THE FITTING OF A TRUNCATED LOG-NORMAL CURVE TO DAILY RAINFALL DATA*

By S. C. Das†

Introduction

In a previous paper (Das 1955) the author discussed a problem of curve fitting which arose in testing the hypothesis proposed by Bowen (1953) concerning daily rainfall data.

In this investigation, the Sydney daily rainfall over the period of 94 years from 1859 to 1952 was examined. A type III probability distribution of the form

$$f(x)=\frac{\mu^{\varkappa}}{\Gamma(\varkappa)}e^{-\mu x}x^{\varkappa-1}$$

provided a good fit to these data.

Because, in the case of rainfall data, we measure the rainfall x to the nearest rounded-off unit on some scale, we are likely to have some zero values of x; in fact we have a large number of zero values in the case of daily rainfall. This made it impossible to use the maximum likelihood equations for estimating the parameters, since these equations involve the sum of the logarithms of the observations. We therefore fitted truncated type III curves to the observations, the truncation being of the following two types. In one case we chose a small interval $(0, \delta)$ and truncated the distribution at δ, ignoring the actual values of x less than δ, but using the fact that we knew their total number. In the other case we fitted a truncated type III curve to the observations which are greater than δ, and ignored all observations which are less than δ. In both cases a very good fit was obtained as judged by the χ^2 test, and there was no significant difference in the expected numbers in the truncated part, so that there was no evidence of singularity at the origin of the distribution.

It is more usual, however, in meteorological practice to fit a log-normal curve of the type

$$f(x)=\frac{1}{(2\pi)^{\frac{1}{2}}\sigma}\cdot\frac{1}{(x-a)}\exp\left[-\frac{1}{2\sigma^2}\{\ln(x-a)-\mu\}^2\right]$$

to rainfall data.

The first attempt to fit a normal distribution to the logarithms of the values of a meteorological element appears to have been made by Blackhouse (1891). He compared the frequencies of annual rainfall amounts and of their logarithms with a normal distribution. He took $a=0$ and found the fit not very convincing, perhaps because only 30 observations (1860–89) were available. More recently

* Manuscript received October 31, 1955.

† Australian National University, Canberra, A.C.T.

Brooks and Carruthers (1953, p. 102) suggest that a log-normal curve will give a good fit for any distribution which is positively skew and sufficiently leptokurtic. They fitted a log-normal curve to the rainfall totals at Camden Square, London, for sets of four consecutive months, between 1870 and 1943, and found the fit to be good. In their data the frequency rises quickly to a maximum and then gradually drops down to low values.

The distribution of rainfall data for Sydney, though slightly leptokurtic, is J-shaped and therefore differs from their data. It seems to be quite commonly believed by meteorologists, however, that the log-normal curve is generally appropriate for rainfall data. In this paper we fit a log-normal curve to the Sydney rainfall data in order to see how good or bad the fit is when compared with that obtained by the use of the type III curve. To apply the maximum likelihood method it is again necessary to truncate the distribution to avoid the zero values. Maximum likelihood equations were complicated for the type III curve, but in this case they are much simpler.

Fiting of a Truncated Log-normal Curve

The log-normal probability density function is the form

$$f(x)=\frac{1}{(2\pi)^{\frac{1}{2}}\sigma(x-a)}\exp\left[-\frac{1}{2\sigma^2}\{\ln(x-a)-\mu\}^2\right] \quad \text{for } x>a\,;$$

in this case we know the origin of our distribution and so take $a=0$.

We now choose a small interval $(0, \alpha)$ and truncate the distribution at α, ignoring the actual values of x less than α, but using the fact that their total number is known. Thus, if n be the number of observations falling in $(0, \alpha)$, the rest $(N-n=m)$ of the observations will all be greater than α.

The likelihood function $\varphi(x_1, x_2, \ldots, x_m)$ in this case is given by

$$\binom{N}{n}\left[\frac{1}{(2\pi)^{\frac{1}{2}}\sigma}\int_0^{\alpha}\frac{1}{x}\exp\left\{-\frac{1}{2\sigma^2}(\ln x-\mu)\right\}^2 dx\right]^n \cdot \frac{1}{\sigma^m(2\pi)^{m/2}}$$

$$\times\frac{1}{\prod_{i=1}^{m}x_i}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(\ln x_i-\mu)^2\right\},$$

where $x_1, x_2, \ldots, x_m \geqslant \alpha$.

Taking logarithms, we obtain

$$L=\ln\varphi$$
$$=\ln\binom{N}{n}+nG(\mu, \sigma)-m\ln\sigma-\tfrac{1}{2}m\ln 2\pi-\sum_{i=1}^{m}\ln x_i-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(\ln x_i-\mu)^2,$$

where

$$G(\mu, \sigma)=\ln\left[\frac{1}{(2\pi)^{\frac{1}{2}}\sigma}\int_0^{\alpha}\frac{1}{x}\exp\left\{-\frac{1}{2\sigma^2}(\ln x-\mu)^2\right\}dx\right].$$

The maximum likelihood equations are given by

$$\frac{\partial L}{\partial \mu}=n\frac{\partial G}{\partial \mu}+\frac{1}{\sigma^2}\sum_{i=1}^{m}(\ln x_i-\mu)=0, \quad \ldots\ldots\ldots\ldots (1)$$

$$\frac{\partial L}{\partial \sigma}=n\frac{\partial G}{\partial \sigma}-\frac{m}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{m}(\ln x_i-\mu)^2=0, \quad \ldots\ldots\ldots (2)$$

where μ and σ are now no longer population parameters, but, for simplicity, stand for their estimates.

Now

$$\frac{\partial G}{\partial \mu}=\frac{\int_0^{\alpha}\frac{1}{x}\frac{(\ln x-\mu)}{\sigma^2}\exp\left\{-\frac{1}{2\sigma^2}(\ln x-\mu)^2\right\}dx}{\int_0^{\alpha}\frac{1}{x}\exp\left\{-\frac{1}{2\sigma^2}(\ln x-\mu)^2\right\}dx}.$$

After changing the variables in the numerator and integrating by parts, this reduces to

$$\frac{\partial G}{\partial \mu}=\frac{-\exp\left\{-\frac{1}{2\sigma^2}(\ln \alpha-\mu)^2\right\}}{\int_0^{\alpha}\frac{1}{x}\exp\left\{-\frac{1}{2\sigma^2}(\ln x-\mu)^2\right\}dx}. \quad \ldots\ldots\ldots\ldots\ldots (3)$$

Further

$$\frac{\partial G}{\partial \sigma}=-\frac{1}{\sigma}+\frac{\int_0^{\alpha}\frac{1}{x}\frac{(\ln x-\mu)^3}{\sigma^3}\exp\left\{-\frac{1}{2\sigma^2}(\ln x-\mu)^2\right\}dx}{\int_0^{\alpha}\frac{1}{x}\exp\left\{-\frac{1}{2\sigma^2}(\ln x-\mu)^2\right\}dx},$$

which may be similarly reduced to

$$\frac{\partial G}{\partial \sigma}=-\frac{1}{\sigma}\frac{(\ln \alpha-\mu)\exp\left\{-\frac{1}{2\sigma^2}(\ln \alpha-\mu)^2\right\}}{\int_0^{\alpha}\frac{1}{x}\exp\left\{-\frac{1}{2\sigma^2}(\ln x-\mu)^2\right\}dx}. \quad \ldots\ldots\ldots\ldots (4)$$

From (3) and (4) we see that

$$\frac{\partial G}{\partial \sigma}=\frac{1}{\sigma}(\ln \alpha-\mu)\frac{\partial G}{\partial \mu}. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Using (5) we rewrite the likelihood equations (1) and (2) as

$$n\frac{\partial G}{\partial \mu}+\frac{1}{\sigma^2}\sum_{i=1}^{m}(\ln x_i-\mu)=0, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

$$\frac{n}{\sigma}(\ln \alpha-\mu)\frac{\partial G}{\partial \mu}-\frac{m}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{m}(\ln x_i-\mu)^2=0. \quad \ldots\ldots\ldots\ldots (7)$$

Now multiplying (**6**) by $\frac{1}{\sigma}(\ln \alpha - \mu)$, and subtracting (**7**) from the result we find

$$\frac{1}{\sigma^3}(\log \alpha - \mu) \sum_{i=1}^{m} (\log x_i - \mu) + \frac{m}{\sigma} - \frac{1}{\sigma^3} \sum_{i=1}^{m} (\log x_i - \mu)^2 = 0,$$

which can be written as

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^{m} (\ln x_i)^2 - \frac{1}{m}(\mu + \ln \alpha) \sum_{i=1}^{m} \ln x_i + \mu \ln \alpha. \quad \ldots\ldots\ldots \quad (8)$$

Thus equations (**6**) and (**8**) can be taken as the likelihood equations for estimating μ and σ.

To make a comparison with our fitting in the previous paper we take $\alpha = 5 \cdot 5$. From the data we obtain

$$m = 437, \qquad \sum_{i=1}^{m} \ln x_i = 1373 \cdot 11748, \qquad \sum_{i=1}^{m} (\ln x_i)^2 = 4717 \cdot 25359.$$

Substituting these values in (**6**) and (**8**), we find that (**8**) reduced to

$$\sigma^2 = 5 \cdot 43807 - (1 \cdot 43739)\mu, \quad \ldots\ldots\ldots\ldots \quad (9)$$

and (**6**) reduces to

$$n \frac{\partial G}{\partial \mu} + \frac{1}{\sigma^2}(1373 \cdot 11748 - 437\mu) = 0. \quad \ldots\ldots\ldots\ldots \quad (10)$$

Solving (**9**) and (**10**) we find $\mu = -0 \cdot 17$ and $\sigma = 2 \cdot 38$ correct to two places of decimals.

The calculation of the expected frequencies for testing the goodness of fit is shown in Table 1. Since the χ^2 test is used for testing goodness of fit, the observations are grouped into classes so that the expected frequency in any class is not less than 5. To facilitate comparison we have here grouped the observations into the same classes as we did in our previous paper for fitting a truncated type III distribution. We have used the Tables of Probability Functions, Vol. I (National Bureau of Standards 1941) to calculate the expected frequencies.

The results are shown in the table. The first column gives the class interval, the column headed f_0 gives the corresponding observed class frequencies, and the column headed $f_E^{(1)}$ gives the expected frequencies based on log-normal distribution. Thus for 16 degrees of freedom the total χ^2 is found to $44 \cdot 0$, which shows that the fit is a very poor one. On the other hand the column headed $f_E^{(2)}$ gives the frequencies expected on the basis of a truncated type III curve. The value of χ^2 is here found to be $7 \cdot 8$, showing that the fit is an extremely good one.

Finney (1941) showed that, if the variable x is such that $\log x$ is normally distributed with mean ξ and variance σ^2, then the x population has the mean $\exp(\xi + \delta\sigma^2)$ and variance $\exp(2\xi + \sigma^2)(\exp \sigma^2 - 1)$. Accordingly the estimates

of the mean and standard error of the daily rainfall distribution based on log-normal curve are given by

$$\text{mean} = 14 \cdot 4 \text{ and standard deviation} = 246 \cdot 2.$$

Based on type III curve these estimates are

$$\text{mean} = 8 \cdot 1 \text{ and standard deviation} = 24 \cdot 9.$$

But calculating directly from the sample we get

$$\text{sample mean} = 8 \cdot 4 \text{ and sample standard deviation} = 27 \cdot 9.$$

Thus the mean and standard deviation calculated directly from the sample agree well with those estimated on the basis of the type III curve, but deviate considerably from those estimated on the basis of the log-normal curve. As is

TABLE 1

FREQUENCIES AS PREDICTED BY THE FITTED CURVES

Class Interval	f_0	$f_E^{(1)}$	$f_E^{(2)}$	Class Interval	f_0	$f_E^{(1)}$	$f_E^{(2)}$
0– 5	1631	1621·6	1638·5	51– 60.. ..	18	13·3	19·6
6–10	115	146·4	106·0	61– 70.. ..	13	10·2	14·7
11–15	67	70·1	62·0	71– 80.. ..	13	7·7	11·6
16–20	42	42·7	43·6	81– 90.. ..	8	6·3	8·9
21–25	27	29·1	32·2	91–100.. ..	8	5·1	7·2
26–30	26	21·1	26·0	101–125.. ..	16	9·4	12·2
31–35	19	16·1	20·7	126–150.. ..	7	6·4	7·2
36–40	14	12·7	17·2	151–225.. ..	9	10·9	9·5
41–45	12	10·5	14·3	226 or more ..	5	19·8	4·4
46–50	18	8·6	12·2				

evident from the table, this is due to the fact that the frequency towards the upper tail of the log-normal curve is very much higher than that observed. This has resulted in increasing the mean, and, more especially, the standard deviation.

On the basis of these results, it would seem that, at any rate when a certain proportion of days have zero rainfall, the better curve to use in fitting rainfall data is the type III curve, not the log-normal, as has been previously assumed by several writers in meteorology.

The author thanks Professor P. A. Moran for suggesting the problem to him.

References

BLACKHOUSE, T. W. (1891).—*Quart. J. R. Met. Soc.* **17**: 286.

BOWEN, E. G. (1953).—*Aust. J. Phys.* **6**: 490–7.

BROOKS, C. E. P., and CARRUTHERS, N. (1953).—" Handbook of Statistical Methods in Meteorology." p. 102. (H.M. Stationery Office: London.)

DAS, S. C. (1955).—*Aust. J. Phys.* **8**: 298–304.

FINNEY, D. J. (1941).—*J. R. Stat. Soc. Suppl.* **7**: 155–61.

NATIONAL BUREAU OF STANDARDS (1941).—" Tables of Probability Functions." Vol. 1.