THE FITTING OF TRUNCATED TYPE III CURVES TO DAILY RAINFALL DATA

By S. C. DAS*

[*Manuscript received November* 24, 1954]

Summary

The method of maximum likelihood has been used to fit a truncated type III(Gamma) distribution to daily rainfall data for Sydney over the period 1859–1952. An approximate test of the hypothesis that there is a singularity at the origin is suggested. This test is based on a comparison of the expected frequency in the truncated part, when the observed frequency in this part is taken into account in the fit, with the expected frequency when these observations are neglected. For Sydney the test shows that there is no evidence in the rainfall data for a singularity at the origin.

I. INTRODUCTION

This paper discusses a problem of curve fitting which arose in testing the hypothesis proposed by Bowen (1953) concerning daily rainfall data. The hypothesis advanced is that meteoritic dust is an important factor in stimulating rainfall.

An analysis of the data to test this hypothesis is presented by Hannan simultaneously with this paper (Hannan 1955). In connexion with this, it is of interest to fit a frequency distribution to the daily rainfall data in order to judge the effect of departure from normality in the distribution of daily rainfall on the test of Bowen's hypothesis.

The rainfall figures (Sydney 1859–1952) which constitute the data for the curve fitting refer to a period of 22 days from October 17 to November 7 for 94 years. The reason for choosing the period between October and November was the small seasonal variation during this period ; a period of 22 days was taken so that the total observations might exceed 2000. In fact there are 2068 observations. The shape of the distribution of rainfall suggests that a type III probability distribution of the form :

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

might provide a good fit. There are, however, a large number of zero observations. This makes it impossible to apply the ordinary maximum likelihood equations which involve the sum of logarithms of the observations. We have therefore modified the ordinary method by truncating the curve, and applying a modified maximum likelihood method which takes account of the number of observations in the truncated part. The resulting theory and the calculations

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

for the particular example are given in Section II. A very good fit is obtained as judged by the χ^2 test, and, in particular, there is a close agreement between the expected and observed numbers in the truncated part.

One might, at first sight, have suspected that such a good fit would not have been found and that the best way to fit observations of this kind would be to fit a mixed probability distribution which had a non-zero concentration at zero, and a continuous distribution for values not equal to zero. This is clearly not so in the present case, but a method is suggested for testing such an alternative. To do this we fit a truncated type III curve to the observations, ignoring the numbers in the truncated part, and compare the observed numbers in the truncated part with the numbers expected from the fitted curve. As an illustration of this method this is done in Section III.

In this connexion it is worth mentioning work done previously in this direction. Cohen (1950) uses the method of moments to develop formulas for testing μ, σ, α_3, the population mean, standard deviation, and the third standard moment respectively from a singly truncated sample when the population is distributed according to Pearson's type III function. Des Raj (1953) discusses the theory of estimation for the same population parameters as Cohen for a type III curve from both singly and doubly truncated samples. He obtains estimating equations by the method of moments and has also shown that they can be obtained by the method of maximum likelihood.

In our present problem we know where the origin of the parent distribution is whereas Cohen and Des Raj are trying to estimate it and so are estimating three parameters. Here we have two parameters and we have evolved a method suitable for our problem which is different from both of theirs.

II. Fitting of Truncated Curve taking the Number of Observations in the Truncated Part into Account

In this section we first discuss the difficulty that is to be encountered in estimating the parameters on account of zero values. Let $x_1, x_2, \ldots, x_N$ be a random sample of size N from a type III distribution given by

$$f(x)=\frac{\mu^{\varkappa}}{\Gamma(\varkappa)}\,\mathrm{e}^{-\mu x}x^{\varkappa-1}. \qquad \ldots\ldots\ldots\ldots\ldots \quad (1)$$

The likelihood of the observations is given by

$$\varphi(x_1, x_2, \ldots x_N)=\frac{\mu^{N\varkappa}}{\{\Gamma(\varkappa)\}^N}\exp\left(-\mu\sum_{i=1}^{N}x_i\right)\prod_{i=1}^{N}x_i^{\varkappa-1}. \qquad \ldots\ldots \quad (2)$$

Taking the logarithm of (2) we get

$$L=\ln\varphi\{x_1, x_2 \ldots x_N\}=N\varkappa\ln\mu-N\ln\Gamma(\varkappa)-\mu\sum_{1}^{N}x_i+(\varkappa-1)\sum_{i=1}^{N}\ln x_i.$$

$$\ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

Thus for estimating the parameters μ and $\varkappa$ we have the following maximum likelihood equations :

$$\frac{1}{N}\frac{\partial L}{\partial \mu}=\frac{\varkappa}{\mu}-\frac{1}{N}\sum_{i=1}^{N}x_i=0, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

$$\frac{1}{N}\frac{\partial L}{\partial \varkappa}=\ln \mu-\frac{\mathrm{d}\ln \Gamma(\varkappa)}{\mathrm{d}\varkappa}+\frac{1}{N}\sum_{i=1}^{N}\ln x_i=0. \quad \ldots\ldots\ldots \quad (5)$$

Now in practice the fact that we measure the x_i to the nearest rounded-off unit on some scale means that in many cases there will be zero values of x_i, and in fact in the case of the daily rainfall there are a large number of zero values. When this happens equation (5) cannot be used.

To avoid this difficulty we choose a small interval $(0, \delta)$ and truncate the distribution at δ. We ignore the actual values of x_i less than δ, but use the fact that we know their total number. Thus if n be the number of observations falling in $(0, \delta)$, the rest $N-n=m$ of the observations will all be greater than δ.

The likelihood function in this case is given by

$$\varphi(x_1, \ldots, x_m)=\binom{N}{n}\frac{\mu^{n\varkappa}}{\{\Gamma(\varkappa)\}^n}\left\{\int_0^{\delta}\exp(-\mu x)x^{\varkappa-1}\mathrm{d}x\right\}^n\frac{\mu^{m\varkappa}}{\{\Gamma(\varkappa)\}^m}$$

$$\times\exp\left(-\mu\sum_1^m x_i\right)\prod_{i=1}^{m}x_i^{\varkappa-1}, \quad \ldots\ldots\ldots\ldots\ldots \quad (6)$$

where $x_1, x_2, \ldots, x_m \geqslant \delta$. From this we get

$$L=\ln\varphi=\ln\binom{N}{n}+N\varkappa\ln\mu-N\ln\Gamma(\varkappa)+n\ln\int_0^{\delta}\mathrm{e}^{-\mu x}x^{\varkappa-1}\mathrm{d}x$$

$$-\mu\sum_{i=1}^{m}x_i+(\varkappa-1)\sum_{i=1}^{m}\ln x_i. \quad \ldots\ldots\ldots \quad (7)$$

Now when $\delta\to 0$

$$\int_0^{\delta}\mathrm{e}^{-\mu x}x^{\varkappa-1}\mathrm{d}x\sim\int_0^{\delta}x^{\varkappa-1}\mathrm{d}x=\frac{\delta^{\varkappa}}{\varkappa},$$

and substituting this value in (7) we get

$$L=\ln\binom{N}{n}+N\varkappa\ln\mu-N\ln\Gamma(\varkappa)+n\varkappa\ln\delta-n\ln\varkappa-\mu\sum_{i=1}^{m}x_i+(\varkappa-1)\sum_{i=1}^{m}\ln x_i.$$

$$\ldots\ldots\ldots\ldots\ldots\ldots \quad (8)$$

The maximum likelihood equations are consequently given by

$$\frac{1}{N}\frac{\partial L}{\partial \mu}=\frac{\varkappa}{\mu}-\frac{1}{N}\sum_{i=1}^{m}x_i=0, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (9)$$

$$\frac{1}{N}\frac{\partial L}{\partial \varkappa}=\ln\mu-\frac{\mathrm{d}\ln\Gamma(\varkappa)}{\mathrm{d}\varkappa}+\frac{n}{N}\ln\delta-\frac{n}{N\varkappa}+\frac{1}{N}\sum_{i=1}^{m}\ln x_i=0. \quad \ldots\ldots\ldots \quad (10)$$

For some values of $\varkappa$ and μ the approximation resulting from replacing

$$\int_0^\delta e^{-\mu x} x^{\varkappa-1} dx \text{ by } \frac{\delta^\varkappa}{\varkappa}$$

may not be good enough and then the solution of the equations is a little more laborious.

It is also easy to show that in most cases the loss of information arising because we do not have exact values x_i and consequently have to use (**6**) instead of (**2**), is quite small.

As a numerical example we give in Table 1 Sydney rainfall values for 2068 days.

TABLE 1

FREQUENCIES AS PREDICTED BY THE FITTED CURVE

Class Interval	f_0	$f_E^{(1)}$	$f_E^{(2)}$
0–5	1631	1638·5	(1614·0)
6–10	115	106·0	103·6
11–15	67	62·0	62·2
16–20	42	43·6	43·6
21–25	27	32·2	32·9
26–30	26	26·0	26·0
31–35	19	20·7	21·1
36–40	14	17·2	17·5
41–45	12	14·3	14·6
46–50	18	12·2	12·5
51–60	18	19·6	20·2
61–70	13	14·7	15·4
71–80	13	11·6	12·0
81–90	8	8·9	9·6
91–100	8	7·2	7·6
101–125	16	12·2	13·5
125–150	7	7·2	8·3
150–425	14	13·2	15·7

Now for estimating μ and $\varkappa$ from equations (**9**) and (**10**) we take $\delta=5$ and consequently we have $n=1631$,

$$\sum_{i=1}^{m} x_i = 16891 \quad \sum_{i=1}^{m} \ln x_i = 1373 \cdot 117.$$

Substituting these values in (**9**) and (**10**) we find that equation (**9**) reduces to $\mu=0 \cdot 1224\varkappa$ and equation (**10**) reduces to

$$\frac{d \ln \Gamma(\varkappa)}{d\varkappa} - \ln \varkappa - \frac{0 \cdot 2114}{\varkappa} + 0 \cdot 1703 = 0.$$

Solving these equations we get $\varkappa=0 \cdot 105$, $\mu=0 \cdot 013$ correct to three places of decimals.

Now for these values of μ, $\varkappa$, and δ we have

$$\int_0^{\delta} e^{-\mu x} x^{\varkappa-1} dx = 11 \cdot 14 \text{ and } \frac{\delta^{\varkappa}}{\varkappa} = 11 \cdot 30.$$

The calculation of the expected frequencies for testing the goodness of fit is shown in the table. 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 calculate the expected frequencies we use tables of the incomplete Γ-function (Pearson 1922) making a double linear interpolation, which is sufficiently accurate for our purpose.

The results are shown in Table 1. The first column gives the class interval, the column headed f_0 gives the corresponding observed class frequencies. The column headed $f_E^{(1)}$ gives the expected frequencies. Thus for 15 degrees of freedom the total χ^2 is found to be 7·8, which shows that the fit is an extremely good one.

III. Fitting of Truncated Curve Ignoring the Observations in the Truncated Part

Here we fit a truncated type III curve to the observations which are all greater than δ and ignore all observations which are less than δ. The probability density in this case is given by

$$f(x) = \frac{\mu^{\varkappa}}{\Gamma(\varkappa)} \cdot \frac{e^{-\mu x} x^{\varkappa-1}}{1 - \frac{\mu^{\varkappa}}{\Gamma(\varkappa)} \int_0^{\delta} e^{-\mu x} x^{\varkappa-1} dx}.$$

The logarithm of the likelihood function is given by

$$L = \ln \varphi(x_1, \ldots, x_m) = m\varkappa \ln \mu - m \ln \Gamma(\varkappa) - mG(\mu, \varkappa) - \mu \sum_{i=1}^{m} x_i + (\varkappa - 1) \sum_{i=1}^{m} \ln x_i, \quad \ldots\ldots\ldots\ldots\ldots\ldots (11)$$

where

$$G(\mu, \varkappa) = \ln \left\{ 1 - \frac{\mu^{\varkappa}}{\Gamma(\varkappa)} \int_0^{\delta} e^{-\mu x} x^{\varkappa-1} dx \right\}.$$

From (11) we obtain the following likelihood equations:

$$\frac{1}{m} \frac{\partial L}{\partial \mu} = \frac{\varkappa}{\mu} - \frac{\partial G}{\partial \mu} - \frac{\Sigma x_i}{m} = 0, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (12)$$

$$\frac{1}{m} \frac{\partial L}{\partial \varkappa} = \ln \mu - \frac{d \ln \Gamma(\varkappa)}{d\varkappa} - \frac{\partial G}{\partial \varkappa} + \frac{\Sigma \ln x_i}{m} = 0. \quad \ldots\ldots\ldots (13)$$

For these equations δ, Σx_i, and $\Sigma \ln x_i$ have the same values as before and $m = 437$. Equations (12) and (13) are generally too complicated to be solved. But we can, however, find approximate solutions and then improve these solutions to any desired degree of accuracy. Thus if $\mu = \mu_0$ and $\varkappa = \varkappa_0$ are an approximate solution for the equations (12) and (13) a better approximate

solution is obtained by taking $\mu=\mu_0+\delta\mu_0$, $\varkappa=\varkappa_0+\delta\varkappa_0$, where $\delta\mu_0$, $\delta\varkappa_0$ are given by the following equations:

$$\left(\frac{\partial^2 L}{\partial\mu^2}\right)_{\substack{\mu_0\\ \varkappa_0}}\delta\mu_0+\left(\frac{\partial^2 L}{\partial\mu\partial\varkappa}\right)_{\substack{\mu_0\\ \varkappa_0}}\delta\varkappa_0=-\left(\frac{\partial L}{\partial\mu}\right)_{\substack{\mu_0\\ \varkappa_0}}, \quad \ldots\ldots\ldots \quad (14)$$

$$\left(\frac{\partial^2 L}{\partial\mu\partial\varkappa}\right)_{\substack{\mu_0\\ \varkappa_0}}\delta\mu_0+\left(\frac{\partial^2 L}{\partial\varkappa^2}\right)_{\substack{\mu_0\\ \varkappa_0}}\delta\varkappa_0=-\left(\frac{\partial L}{\partial\varkappa}\right)_{\substack{\mu_0\\ \varkappa_0}}. \quad \ldots\ldots\ldots \quad (15)$$

In this particular example we took $\varkappa=0\cdot10$ and $\mu=0\cdot01$ as the approximate solution for equations (**12**) and (**13**): then with the help of equations (**14**) and (**15**) we found $\varkappa=0\cdot105$ and $\mu=0\cdot012$ correct to three places of decimal. The Newton-Gregory formula was used for numerical differentiation in calculating the $\partial G/\partial\mu$, $\partial^2 G/\partial\mu^2$, $\partial G/\partial\varkappa$, $\partial^2 G/\partial\varkappa^2$, and $\partial^2 G/\partial\mu\partial\varkappa$, etc., which occur in the solutions of equations (**14**) and (**15**). In this case, the approximation previously given for the integral in G is not sufficiently accurate. The expected frequencies are given in Table 1 in the column headed $f_E^{(2)}$. The fit of the observed frequencies to the expected values, ignoring the interval $(0, \delta)$ gives a $\chi^2=8\cdot47$ which shows that the fit is a good one.

The expected frequency in the interval $(0, \delta)$ as predicted by this fitted distribution is 1614. To test whether this is significantly different from the observed frequency in this interval we use the following statistic:

$$E=\frac{N_1-Np}{\sqrt{Npq+V(Np)}},$$

where N_1 stands for the number observed in the range $(0, \delta)$, N stands for the number of observations in the whole sample, and p is the probability of an observation falling in the range $(0, \delta)$ given by

$$p=\frac{\mu^\varkappa}{\Gamma(\varkappa)}\int_0^\delta \mathrm{e}^{-\mu x}x^{\varkappa-1}\mathrm{d}x.$$

Now according to our notation $\ln(1-p)=G(\mu,\varkappa)$,

$$\frac{-\mathrm{d}p}{1-p}=\frac{\partial G}{\partial\mu}\mathrm{d}\mu+\frac{\partial G}{\partial\varkappa}\mathrm{d}\varkappa,$$

$$\frac{V(p)}{(1-p)^2}=\left(\frac{\partial G}{\partial\mu}\right)^2 V(\mu)+\left(\frac{\partial G}{\partial\varkappa}\right)^2 V(\varkappa)+2\left(\frac{\partial G}{\partial\mu}\right)\left(\frac{\partial G}{\partial\varkappa}\right)\mathrm{cov}\,(\mu\varkappa).$$

For finding variance and covariances of μ and $\varkappa$ we have

$$\begin{pmatrix} \mathrm{v}(\mu) & \mathrm{cov}\,(\mu\varkappa) \\ \mathrm{cov}\,(\mu\varkappa) & \mathrm{v}(\varkappa) \end{pmatrix}=\begin{pmatrix} -E\left(\frac{\partial^2 L}{\partial\mu^2}\right) & -E\left(\frac{\partial^2 L}{\partial\mu\partial\varkappa}\right) \\ -E\left(\frac{\partial^2 L}{\partial\mu\partial\varkappa}\right) & -E\left(\frac{\partial^2 L}{\partial\varkappa^2}\right) \end{pmatrix}^{-1},$$

$$\begin{pmatrix} (2179\cdot4)m & (-87\cdot2)m \\ (-87\cdot2)m & (14\cdot3)m \end{pmatrix}^{-1}=\begin{pmatrix} \frac{0\cdot0006}{m} & \frac{-0\cdot0037}{m} \\ \frac{-0\cdot0037}{m} & \frac{0\cdot0924}{m} \end{pmatrix},$$

from which we get

$$V(p)=\frac{0\cdot477}{m}=0\cdot0011.$$

Thus we find $t=0\cdot24$ which gives no evidence of a concentration of probability at zero.

The distribution of mean rainfall based on 94 years from this type III population is given by

$$f(\bar{x})=\frac{(n\mu)^{n\varkappa}}{\Gamma(n\varkappa)}\mathrm{e}^{-n\mu\bar{x}}\bar{x}^{n\varkappa-1}=\frac{(1\cdot128)^{9\cdot870}}{\Gamma(9\cdot870)}\mathrm{e}^{-1\cdot128\bar{x}}\bar{x}^{8\cdot870}.$$

It corresponds approximately to a χ^2 distribution with 10 d.f. with skewness

$$\tfrac{1}{2}\frac{\mu_3}{\sigma^3}=0\cdot318.$$

Thus the distribution of the mean of 94 observations is far from normal and a test of significance of such a mean, based on a normal distribution, may be quite misleading.

IV. References

Bowen, E. G. (1953).—*Aust. J. Phys.* **6**: 490-7.
Cohen, A. C., Jr. (1950).—*J. Amer. Statist. Ass.* **45**: 411-23.
Des Raj (1953).—*J. Amer. Statist. Ass.* **48**: 336-49.
Hannan, E. J. (1955).—*Aust. J. Phys.* **8**: 289-97.
Pearson, K. (1922).—" Tables of the Incomplete Γ-Function." (Cambridge Univ. Press.)