DRUG DEVELOPMENT AND INDUSTRIAL PHARMACY, 18(19), 2105-2115 (1992)

COMMENTS ON THE INTERPRETATION OF DATA FROM LOG NORMAL DISTRIBUTIONS AND DEFINITION OF GEOMETRIC STANDARD DEVIATION

Rose E Gaines Das
National Institute for Biological Standards and Control
Blanche Lane, South Mimms, Potters Bar
Hertfordshire, EN6 3QG

ABSTRACT

Estimates of relative potency (commonly obtained as antilogs of log relative potencies) and the distribution of many pharmaceutical measurements are positively skewed and appear to follow a log normal distribution. The geometric mean arises naturally as a measure of location in these circumstances. There is not, however, a measure of dispersion which corresponds to the geometric mean and conflicting definitions have been used for the term geometric standard deviation. Examples of data for which the use of geometric means may be appropriate are given, together with comments on inferences which may be drawn regarding the population from which the data arise. It is suggested that the term geometric standard deviation leads to confusion and inappropriate inferences, and that it should therefore be avoided.

INTRODUCTION

Data are frequently observed in pharmaceutical work which are, by their nature or from the way in which they are obtained, greater than 0, but which

may not have a fixed upper limit. Such data do not conform to a normal distribution, but are often positively skewed. Examples of such data include estimates of potency, morphological measurements and bioavailability estimates. These data are nevertheless widely summarized in terms of their mean and standard deviation which are then often interpreted as meaning that approximately 95% of values lie in an interval of ± 2 standard deviations about the mean, or that values lying further than 3 standard deviations from the mean are outliers. However, such summary and interpretation are based on statistical theory developed under the assumption that the observed data represent a random sample of values of a normally distributed variable. Depending on the degree of non-normality of the actual distribution, these inferences may be misleading.

Where the non-normality of these distributions is recognised, the observed values may be transformed, commonly using a logarithmic transformation, to give transformed values which approximate more closely to, and may thus be assumed to follow, a normal distribution. When a logarithmic transformation is used, the distribution of the observed variable is said to be a log normal distribution. The log normal distribution and its properties have been extensively described by, amongst others, Aitchison and Brown (1) and Johnson (2).

When a satisfactory 'normalizing' transformation such as the logarithmic transformation has been found, it is then appropriate to carry out the usual statistical processes of estimation and testing using the transformed values. However, it is often desirable to express the results of these statistical processes in terms of the scale of the observed (untransformed) variable. In the case of the log normal distribution, application of the reverse transformation to the mean gives the geometric mean. There is unfortunately no simple analogous statistic related to the reverse transformation of the variance (or standard deviation).

DESCRIPTIVE MEASURES OF DATA

The 'location' of a set of data is customarily described in terms of its arithmetic mean, although other measures such as the median may also be used. Let $y_1, y_2, \ldots y_n$ denote a random sample of size n from a log normal distribution such that $x_i = \log(y_i)$, $i = 1,\ldots,n$ are a random sample from a normal distribution with mean ξ and variance σ^2. (Log denotes logarithm to the base e; results are similar for logarithms to base 10 with appropriate inclusion of the constant log 10.) Estimates for ξ and σ^2 are the usual sample mean, $\bar{x} = \Sigma x_i/n$, and variance , $s^2 = \Sigma(x_i - \bar{x})^2/(n-1)$, respectively (where Σ denotes the sum over $i = 1,2,\ldots,n$) and the variance of $\bar{x}$ is σ^2/n.

Confidence limits for the mean, ξ, of the normally distributed x are readily obtained using 'normal statistical theory', namely selecting a and b so that

$$\Pr\left(a < \frac{\bar{x} - \xi}{s\sqrt{n}} < b\right) = 95\%.$$

This can be readily done since the distributions of $\bar{x}$ and s^2 are known, and the ratio given above can be shown to follow a Student's t distribution. Moreover, the probability statement is true if antilogs are taken of each of the terms in the inequality, thus giving a confidence interval for the geometric mean of the y. The relationships (noted in the Appendix) between the mean of the normally distributed x and that of the log normally distributed y taken together with the information given in Tables A1 and A2 show that for small values of σ (in particular, for $\sigma < 0.1$) the geometric mean is a good approximation to the arithmetic mean of y. In this case, the log normal distribution also approximates very closely to a normal distribution. However as σ increases, the approximation becomes increasingly poor.

The standard deviation is customarily taken to be the measure which is descriptive of the dispersion of a set of data, often with the unstated assumption

(as indicated by subsequent use of this measure) that the values are a sample from a normal distribution. This measure is also computed if the data have been logarithmically transformed. However, in the case of the log normal distribution the 'reverse transformation' i.e. the anti log, of the standard deviation of the normally distributed x, unlike the antilog of the mean, cannot be interpreted analogously to σ as an *additive* factor when applied to the log normally distributed y. It can, however, be interpreted as a *multiplicative* factor. This multiplicative factor has been called the geometric standard deviation (5), and use of this factor to obtain confidence intervals is equivalent to obtaining the confidence interval for ξ and then using the reverse transformation. This definition appears to have been misunderstood by Bohidar (6) who defines the 'geometric standard error' as the square root of the variance, approximated by a Taylor series expansion, of the geometric mean and then uses it to obtain 'confidence intervals' for the geometric mean. He also defines the geometric standard deviation as the square root of the sample size times this geometric standard error. Although these terms are used with a Student's t value in this procedure (6), no evidence is given as to the conformity of these values to the implied distributions.

The intention underlying taking the antilog of $\bar{x}$ and then attempting to find confidence intervals by this method (6) is not clear. Does the interest lie in determining a confidence interval of some other sort than that with equal areas at either end of the distribution, since such a confidence interval is most readily obtained as already described using the symmetry of normal and Student's t distributions, or does the interest lie in estimating the arithmetic mean rather than the geometric mean of the log normally distributed y variable and then obtaining a confidence interval of some sort for this estimate? This latter problem has been described elsewhere (2,4).

APPLICATION OF VARIOUS DESCRIPTIVE MEASURES TO DATA

Total output of gastric acid in rats, expressed as microequivalents of acid/ml in response to injection of salmon calcitonin was studied as described elsewhere (7). The values obtained for two of the control groups in this study, denoted A and B, are shown in Figure 1a and c and the log transformed values are shown in Figure 1b and d. Results of the Shapiro-Wilk test for normality, together with the means and 95% confidence intervals obtained by applying the various procedures described above are shown in Table 1. Thus for control group A, the mean of the log transformed values was 3.738 with standard error 0.184 ($=0.610/\sqrt{11}$). The 95% confidence interval for the mean of the log transformed values (obtained as mean $\pm$ the standard error multiplied by the value of Student's t with 10 degrees of freedom and area 0.025 to its right) is 3.328-4.148, and hence, taking antilogs, the 95% confidence interval for the geometric mean, 42.0, is 27.9-63.3 micro-equivalents of acid/ml.

The data for Batch-R as given by Bohidar (6) have been similarly analyzed. It is not clear, on examination of this particular set of data, why a log transformation was used since it has the effect of increasing the negative skewness of the distribution (Figure 1, e and f). This is reflected in the increased departure of the log transformed data from a normal distribution compared with the untransformed data (Table 1, PNL < PN).

The interval denoted GMCB (Table 1) is not, assuming that the log transformed values are normally distributed, a 95% confidence interval but somewhat less than 95%, and is also virtually a 'one sided interval' since, if the theoretical mean ξ actually had value $\bar{x}$ then the probability of a geometric mean less than this lower limit is less than 1% and the probability of a geometric mean greater than the upper limit is virtually 5%. For control group B the deviation from 95% is more notable with the probability of a geometric

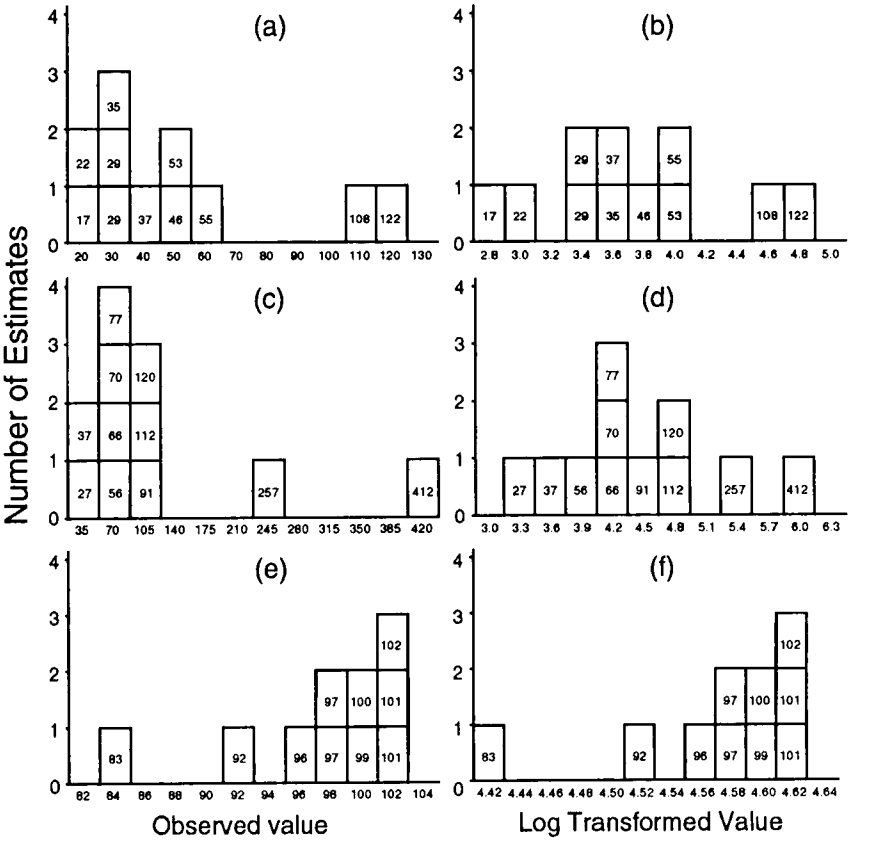

Figure 1

Observed (a,c,e) and log transformed (b,d,f) values for the gastric acid data sets A (a,b) and B (c,d) and for the data from 'Batch-R' (reference 6) (e,f). Observed values, are given inside the square.

mean greater than the upper limit being larger than 5%. Comparison of these intervals with intervals having 'equal tail areas' (Table 1) shows that for data for which conformity to a normal distribution is improved by a log transformation these 'GMCB' intervals may substantially underestimate the upper limit.

TABLE 1.

Arithmetic and geometric means and confidence intervals for various sets of data. Statistics given are n = sample size, PN = probability of a more extreme value of the Shapiro-Wilk statistic under the assumption that the observed data follow a normal distribution, PNL = probability as PN except for log transformed observed data, AM = arithmetic mean of observed data, GM = geometric mean of observed data, AMC = '95% confidence interval' about the AM calculated assuming the observed data follow a normal distribution, GMC = 95% confidence interval about the geometric mean of the observed data calculated as the antilog of the 95% confidence interval about the arithmetic mean of the log transformed data, GMCB = '95% confidence interval' about the geometric mean of the observed data calculated using the technique given by Bohidar (6).

Statistic	Gastric Acid Data		Batch R
	(A)	(B)	
n	11	11	10
PN	0.014	0.002	0.019
PNL	0.658	0.712	0.011
AM	50.3	120.5	96.80
AMC	27.2-73.3	43.3-197.7	92.73-100.87
GM	42.0	88.4	96.64
GMC	27.9-63.3	52.2-150.0	92.47-101.00
GMCB	24.8-59.2	41.7-135.1	92.38-100.90

CONCLUSIONS

It is suggested that before a logarithmic transformation is applied to a set of observed values y_i, i = 1,....,n, the distribution of these values should be assessed, even if only informally or graphically, for its conformity to a normal distribution. Where the distribution of the values is positively skewed or where there is evidence that the values may be divisible into subpopulations with

different means and standard deviations but with the standard deviation of the subpopulation being proportional to the mean of the subpopulation, then a logarithmic transformation may give values $x_i = \log y_i$ which conform more closely than the observed values to a normal distribution or have more homogeneous variances. In this case it is suggested that hypothesis testing concerning, or construction of confidence intervals for, the geometric means may most appropriately be carried out by working with the presumed normally distributed x values and then translating conclusions about ξ, the mean of the x values, into conclusions about the geometric mean if this is the number of interest (see e.g. (8)). This avoids the need to use a 'geometric standard deviation' unless there is some feature of the dispersion of the values which is of particular interest. In that case, a direct approach to the problem of interest with a clear definition of the terms used is suggested.

REFERENCES

(1) J. Aitchison and J.A.C Brown, The Lognormal Distribution, Cambridge University Press, 1957.

(2) N.L. Johnson and S. Kotz, Distributions in Statistics: Continuous Univariate Distributions Vol. 1, Houghton-Mifflin Co., New York, 1970.

(3) D.J. Finney, On the distribution of a variate whose logarithm is normally distributed, Journal of the Royal Statistical Society, Series B, Vol. 7, 155-161 (1941).

(4) C.E. Land, Standard confidence limits for linear functions of the normal mean and variance, Journal of the American Statistical Association, Vol. 68, 960-963 (1973).

(5) T.B.L. Kirkwood, Geometric means and measures of dispersion, Biometrics Vol. 35, 908-909 (1979).

(6) N.R. Bohidar, Determination of geometric standard deviation for dissolution, Drug Development and Industrial Pharmacy, Vol. 17, 1381-1387 (1991).

(7) F. Guidobono, C. Netti, A. Pecile, J.M. Zanelli and R.E. Gaines Das, Specific inhibition of basal gastric acid secretion by salmon calcitonin in rats does not necessarily involve a central pathway, Pharmacological Research Vol. 22, 287-295 (1990).

(8) W.G.S. Hines, Geometric mean, in Encyclopedia of Statistical Sciences, Vol. 5, S. Kotz and N.L. Johnson (editors) John Wiley and Sons, New York, 1985, p. 397-400.

APPENDIX

RELATIONSHIP OF MEANS AND VARIANCES OF NORMAL AND LOG NORMAL VARIABLES.

Let $y_1, y_2, \ldots y_n$ denote a random sample of size n from a log normal distribution such that $x_i = \log(y_i)$, $i = 1,\ldots,n$ are a random sample from a normal distribution with mean ξ and variance σ^2. The mean and variance of the log normal variables, y, cannot be obtained by taking antilogs of ξ and σ^2. Specifically the mean of y, μ, and the variance of y, ν^2, are given by

$$\mu = e^{\xi + \frac{1}{2}\sigma^2} \quad , \quad \nu^2 = e^{2\xi + \sigma^2}(e^{\sigma^2} - 1) .$$

Finney (3) has shown that efficient estimates of μ and ν^2 are given by

$$m = e^{\bar{x}} g(\tfrac{1}{2}s^2) \quad , \quad v = e^{2\bar{x}}\{g(2s^2) - g((n-2)s^2/(n-1))\}$$

where g(t) denotes the infinite series

$$g(t) = 1 + \frac{n-1}{n}t + \sum_{j=2}^{\infty} \frac{(n-1)^{2j-1}}{n^j(n+1)(n+3)\ldots(n+2j-1)} \frac{t^j}{j!} = 1 + \sum_{i=1}^{\infty} c_i t^i .$$

Table A1 shows the coefficients, c_i, of t^i for i = 1 to 6 for selected values of n, and Table A2 shows the coefficients by which the geometric mean is multiplied to give the mean μ obtained by evaluating these first terms of the series for selected values of σ (the standard deviation of the normally distributed x_i). Finney also gives an approximation to the variance of m as

$$\text{Var}(m) \simeq e^{2\xi + \sigma^2}\{\sigma^2 + \frac{\sigma^4}{2} + \frac{1}{2n}(\sigma^6 + \frac{\sigma^8}{4})\}/n.$$

TABLE A1.

Coefficients C_i of t to the power of i in the function g(t) for selected values of N.

N	C_1	C_2	C_3	C_4	C_5	C_6
3	0.6667	0.1111	0.0082	0.00034	0.000009	0.0000002
4	0.7500	0.1688	0.0181	0.00113	0.000046	0.0000013
5	0.8000	0.2133	0.0284	0.00228	0.000121	0.0000046
6	0.8333	0.2480	0.0383	0.00362	0.000232	0.0000108
7	0.8571	0.2755	0.0472	0.00506	0.000372	0.0000199
8	0.8750	0.2977	0.0553	0.00651	0.000532	0.0000319
9	0.8889	0.3160	0.0624	0.00793	0.000705	0.0000464
10	0.9000	0.3314	0.0688	0.00929	0.000885	0.0000629
15	0.9333	0.3811	0.0922	0.01506	0.001789	0.0001624
20	0.9500	0.4083	0.1068	0.01928	0.002578	0.0002674
25	0.9600	0.4254	0.1167	0.02240	0.003226	0.0003643
30	0.9667	0.4371	0.1238	0.02478	0.003755	0.0004499
35	0.9714	0.4456	0.1291	0.02665	0.004192	0.0005244
40	0.9750	0.4521	0.1333	0.02815	0.004556	0.0005892
45	0.9778	0.4572	0.1366	0.02939	0.004862	0.0006457
50	0.9800	0.4614	0.1393	0.03041	0.005124	0.0006951
60	0.9833	0.4676	0.1435	0.03203	0.005547	0.0007774
70	0.9857	0.4721	0.1466	0.03324	0.005873	0.0008427
80	0.9875	0.4755	0.1490	0.03419	0.006131	0.0008956
90	0.9889	0.4782	0.1509	0.03494	0.006340	0.0009394
100	0.9900	0.4803	0.1524	0.03555	0.006513	0.0009761

TABLE A2

Values of the function $g(\frac{1}{2}\sigma^2)$, approximated by the first seven terms, for selected N and selected values of the standard deviation, σ.

N	0.01	0.025	0.05	0.075	0.1	0.25	0.5	0.75	1	1.5	2	3
3	1.00003	1.00021	1.00083	1.00188	1.00334	1.0209	1.0851	1.196	1.362	1.903	2.849	7.159
4	1.00004	1.00023	1.00094	1.00211	1.00375	1.0236	1.0964	1.225	1.420	2.085	3.339	10.000
5	1.00004	1.00025	1.00100	1.00225	1.00401	1.0252	1.1034	1.243	1.457	2.214	3.721	12.707
6	1.00004	1.00026	1.00104	1.00235	1.00417	1.0263	1.1081	1.255	1.484	2.312	4.031	15.264
7	1.00004	1.00027	1.00107	1.00241	1.00429	1.0271	1.1115	1.264	1.504	2.389	4.288	17.667
8	1.00004	1.00027	1.00109	1.00246	1.00438	1.0276	1.1141	1.271	1.519	2.451	4.506	19.918
9	1.00004	1.00028	1.00111	1.00250	1.00445	1.0281	1.1162	1.276	1.532	2.503	4.694	22.025
10	1.00005	1.00028	1.00113	1.00253	1.00451	1.0285	1.1178	1.281	1.542	2.546	4.857	23.997
15	1.00005	1.00029	1.00117	1.00263	1.00468	1.0295	1.1228	1.295	1.574	2.691	5.438	32.147
20	1.00005	1.00030	1.00119	1.00268	1.00476	1.0301	1.1253	1.302	1.592	2.774	5.796	38.156
25	1.00005	1.00030	1.00120	1.00270	1.00481	1.0304	1.1269	1.306	1.602	2.827	6.040	42.728
30	1.00005	1.00030	1.00121	1.00272	1.00484	1.0306	1.1279	1.309	1.610	2.864	6.217	46.307
35	1.00005	1.00030	1.00121	1.00274	1.00487	1.0308	1.1287	1.312	1.615	2.892	6.352	49.179
40	1.00005	1.00030	1.00122	1.00275	1.00489	1.0309	1.1292	1.313	1.619	2.913	6.459	51.531
45	1.00005	1.00031	1.00122	1.00275	1.00490	1.0310	1.1296	1.314	1.622	2.930	6.544	53.492
50	1.00005	1.00031	1.00123	1.00276	1.00491	1.0311	1.1300	1.315	1.625	2.944	6.615	55.150
60	1.00005	1.00031	1.00123	1.00277	1.00493	1.0312	1.1305	1.317	1.629	2.965	6.725	57.800
70	1.00005	1.00031	1.00123	1.00278	1.00494	1.0313	1.1309	1.318	1.631	2.981	6.807	59.824
80	1.00005	1.00031	1.00124	1.00278	1.00495	1.0313	1.1312	1.319	1.634	2.993	6.870	61.418
90	1.00005	1.00031	1.00124	1.00279	1.00496	1.0314	1.1314	1.320	1.635	3.002	6.919	62.707
100	1.00005	1.00031	1.00124	1.00279	1.00496	1.0314	1.1316	1.320	1.637	3.009	6.960	63.769