, D.** On quick choice of power transformation. Appl. Statist. 26: 67-69. 1977.
6. **LAND, C.E.** Confidence interval estimation after data transformations to normality. J. Am. Statist. Ass. 69: 795-802. 1974.
7. **LUMSDEN, J.H. and K. MULLEN.** On establishing reference values. Can. J. comp. Med. 42: 293-301. 1978.
8. **LUMSDEN, J.H. K. MULLEN and B.J. McSHERRY.** Canine hematology and biochemistry reference values. Can. J. comp. Med. 43: 125-131. 1979.
9. **LUMSDEN, J.H., K. MULLEN and R. ROWE.** Hematology and biochemistry reference values for female Holstein cattle. Can. J. comp. Med. 44: 24-31. 1980.
10. **SCHLESSELMAN, J.** Power families: a note on the Box and Cox transformation. J.R. Statist. Soc. B 33: 307-311. 1971.
11. **SNEDECOR, G.W. and W.G. COCHRAN.** Statistical Methods, 6th Edition. pp. 325-330, 493-502. Ames: Iowa State University Press. 1967.
12. **STEEL, R.G.D. and J.H. TORRIE.** Principles and Procedures of Statistics. Toronto: McGraw-Hill. 1960.
13. **THONI, H.** Transformations of variables used in the analysis of experimental and observations date. A review. Tech. Report 7. Statistical Laboratory, Iowa State University, Ames, Iowa. 1967.