A TEST FOR SINGULARITIES IN SYDNEY RAINFALL

By E. J. HANNAN*

[*Manuscript received November* 24, 1954]

Summary

A statistical test of the homogeneity of daily mean rainfalls (after the removal of a smooth seasonal trend) for Sydney over the years 1859 to 1952 does not contradict the hypothesis of homogeneity. The test was suggested by the theory that meteoric dust produces rainfall singularities, proposed by Bowen (1953). The test is made difficult by the complicated stochastic nature of the process generating the observations so that the significance level may in fact be lower than presumed. As the test statistic does not fall in the critical region, however, there can be no doubt that the test does not contradict the hypothesis (of homogeneity of means) at the presumed significance level.

I. INTRODUCTION

In the paper " The Influence of Meteoritic Dust on Rainfall ", by E. G. Bowen (1953), it was suggested that there were certain singularities in the daily rainfall patterns of Sydney (and other cities, mainly in the southern hemisphere) and that these singularities (which represented days of exceptionally high rainfall) were due to the annually recurring deposit of dust in the atmosphere from the passage of meteor showers.

A statistical test of the theory can be obtained in two ways. One way is is relate the days of high meteor activity to the days of high rainfall, these last having been objectively chosen. The second method is to test the homogeneity of the daily means, for a station or a number of stations, after the means have been adjusted for any smooth seasonal variation. In this paper the second method will be applied to the Sydney rainfall. In the process of preparing the data for this test a sound basis for objectively choosing the days of high rainfall will be obtained, so that the first method may later be used.

II. THE NATURE OF THE DATA AND OF THE NULL HYPOTHESIS

The distribution of daily rainfall is J-shaped. In an accompanying paper (Das 1955) it has been shown that, for a period of 22 days in October and November, a type III distribution fits very well. This best-fitting distribution is so extremely skew that observations more than 16·5 standard deviations away from the mean occur with a probability of 10^{-4}. If the distribution were normal with the same mean and standard deviation the occurrence of an observation more than 6·1 standard deviation units from the mean would have the probability 10^{-9}.

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

An indication of the very great departure from normality is also obtained by considering the daily variances. For example, over half of the January and February variances are " significantly " different from what is clearly a well-fitted trend, when treated as being distributed as χ^2 with 94 degrees of freedom. The explanation for this phenomenon is of course found in the high fourth moment of the distribution (see Box 1953).

This non-normality is the first of a number of difficulties which affect a test of the significance of the rainfall singularities. A second is the lack of independence between rainfalls on days near each other in time. A final difficulty is the seasonal variation in the data which affects not only the daily means and variances but also the degree of dependence between days.

A study of the correlations for lags of up to 4 days suggests that

$$\rho_{kj} \approx \rho_{1j}^{k}, \qquad k=1, 2, 3, 4.$$

Here ρ_{kj} is the correlation between rainfall on days j and $j-k$. For $k>4$ the correlation appears to be, effectively, zero. For example the first four lag correlations were computed from the observations over the 94 years between the dates October 17 and November 7 (2068 observations in all). Over this period the seasonal variation is small. The correlations were computed using the formulae

$$r_k = \frac{22}{22-k} \frac{\sum_{i=1}^{94} \sum_{j=k+1}^{22} (x_{ij}x_{ij-k}) - 94(22-k)\bar{x}^2}{\sum_{i=1}^{94} \sum_{j=1}^{22} (x_{ij}^2) - 94(22)\bar{x}^2}. \qquad (k=1, 2, 3, 4)$$

Here

$$\bar{x} = \frac{1}{2068} \sum_{i=1}^{94} \sum_{j=1}^{22} x_{ij},$$

and x_{ij} is the rainfall on the jth day of the ith year.

This particular formula was used mainly for computational convenience, the $\sum_{i=1}^{94} x_{ij}x_{ij-k}$ and $\sum_{i=1}^{94} x_{ij}$ having already been computed on the C.S.I.R.O. electronic computer so that the seasonal variation in the means, variances, and lag correlations could be estimated. The statistics r_k will provide consistent estimators of the corresponding ρ_k. A formula such as $(22-k)^{-1} \sum_{j=k}^{22} r_{kj}$, where r_{kj} is the observed correlation between days j and $j-k$, would not be satisfactory since the bias will remain the same as the bias of the individual r_{kj}. The statistics r_k, together with r_1^k, are given below.

$$r_1 = 0 \cdot 219\,; \quad \begin{cases} r_2 = 0 \cdot 057 \\ r_1^2 = 0 \cdot 048 \end{cases}; \quad \begin{cases} r_3 = -0 \cdot 007 \\ r_1^3 = 0 \cdot 011 \end{cases}; \quad \begin{cases} r_4 = 0 \cdot 048 \\ r_1^4 = 0 \cdot 002 \end{cases}$$

If these were computed from normally distributed observations their standard deviations would be about 0·02. The variance in the correlation coefficient depends upon the moments of the fourth order (Cramer 1946, p. 359)

and in the present case is likely to be underestimated by the formula, $(1-\rho^2)^2/n$, which applies for a normal distribution.

The high value of r_4 (compared with r_1^4) is entirely due to one cross product, that between the rainfall of 181 points on October 25, 1882 and 423 points on October 29, 1882, which is probably fortuitous and a result of the extraordinary skew parent distribution.

It is interesting to note that the serial correlations r_k in the rainfall data are here estimated by ordinary correlations (space averages), the data for the several years being treated as realizations of a stochastic process. This is a rather extraordinary situation for usually only one realization is available, from which the serial correlation is estimated as a phase average.

The relation $\rho_{kj}=\rho_{1j}^k$, which holds reasonably well for k small, at first sight suggests that the process can be represented by a model of the form

$$(x_j-m)=\rho(x_{j-1}-m)+\varepsilon_j, \qquad |\rho|<1.$$

Here m is the mean of the process, ρ the correlation between x_j and x_{j-1}, and ε_j is an independent random process with zero mean and variance $\sigma^2(1-\rho^2)$.

It is at once evident that this process cannot be used to describe rainfall data for not only is there a seasonal variation in ρ, m, and σ^2 but the variate x_j is necessarily positive so that $\varepsilon_j+m(1-\rho)$ would have to have a distribution admitting only positive values. This would, however, imply a smoothness in the series of x_j which would not be supported by the data. (A high x_j could not be followed by an x_{j+1} near to zero.)

A modified process which suggests itself is

$$\left(\frac{x_j-m_j}{\sigma_j}\right)=\rho_{1j}\left(\frac{x_{j-1}-m_{j-1}}{\sigma_{j-1}}\right)+\varepsilon_j. \qquad \ldots\ldots\ldots\ldots \quad (1)$$

Here m_j is the mean and σ_j^2 the variance of x_j, ρ_{1j} is the correlation between x_j and x_{j-1}, and ε_j is an independent random process with zero mean and variance $(1-\rho_{1j}^2)$.

It can be seen that the correlation between x_j and x_{j-k} will be

$$\rho_{kj}=\rho_{1j}\rho_{1j-1}\cdots\rho_{1j-k+1}.$$

If the lag k is not too large and the seasonal effect is small, over short periods, then

$$\rho_{kj}\approx\rho_{1j}^k.$$

Again this process will not be satisfactory for the daily rainfalls (since they are positive) but it may provide a good approximation to the process generating the daily mean rainfall (over 94 years) since the range of variation of the sample mean below the true mean will be much greater. Moreover, if the m_j, σ_j^2, and ρ_{1j} are equated to the means, variances, and first serial correlations of daily rainfall for day j (as estimated from a graduation of the observed values) the process (1) will have, approximately, the same means, variances, and serial correlations (for all lags) as the actual series.

A better approximation to the process generating the daily means could be got by including a term $(x_{j-2}-m_{j-2}/\sigma_{j-2})$. However, the nearness of the r_{2j} to r_{1j}^2 suggests that the coefficient of this term will be very small, for this coefficient will be, approximately, $(r_{2j}-r_{1j}^2)(1-r_{1j}^2)^{-1}$. The additional computations involved would not be justified by the very small effect from the inclusion of the extra term.

The model of the process generating the daily means which will be used is, therefore,

$$\sqrt{94}\frac{\bar{x}_j-m_j}{\{\sigma_j^2(1-\rho_{1j}^2)\}^{\frac{1}{2}}}=\rho_{1j}\sqrt{94}\frac{\bar{x}_{j-1}-m_{j-1}}{\{\sigma_{j-1}^2(1-\rho_{1j}^2)\}^{\frac{1}{2}}}+\varepsilon_j. \quad \ldots\ldots \quad (2)$$

Here $\bar{x}_j$ is the mean rainfall observed on the jth day of the year while m_j, σ_j^2, and ρ_{1j} are the true means, variances, and first serial correlations of rainfall on that day. The ε_j come from a process with zero mean and unit variance.

The specification of the null hypothesis will be completed if the nature of the seasonal variation in the means, variances, and correlation coefficients is laid down. All that needs to be said here is that these seasonal variations should be described by reasonably smooth curves such as a low order polynomial in j or a trigonometric polynomial formed from the first few harmonics. The choice among these alternatives depends on the goodness of their fit to the data, with the proviso that the fit should not be carried anywhere near to the point where individual singularities, confined to a period of a relatively few days, would be eliminated.

This restriction on the order of the polynomials means that a very large number of degrees of freedom are available for the estimation of the relatively few constants involved so that the m_j, $\hat{\sigma}_j^2$, and $\hat{\rho}_{1j}$ given by the estimated curves can be treated as the true values to a satisfactory degree of approximation and, when used to transform the $\bar{x}_j$ into the corresponding $\hat{\varepsilon}_j$, these can be regarded as uncorrelated random variates with zero mean and unit variance. Since the means $\bar{x}_j$ should have a distribution near to normality the sum $\chi_1^2=\Sigma\hat{\varepsilon}_j^2$ should be approximately distributed as χ^2 with $(365-p)$ degrees of freedom, where p is the number of constants fitted in the process of estimating the seasonal variation of the m_j, σ_j^2, and ρ_{1j}.

This test is the generalization of the classical test of the homogeneity of a set of means (from observations subject to different treatments) which is obtained by comparing the variance within treatments with that between the means themselves. In the present case the comparison is being made between the within-days variance of rainfall and the variance as estimated from the daily means. In addition, however, the data have had to be transformed to independence by the use of the ρ_{1j} and the within-days variances made homogenous while the effect of the seasonal variation in the means has been removed so that it will not void the test against the effect of the meteor showers.

This test is subject to a number of qualifications, however, as has already been indicated :

(*a*) In fact the seasonal pattern of the m_j, σ_j^2, and ρ_{1j} will not be known exactly, but will have been estimated, so that the variance of the statistic χ_1^2 will be increased by a component due to the variance of these estimates.

(*b*) The true nature of the underlying process generating the x_j will be more complicated than is indicated by (**2**) above, so that the residuals $\hat{\varepsilon}_j$ will not be independent. Their correlations may also be different from zero. This may increase the variance of χ_1^2 also. For the variance of this quantity will be

$$E\{\Sigma\hat{\varepsilon}_i^2\}^2 - \{E(\Sigma\hat{\varepsilon}_i^2)\}^2,$$

where E denotes expected value.

This expectation will include such terms as

$$E\{\hat{\varepsilon}_i^2\hat{\varepsilon}_j^2\}, \quad j \neq i.$$

If the $\hat{\varepsilon}_i$ were truly independent with zero mean and unit variance this expectation would be 1. If $\hat{\varepsilon}_i$ and $\hat{\varepsilon}_j$ are not independent the expectation may be different from 1. The high fourth moment of the daily rainfalls suggests that the second moment of the individual ε_i^2 will be high and that $E\{\hat{\varepsilon}_i^2\varepsilon_j^2\}$ may be greater than 1. This will increase the variance of χ_1^2 above its theoretical value.

(*c*) The extreme non-normality of the daily rainfall distribution will also increase the variance of χ_1^2. Since χ_1^2 is itself a sum of squares of standardized observations, it is clear that its variance will depend upon the fourth moment of these observations. If this fourth moment is greater than that of a standard normal variate the variance will be greater than it would be if it were exactly distributed as χ_{365-p}^2. In the present case the nature of the parent distribution makes it certain that the fourth moment will be high. The effect will, of course, be reduced by the fact that each $\hat{\varepsilon}_j$ is the sum of a large number of observations but, because of the very extreme parent distribution, it will probably persist.

(*d*) Since the large deviations in the present distribution are all positive the distribution of χ_1^2 will also have a greater positive skewness than it should theoretically have.

The effect of all of these factors will therefore be to increase the variance of χ_1^2 above its theoretical value. The fourth factor, which with the third seems to be the most important, will also make the distribution more skew (with a longer tail to the right) than it should have. The test of significance will consist of choosing a positive number, t, such that the probability of an observed χ_1^2 being greater than or equal to t (and therefore significant) is α. Both the increase in variance and the increase in positive skewness will make the probability of a significantly large χ_1^2 occurring (if the null hypothesis is true) greater than α.

If, therefore, the observed χ_1^2 is just beyond the critical point it will not follow that the result is significant at the level α since the true critical point may be further to the right by a sufficient amount to render this conclusion invalid. On the other hand, if the observed χ_1^2 is to the left of the critical point it can be said that the result is not significant and the null hypothesis will not be rejected at the level α.

This element of indeterminacy is of course a defect in the test. It is one which cannot easily be removed, however. Even if an exact distribution-free test, based, for example, on rearrangements of the observations, could be found it would almost certainly be much less powerful than the present test.

III. Removal of Seasonal Variation

In order to get a closer agreement between calendar and sidereal years the leap year was reinserted in 1900 so that March 1, 1900 became February 29, 1899, and so on.

Since a strictly periodic seasonal effect was being removed it seemed clear that a harmonic curve should be used to graduate the series. At the same time no attempt was made to fit a harmonic curve including terms with too short a period, as this would tend to remove the effect being sought.

For both means and variances the year was divided into 24 periods of equal lengths and the first four harmonics were fitted to the resulting 24 values. The intensities S^2 of these four harmonics, expressed as multiples of $4V^2/24$ are shown in Table 1. (Here V^2 is the estimate of the variance of the 24 values.)

The intensity corresponding to the period μ is derived from the relation

$$S^2=A^2+B^2,$$

$$A^2=\frac{2}{24}\Sigma_j u_j \cos\frac{2\pi j}{\mu}, \qquad B^2=\frac{2}{24}\Sigma_j u_j \sin\frac{2\pi j}{\mu},$$

where u_j is the mean (or variance) on the jth day.

Table 1

$\varkappa=24S^2/4V^2$

Period (days)	Means	Variances
365	8·53	9·57
$\frac{365}{2}$	0·09	0·06
$\frac{365}{3}$	0·17	0·10
$\frac{365}{4}$	0·13	0·24

There is, of course, no doubt of the significance of the period of 365 days. On the other hand the quantities $\varkappa$ for the other three periods are not significant at any reasonable level. It was also clear from an examination of the data that little could be gained by adding further terms unless their periods became quite short.

The first harmonic was therefore fitted to the means and variances using the individual daily averages. The resulting curves were :

$$\text{Means :} \qquad \hat{m}_j=12\cdot865+4\cdot377\sin\left(\frac{2\pi j}{365\cdot25}+0\cdot527\right)$$

$$\text{Variances :} \qquad \sigma_j^2=1755\cdot0+1015\cdot9\sin\left(\frac{2\pi j}{365\cdot25}+0\cdot571\right)$$

Here $j=1$ on March 1. The phase angles do not differ significantly.

A simple harmonic was also fitted to the between-days correlations, the resulting curve being

$$\hat{\rho}_{1j} = 0\cdot338 + 0\cdot103 \sin\left(\frac{2\pi j}{365\cdot25} - 0\cdot285\right).$$

Again $j=1$ on March 1.

The fit of this curve is not as good as that to the means and variances and it seems, in retrospect, that higher order terms should have been taken. In fact if the first harmonic were fitted to the covariances rather than the correlations a better fit would be obtained and it seems that the phase angle would not differ significantly from those for the means and variances. The effect on the residuals of the poor fit to the r_{1j} will be small, however (it will increase the variance of χ_1^2).

IV. Test of Homogeneity of Means

The residuals

$$\hat{\varepsilon}_j = \left(\frac{94}{1-\hat{\rho}_{1j}^2}\right)^{\frac{1}{2}} \left\{ \frac{\bar{x}_j - \hat{m}_j}{\hat{\sigma}_j} - \hat{\rho}_{1j} \frac{\bar{x}_{j-1} - \hat{m}_{j-1}}{\hat{\sigma}_{j-1}} \right\} \quad \ldots\ldots\ldots\ (3)$$

were formed. The $\hat{\varepsilon}_j$ had a small positive mean (0·018) and a small negative serial correlation (−0·05), both of which were far from significant.

The distribution of the $\hat{\varepsilon}_j$ approaches normality, as could be expected, but there is positive skewness.

There are 366 $\hat{\varepsilon}_j$, one resulting from the 24 rainfalls on February 29. (For this $\hat{\varepsilon}_j$ the factor $(24)^{\frac{1}{2}}$ will replace $(94)^{\frac{1}{2}}$ in (3).) Three constants have been fitted to each of the means, variances, and serial correlations so that χ_1^2 may be treated as distributed as χ^2 with 357 degrees of freedom. The value of $\Sigma\hat{\varepsilon}_j^2 = \chi_1^2$ is 394·1. Using Fisher's approximation to the distribution of χ^2 the quantity

$$\{2(394\cdot1)\}^{\frac{1}{2}} - \{2(357)-1\}^{\frac{1}{2}} = 1\cdot37$$

was computed. Treated as a standard normal variate and using one tail of the normal distribution this is well inside the 5 per cent. point (1·65).

The null hypothesis, therefore, cannot be rejected at this level of significance and the result of the test, while not of course disproving Bowen's hypothesis, cannot be said to justify it.

V. Alternative Tests

The test which has been given might not be very powerful against the alternative of only a few very widely scattered singularities since the effect of these in increasing χ_1^2 may be lost among the accompanying " noise ".

At first sight an alternative would be to pick out the high peaks in the residuals and test them as the largest among 366. For example, the highest peak in the residuals is on July 23 (24, 1901–1952). This residual is 4·037 times its (trend) standard deviation. This lies almost exactly on the 1 per cent. point for the largest out of 366 normally and independently distributed observations. In computing the probability of obtaining a *largest* observation, as

extreme as or more extreme than the largest observed, only the extreme tail area of the distribution of the individual observations is used. Moreover any error here will be multiplied, in the process of computing the first probability, by the number of observations (approximately). It is certain, from the shape of the parent distribution, that the tail area used will be too small compared with the true, unknown, probability. How badly misleading the assumption of normality can be here can be seen by considering the distribution of χ^2 for 102 degrees of freedom. (The largest number of degrees of freedom available in Pearson's tables of the "Incomplete Gamma Function".) This distribution, to the eye, would be indistinguishable from the normal distribution. In the

TABLE 2

DAYS FOR WHICH THE RESIDUALS ε_j ARE INDIVIDUALLY SIGNIFICANT AND POSITIVE

Date (prior to 1900)	Total Rainfall on this Date over 94 Years (points)	Total of Four Highest Rainfalls over 94 Years (points)	$\frac{\text{Col. 3}}{\text{Col. 2}} \times 100$
Jan. 11..	1792	1170	65
12..	2293	1296	57
Mar. 20..	2603	1307	50
Apr. 6..	3293	1943	59
May 6..	2382	1159	49
25..	2758	1412	51
July 23..	2826	1043	37
Aug. 2..	1866	1148	62
23..	1635	613	37
Sept. 24..	1512	836	55
28..	1754	1067	61
Oct. 12..	1452	1066	73
29..	1186	743	63
Nov. 17..	1315	540	41
18..	1897	871	46
Dec. 1..	1505	905	60

accompanying paper Das has shown that for the period from October 17 to November 7 the distribution of the daily means is close to χ^2 with 10 degrees of freedom. The distribution of the residuals $\hat{\varepsilon}_j$ is therefore much further from normal than χ^2_{102}. The 5 per cent. point of the greatest out of 366 observations from χ^2_{102} is $4 \cdot 2$ standard deviation units from the mean. The corresponding 5 per cent. point for the normal distribution is $3 \cdot 6$ standard deviation units from the mean. The largest observed rainfall residual, while significant at the 1 per cent. point on the basis of normality, is not significant at the 5 per cent. point on the basis of χ^2_{102}. This result suggests that true probability of a largest residual $\geqslant 4 \cdot 037$ is much higher than $0 \cdot 05$. The large effect on the computed probabilities of small variations in the nature of the parent distribution certainly makes the testing of the significance of the large observations an impossible task.

Further and more powerful tests can perhaps be obtained by a comparison of the sequence of meteor showers throughout the year with the days of high residuals or by considering data for a number of different stations.

It is of some interest to examine the days for which the residuals $\hat{\varepsilon}_j$ are individually " significant " (and positive), say at the 1 per cent. point. These days are listed in Table 2 (under their dates prior to 1900), which also shows the total rainfall on each of these days over the 94 years and the total rainfall on the 4 occasions of highest rainfall.

The highest rainfall over the years for these 16 days is, on an average, 22·7 per cent. of the total rainfall. Most or all of the deviation from the mean seems therefore to be due to a relatively few very heavy falls of rain. This phenomenon is, to some extent at least, explained by the very skew distribution of daily rainfall.

VI. Acknowledgments

I wish to thank Dr. E. G. Bowen for suggesting this investigation, and Mr. Pearcey and Mr. Beard of C.S.I.R.O. for the assistance they gave with the C.S.I.R.O. electronic computer.

VII. References

Bowen, E. G. (1953).—*Aust. J. Phys.* **6**: 490.
Box, G. E. P. (1953).—*Biometrika* **40**: 318.
Cramer, H. (1946).—" Mathematical Methods of Statistics." (Princeton Univ. Press.)
Das, S. C. (1955).—*Aust. J. Phys.* **8**: 298.