This paper discusses general considerations that arise in the statistical analysis of point stochastic processes (series of events) and a computer program called SASE *designed to implement such an analysis.*

The program is written as a sequence of independent subroutines. The computations performed in each subroutine are described and an example of an analysis of a series of events is presented and discussed.

A computer program for the statistical analysis of series of events

by P. A. W. Lewis

The purpose of this paper is to familiarize the reader with a computer program for performing a statistical analysis of a series of events (point stochastic process). First, we define what is meant by a series of events, give a general outline of the program, and discuss in general terms the types of analysis that might be performed on a series of events.

series of events

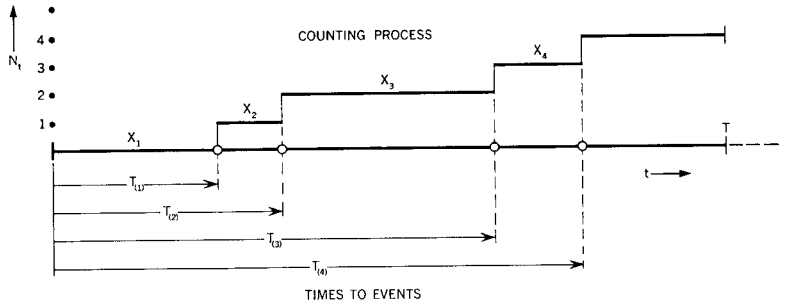
Series of events or point stochastic processes arise in many technological and scientific contexts. Typical examples are:

- The series of failures of a computer
- The series of arrivals at a queue or service facility
- The series of times of vehicles passing a point on a road
- The occurrence of pulses at a nerve junction
- The successive level crossings of a continuous time-parameter stochastic process
- The series of emissions of particles from a radioactive source

There are basically two situations in which an analysis of a series of events is required. In the first, a physical system may be observable only in terms of its output, the output being a series of events. It is then required to infer something about the structure of the system from a statistical analysis of the output series of events.

An example of this first type of situation is an analysis of three series of computer failures performed by Lewis,[1] who found significant deviations from the predictions of a standard reliability

Figure 1 Typical series of events



model. The deviations were found to be due to imperfect maintenance of the computers; the size of this maintenance effect was determined in the statistical analysis of the series.

Again, there is much interest in the statistical analysis of nerve pulse data,[2] because the series of pulses are observable, while the mechanisms which generate them are not. Models of the generating mechanism can be constructed and used to predict the probabilistic structure of the series of events. Comparison of this predicted structure with the structure of the observed series then provides a means of verifying the model of the generating mechanism.

The second situation requiring an analysis of a series of events is the design of systems whose structures are known and whose inputs are series of events. Two examples of such analyses are the following.

In designing queuing and service systems, the adequacy of the proposed service system to handle the input traffic depends critically upon the statistical structure of the input. The input is a series of events, the events being the arrivals of customers at the queuing system. The required analysis of the input can be performed with the aid of the program described in this paper.

Again, errors occur in transmitting digital data over telephone or other types of circuits—the errors being a series of events. The performance of an error-detecting and -correcting code from such data depends upon the statistical structure of the series of errors. Analyses of such error data have been given by Berger and Mandelbrot[3] and Lewis and Cox.[4]

We consider only univariate series of events; that is, series in which the events are distinguishable only by where they occur in time, as shown by Figure 1. In other words, quantitative or qualitative information associated with each event, e.g., type of failure, height of nerve pulse, or speed of a car passing a point on a road, is unavailable or ignored for present purposes. Consequently, the times-to-events $\{T_{(i)}\}$ or times-between-events $\{X_i\}$ completely characterize the process. Thus we have

$$0 < T_{(1)} < T_{(2)} < T_{(3)} < \cdots$$

(Throughout the paper parenthetical indices denote quantities ordered by magnitude).

Another equivalent characterization of the series of events is in terms of the counting process N_t, the number of events occurring in the interval $(0, t]$. The counting process N_t is a continuous time-parameter stochastic process whose sample functions are jump functions. We have $N_t < n$ if and only if

$$T_{(n)} = \sum_{i=1}^{n} X_i > t \qquad n = 1, 2, \cdots \tag{1}$$

and

$$\text{prob}\,(N_t < n) = \text{prob}\,(T_{(n)} > t) \qquad n = 1, 2, \cdots \tag{2}$$

Equations 1 and 2 specify the fundamental relationship between the counting process representation of a series of events and the interval representation. The main implication of this relationship for a statistical analysis of a series of events is that an analysis based on second-order correlational properties of the counting process, N_t, is in general not equivalent to an analysis based on the second-order correlational properties of the interval process, $\{X_i\}$. Both of these types of analysis are discussed in this paper.

The analysis of the interval process, $\{X_i\}$, is basically the analysis of a time series consisting of positive random variables, so that the usual normal theory does not hold. However, the analysis of the counting process N_t has no counterpart in ordinary time series analysis.

SASE program

We discuss now the various types of analyses implemented by a computer program called SASE,[5] written to assist in the statistical analysis of series of events. The theory behind the analyses is given in a recently published monograph.[6] This monograph is the first comprehensive account to appear on a relatively unexplored area of statistical analysis. As far as is known, the SASE program is the only program available to implement this type of analysis.

Two cases arise in practice in the statistical analysis of series of events that are differentiated by the program and give rise to fairly subtle differences in the formal analysis.

Case 1. The series is observed for a fixed length of time, T, and n events are observed in this time period. Here n is the observed value of the random variable N_T.

Case 2. The series is observed up to the occurrence of a fixed number, n, of events. The total time of observation, $t_{(n)}$, is the observed value of the random variable, $T_{(n)}$. This situation is indicated to the program by setting $T = 0$.

In order to accommodate the large number of possibilities that arise in analyzing series of events, the computer program was broken up into subroutines, most of them being independent

Table 1 The SASE program

Subroutine	*Computation*
MAIN	Control of data and subroutines
TREND*	Tests for trends in series
EXPO DURB*	Tests for a Poisson process
INTER	Marginal distribution of times-between-events $\{X_i\}$
RHO* SPEC*	Second-order joint properties of $\{X_i\}$ and tests for renewal processes
VART*	Second-order properties of the counting process N_t
COV* DENS* BART*	Second-order joint properties of N_t

*Subroutine executed upon control card indication only

of the other subroutines and capable of being suppressed if not needed. These subroutines and their general functions are shown in Table 1. An asterisk indicates that the subroutine is performed only when a suitable indication is given on a control card. Subroutines EXPO and INTER are always carried out.

general considerations

The computations performed by these subroutines and their interrelations are discussed in more detail later in this paper. First, we discuss some general considerations in the statistical analysis of series of events.

There are roughly two situations which arise in the analysis of series of events:

- Exploratory analyses in which no particular model is being put forward and the gross features of the data are being examined
- Analyses in which specific models are to be tested against data and parameters are to be estimated

In the first situation, the exploratory analysis may be used to suggest a pertinent model, or may be used to give direction to a further search for the physical mechanisms which generate the data. Graphical analysis is particularly important here, and the output of the program has been designed to facilitate this. In particular, it is always appropriate to examine the data for trends. For example, a plot of the cumulative number of events against time (the observed realization of N_t) may sometimes show these trends immediately. The existence of several specific types of trend in the series can be checked formally by computations performed in subroutine TREND.

If no trends are found in the data, the assumption is made that the series of events is stationary, which implies that the marginal distributions of the X_i's are identical. The next step

in the exploratory analysis is generally to determine whether serial correlation exists between the successive X_i's (subroutines RHO and SPEC). If no positive indications are obtained, one can assume that the X_i's are independent and identically distributed with an unknown distribution $F(x)$, i.e., the sequence $\{X_i\}$ is a renewal process. This is the usual case of a random sample considered in ordinary statistical analysis. The only remaining problem is to find a suitable model for $F(x)$. This modeling is facilitated by the output of subroutine INTER. Given a suitable model, standard methods such as maximum likelihood may be used to estimate parameters in the model.

A central role is played in the analysis of series of events by the Poisson process, which is a special case of a renewal process where

$$F(x) = \text{prob}\,(X \leq x) = 1 - e^{\lambda x} \tag{3}$$

Then, as is well known and may be verified from the fundamental relationship given by Equations 1 and 2

$$\text{prob}\,(N_t = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \qquad n = 0, 1, 2, \cdots \tag{4}$$

In addition to the property of independent intervals, the Poisson process has the property of independent increments. This property is that the numbers of events in any set of nonoverlapping intervals are independent random variables with the Poisson distribution given by Equation 4. Tests for a Poisson process are performed in subroutines EXPO, DURB, and to some extent in DENS and BART, which are concerned with estimating the second order properties of the counting process, N_t.

In the case of analyses involving specific models, the models are generally put forward as a result of an exploratory analysis or on the basis of prior knowledge. Renewal and Poisson models were discussed earlier in this paper. For other models which postulate serially correlated intervals, an initial analysis is the confirmation that a renewal model is not an adequate representation of the data. Tests-of-fit of these models with serially correlated intervals use the second-order properties of the interval process $\{X_i\}$ (subroutines RHO and SPEC) and the second-order properties of the counting process N_t (subroutines VART, COV, DENS, and BART). The fundamental relationship given by Equations 1 and 2 shows that these processes are equivalent only in terms of their complete distributions. Consequently, analyses based on second-order properties of counts and intervals are not, in general, equivalent. Estimation of parameters in these models is done usually in an ad hoc manner. The role of prior knowledge is very strong here, but is difficult to formalize because likelihood functions usually cannot be written down.

The detailed analysis given by Lewis for a series consisting of the successive times of failures of a computer is an example of analysis involving a specific model.

Computational aspects of the program

We discuss now in more detail the computations performed in each subroutine and the analyses based on these computations.

tests for trend

Subroutine TREND. The first computation in this subroutine gives a test for a trend in the rate of occurrence of events represented by a smooth change in time. Instead of the rate parameter in a Poisson process being assumed constant in time, it is assumed to have the functional form

$$\lambda(t) = e^{\alpha+\beta t} \tag{5}$$

so that

$$\text{prob}\,(N_t = n) = \frac{u^n e^{-u}}{n!}$$

where

$$u = \int_0^t \lambda(v)\, dv$$

The test is for the hypothesis $\beta = 0$, α being essentially a nuisance parameter since the test is for the null hypothesis of a Poisson process per se. Note that locally, near $\beta = 0$, Equation 5 is equivalent to a linear trend. From the likelihood function for observations of a series over a time T, it can be shown that the best test for $\beta = 0$ against $\beta \neq 0$ is based on the distribution of the statistic

$$S = \sum_{i=1}^{n} t_{(i)}/n \tag{6}$$

conditionally on the observed value n of N_T. Given n, and for $\beta = 0$, S has the distribution of the sum of n independent rectangular random variables. Consequently, the distribution of the standardized random variable

$$U = \frac{S - \frac{1}{2}T}{T/(12n)^{\frac{1}{2}}} \tag{7}$$

goes very rapidly to the standardized normal form as n increases. Essentially, the centroid of the observed times-to-events, $t_{(i)}$, is compared to the midpoint of the period of observation. The test based on Equation 7 is an optimum test against the trend given by Equation 5.

The program prints out the value of U observed and the separate quantities in it. If U is greater than 1.96, corresponding to significance at a 5 per cent level, the program stops and prints out INDICATION OF TREND AT 5% LEVEL.

The remainder of this subroutine computes quantities that are useful in tests for trend based on standard least squares regression methods. These methods are flexible, allowing for various types of trends to be tested, and work reasonably well under fairly weak assumptions about the detailed structure of the series in question.

The basic idea is that if the time between events X has a gamma distribution with parameters λ and a,

$$f_X(x) = \frac{\lambda^a x^{a-1} e^{-\lambda x}}{\Gamma(a)}$$

then $\log X$ has a log chi-square distribution with

$$E(\log X) = -\log \lambda + \psi(a)$$

$$\text{var } (\log X) = \psi'(a)$$

Here $\psi(a)$ is the derivative of the logarithm of the gamma function $\Gamma(a)$, and is called the digamma function,

$$\psi(a) = \frac{d}{da} \log \Gamma(a)$$

Now, let X_i have a gamma distribution with parameters a and

$$\lambda_i = e^{\alpha + \beta y_i}$$

Then if the times between events X_i (or contiguous groups of times-between-events) are independent, the series of values $\log X_i$ have the usual linear model with uncorrelated residuals and variances independent of the mean values $\psi(a) - \alpha - \beta y_i$. Thus it is possible to obtain least-squares estimates of α and β, and to test approximately the null hypothesis $\beta = 0$ by standard regression methods.

In the model, y_i can be defined as the time at the center of the interval if λ is considered a function of time. Another possibility is to define y_i as the serial number i if λ is considered a function of serial number. Still another possibility is to define y_i as the average of an independent variable that controls the rate of occurrence λ.

It is also possible to consider multiple regression models such as one in which λ_i is a quadratic function of Z_i. To assist in the regression analysis, the subroutine computes and prints out the following quantities for a constant K specified in the input to the program:

$$ZI = t_{(Ki)} - t_{(Ki-K)} = x_{Ki} + x_{Ki+1} + \cdots + x_{Ki-K}$$

ln ZI

$$MXI = \tfrac{1}{2}[t_{(Ki)} + t_{(Ki-K)}] \qquad I = i = 1, 2, 3, \cdots, [n/K]$$

where $[n/K]$ is the greatest integer less than or equal to n/K, and $t_0 = 0$. Computations are repeated using the values $2K$ and $3K$. The subroutine also prints out the estimated mean and coefficient of variation of successive sets of K, $2K$, and $3K$ intervals. These quantities are useful in determining a suitable trend model for the data.

Subroutine EXPO. Tests of a Poisson hypothesis for an observed series of events are computed by subroutine EXPO and subroutine

DURB, described later in this paper. There are three broad categories of alternatives to the Poisson hypothesis:

tests for Poisson process

- Non-stationarities or trends in the series
- Stationary series having independent intervals with a non-exponential distribution, i.e., renewal processes
- Stationary series with serially correlated intervals between events

The first alternative is specifically taken into account in subroutine TREND, but tests based on the statistics computed in subroutine EXPO are also somewhat sensitive to this alternative. Tests against general alternatives are based on the following idea.

In testing for a Poisson hypothesis for a series observed for a fixed period T, the parameter λ in Equation 4 is a nuisance parameter with a sufficient statistic n. (The quantity n is the observed number of events in the interval of length T.) The test should therefore be based on the distribution of the observations, conditionally upon the observed value of n. With this condition, the quantities

$$y_{(i)} = t_{(i)}/T \qquad i = 1, 2, \cdots, n$$

are, under the null hypothesis, the order statistics from a random sample of size n from a population uniformly distributed over (0, 1):

$$\text{prob}\,(Y_i \leq y) = \begin{cases} 0 & y \leq 0 \\ y & 0 < y \leq 1 \\ 1 & y > 1 \end{cases}$$

Similarly, when the series is observed up to the nth event ($T = 0$), the quantities

$$y_i = t_i/t_{(n)} \qquad i = 1, \cdots, (n-1)$$

are independent observations from a uniformly distributed population.

A test for the Poisson hypothesis based on testing the uniform distribution of the y_i's is called a *uniform conditional test.* This, however, is the canonical form of all distribution-free tests of goodness-of-fit, and four of the many possible distribution-free statistics are computed by the EXPO subroutine.

The one-sided Kolmogorov-Smirnov statistics. Denote by $F_n(y)$ the empirical distribution function of the observations y_i:

$$F_n(y) = \frac{\text{number of } y_i \leq y}{n} \qquad \text{where} \qquad 0 \leq y \leq 1$$

Subroutine EXPO computes

$$KS+ = D_n^+ = (n)^{1/2} \sup_{0 \leq y \leq 1} [F_n(y) - y] = (n)^{1/2} \max_{1 \leq i \leq n} \left[\frac{i}{n} - y_{(i)}\right]$$

and

$$KS- = D_n^- = (n)^{1/2} \sup_{0 \le y \le 1} [y - F_n(y)]$$

$$= (n)^{1/2} \max_{1 \le i \le n} \left[y_{(i)} - \frac{(i-1)}{n} \right]$$

The two-sided Kolmogorov-Smirnov statistic.

$$KS = D_n = (n)^{1/2} \sup_{0 \le y \le 1} |F_n(y) - y| = \max (D_n^+, D_n^-)$$

The Anderson-Darling statistic.

$$WN2 = W_n^2 = n \int_0^1 \frac{[F_n(y) - y]^2}{y(1-y)} \, dy$$

$$= -n - \frac{1}{n} \sum_{i=1}^{n} \{(2i-1) \ln y_{(i)} + [2(n-i)+1] \ln (1 - y_{(i)})\}$$

Percentage points of the Kolmogorov-Smirnov statistics are given in most statistical tables; those for W_n^2 are given by Cox and Lewis.[6] The four tests given are not consistent against certain stationary alternatives,[7] and are most sensitive to trend alternatives. The modified tests given later in subroutine DURB, however, give relatively powerful tests of the Poisson hypothesis.[7]

Tests of the Poisson hypothesis based on the property of independent increments are sometimes used, the best known of these being the test based on the index of dispersion. This test, however, can be shown to be equivalent to the uniform conditional test when the uniformity of the y_i's is tested with the chi-square test of goodness-to-fit. The drawback of this test is the need to choose a suitable grouping interval.

Specific tests of the Poisson hypothesis against renewal hypotheses are known. Of these, the most useful is based on the Moran statistic

$$MORAN = -2 \sum_{i=1}^{n} \ln [y_{(i)} - y_{(i-1)}] - 2n \ln (n)$$

where

$$y_0 = 0$$

The value of the statistic is computed and printed out by the EXPO subroutine. The test for a Poisson hypothesis based on this statistic is asymptotically the most powerful test against a renewal alternative in which the intervals have a gamma density

$$f(x) = \lambda^a x^{a-1} e^{-\lambda x} / \Gamma(a) \qquad a > 0$$

The test is for $a = 1$ against $a \neq 1$, and the statistic has a chi-square distribution with $n - 1$ degrees of freedom for large n.

Subroutine DURB. Denoting the intervals between events by $x_1, x_2, \cdots, x_n$ and the interval between the last observed event

and the end of the observation period by $x_{n+1} = T - t_{(n)}$, we order the $(n + 1)$ x_i's by magnitude to obtain the observed order statistics

$$0 < x'_{(1)} \leq x'_{(2)} \leq \cdots \leq x'_{(n)} \leq x'_{(n+1)}$$

The DURB subroutine then calculates the quantities

$$w_{(i)} = \frac{x'_{(1)}}{T} + \frac{x'_{(2)}}{T} + \cdots + \frac{x'_{(i-1)}}{T} + (n + 2 - i)\frac{x'_{(i)}}{T} \qquad (8)$$

$$i = 1, \cdots, n$$

One interval, $x_{(n+1)}$, is not used in the computation.

Under the null Poisson hypothesis, the w_i's have the same distributional properties as the y_i's computed in subroutine EXPO. The null hypothesis is again tested by computing the statistics $KS+$, $KS-$, KS, and $WN2$. The reason for using the transformation is that it is conjectured to give a large increase in power, relative to the uniform conditional tests, for a broad class of alternatives.

In the case where observation is up to the nth event ($T = 0$), the n intervals are ordered to give the $x'_{(i)}$'s. The $x'_{(i)}$'s, divided by $t_{(n)}$, are used as in Equation 8 to give $(n - 1)$ $w_{(i)}$'s.

marginal distribution of intervals

Subroutine INTER. Computations performed in this subroutine are designed to facilitate the graphic and numerical examination of the marginal distribution of the sequence of intervals between events $\{X_i\}$. The first part of the subroutine orders the n observed x_i's to obtain the observed order statistics

$$0 < x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

Note that the interval $x_{(n+1)}$ is not included as in subroutine DURB where observation of the series is for a fixed period. These order statistics are used in graphic displays of the empirical distribution function $F_n(x)$ for the observed intervals

$$F_n(x) = \frac{\text{number of } x_i\text{'s} \leq x}{n}$$

$$= \begin{cases} 0 & x < x_{(1)} \\ \dfrac{i}{n} & x_{(i-1)} \leq x < x_{(i)} \qquad i = 2, 3, \cdots, n \\ 1 & x_{(n)} \leq x \end{cases}$$

The subroutine prints out i, $x_{(i)}$, and i/n in successive columns. The next column lists $i/(n + 1)$; conventionally, this quantity is the point plotted on the ordinate against $x_{(i)}$ when n becomes so large that it is inconvenient to show the steps of size $1/n$.

The function $F_n(x)$ is a non-parametric estimate of the unknown marginal distribution function $F(x)$. It is often convenient to work with the empirical survivor function

$$R_n(x) = 1 - F_n(x)$$

and the logarithm of that function. To facilitate this, the subroutine prints out columns containing the quantities $(n - i)/n$, $(n - i + 1)/(n + 1)$, $\ln [(n - i)/n]$, and $\ln [(n - i + 1)/(n + 1)]$ respectively. The logarithmic plot is useful because if $F(x)$ is an exponential distribution, then

$$\ln R(x) = -\lambda x$$

In certain cases, systematic departures from linearity can be given specific interpretations.

The last three columns of the printout contain the exponential scores, the serially ordered intervals x_i, and the successive times-to-events $t_{(i)}$. The exponential scores are the expected values of the order statistics from an exponential population of size n given by

$$ESi = ESI = \sum_{l=1}^{i} \frac{1}{(n + 1 - l)}$$

These exponential scores have various uses in formal statistical procedures.[6]

The second part of the INTER subroutine computes and prints out the first three sample moments of the intervals between events

$$MU = \tilde{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$VAR = \tilde{\mu}_2 = \frac{1}{(n - 1)} \sum_{i=1}^{n} (x_i - \tilde{\mu})^2$$

$$MU3 = \tilde{\mu}_3 = \frac{n}{(n - 1)(n - 2)} \sum_{i=1}^{n} (x_i - \tilde{\mu})^3$$

and the related quantities

$$SIGMA = \tilde{\sigma} = (\tilde{\mu}_2)^{\frac{1}{2}}$$

$$C = \tilde{\sigma}/\tilde{\mu}$$

$$SKEW = \tilde{\gamma}_1 = \tilde{\mu}_3/(\tilde{\sigma})^3$$

The quantity C is an estimate of the coefficient of variation having the value unity for an exponential population, while $SKEW$ is an estimate of the standard measure of skewness.

Subroutine RHO. Subroutines RHO and SPEC are concerned with the computation of estimates of quantities which characterize the second order, joint properties of the intervals between events. These quantities are the serial correlation coefficients

serial correlation coefficients

$$\rho_j = \frac{E\{[X_i - E(X)], [X_{i+j} - E(X)]\}}{\text{var } (X)} = \frac{\text{cov } (X_i, X_{i+j})}{\text{var } (X)} \qquad j = \cdots, -1, 0, 1, \cdots \tag{9}$$

which are the Fourier coefficients of the second quantity of interest, $f_+(\omega)$, the spectral density function:

$$f_+(\omega) = \frac{1}{\pi}\left[1 + 2\sum_{i=1}^{\infty} \rho_i \cos(j\omega)\right] \qquad 0 \leq \omega \leq \pi \qquad (10)$$

The estimates of the serial correlation coefficients computed in subroutine RHO are the standard ones

$$\tilde{\rho}_j = \frac{\frac{1}{(n-j)}\sum_{i=1}^{n-j}(x_i x_{i+j}) - \frac{1}{(n-j)^2}\left(\sum_{i=1}^{n-j} x_i\right)\left(\sum_{i=1}^{n-j} x_{i+j}\right)}{\left[\frac{1}{(n-j)}\sum_{i=1}^{n-j} x_i^2 - \frac{1}{(n-j)^2}\left(\sum_{i=1}^{n-j} x_i\right)^2\right]^{1/2}\left[\frac{1}{(n-j)}\sum_{i=1}^{n-j} x_{i+j}^2 - \frac{1}{(n-j)^2}\left(\sum_{i=1}^{n-j} x_{i+j}\right)^2\right]^{1/2}} \qquad (11)$$

The $\tilde{\rho}_j$'s and the values $(n-j)^{\frac{1}{2}}\tilde{\rho}_j$ are computed and printed out for $j = 1, 2, \cdots$, up to the greatest integer less than $n/2$, or 100, whichever is smaller.

When indicated by special input instructions, the subroutine also computes estimates of the serial covariances normalized by the value of the estimate of var(x) computed in subroutine INTER, i.e., Equation 11 with the denominator replaced by VAR. This quantity, $RHOVJ$, is computed for all $j \leq n-1$ for use in subroutine SPEC when the special input indicator is given to the program.

The estimated correlation coefficients, $\tilde{\rho}_j$, provide a simple but rough means of testing for the presence or absence of serial correlation. Under the null hypothesis $\rho_j = 0$, $j = 1, 2, \cdots$, and provided the marginal distribution of the intervals is not too highly skewed, the $\tilde{\rho}_j$ may be considered as observed values of a standardized normal variate, i.e., their distribution is $N[0, 1/(n)^{1/2}]$. This approximation is reasonable for $n \geq 100$, say, if the skewness is moderate.

Subroutine SPEC. In this subroutine we compute smoothed estimates of the spectral density function, $f_+(\omega)$. Such estimates have the general form

$$\tilde{f}_+(\omega) = \frac{1}{\pi}\left[1 + 2\sum_{j=1}^{\infty} \tilde{\rho}_j \lambda_j \cos(j\omega)\right]$$

where the λ_j's are suitably chosen weights. The weight sequence used in subroutine SPEC is due to Bartlett.[8]

$$\lambda_j = \frac{m-j}{n} \quad \text{if} \quad j \leq m$$

$$= 0 \quad \text{if} \quad j > m$$

Therefore, the computation gives

$$FWI = \tilde{f}_+(\omega) = \frac{1}{\pi}\left[1 + 2\sum_{j=1}^{m-1}\left(\frac{m-j}{m}\right)\tilde{\rho}_j \cos\left(\frac{2j\pi I}{n}\right)\right]$$

The *FWI*'s are computed for

$$I = \begin{cases} 0, 1, 2, \cdots, n/2 & n \text{ even} \\ 0, 1, 2, \cdots, \dfrac{n-1}{2} & n \text{ odd} \end{cases}$$

and for three different values of m, (m_1, m_2, m_3), specified in the input to the subroutine. The coefficient of variation of the individual estimates is approximately $[(2m)/(3n)]^{1/2}$. Decreasing the coefficient of variation by decreasing m gives greater smoothing in the estimated spectrum, but less ability to resolve distinctive features in the underlying spectral density. Generally, it is possible to see empirically what the limit of resolution is for a given sample size n by using three different values of m.

The particular weight sequence used here is adequate for most purposes; if necessary, estimates using more elaborate weight schemes can be performed with standard programs for time-series analysis.

It is sometimes required to compute the unsmoothed estimate of the power spectral density, $\text{var}(X)f_+(\omega)$, for use in testing the independence of the sequence of intervals $\{X_i\}$. This estimate is sometimes called the *periodogram* and is computed on special input instructions to the subroutine. The computation is as follows:

$$PERIO = \frac{1}{\pi}\left[1 + 2\sum_{j=1}^{n-1}\left(\frac{n-j}{n}\right) RHOVJ \cos\left(\frac{2j\pi I}{n}\right)\right]$$

The quantity *RHOVJ* is computed in subroutine RHO (previously discussed) and has as divisor the estimated variance of X. This division is used to normalize the periodogram for convenience of computation. In test procedures, ratios of the periodogram values, *PERIO*, at different values of I are used and therefore the normalization does not affect the results. The computed values *PERIO* are printed out under this heading in a column adjacent to the smoothed spectral estimates.

The SASE program is limited to a sample size of $n = 999$. This limitation, together with the form of correction for a non-zero mean used in Equation 11, means that little computation speed is gained by using the fast Fourier transform of Cooley and Tukey[9] in the computation of the *FWI*'s and *PERIO*.

Tests for independence of the intervals $\{X_i\}$ are based on the following result. If the sequence of observed intervals are observations on independent, normally distributed random variables, then the values computed by *PERIO* (for $I \neq 0$ and $I \neq n/2$ if n is even), multiplied by *VAR*, are observations on independent exponentially distributed random variables with parameter

$$\lambda = \pi/[\text{var } (X)]$$

This result is asymptotically true for independent but non-normally distributed sequences $\{X_i\}$. If the fourth cumulant κ_4 of the X_i's is small, the result is approximately true even for

moderate sample size, say $n \geq 100$. Thus the test for independence consists in testing whether the values computed by *PERIO* for successive I are observations on a Poisson process.

The computational procedure is as follows:

Case 1. If n is even, then drop off *PERIO* for $I = 0, n/2$.

Case 2. If n is odd, then drop off *PERIO* for $I = 0$.

The remaining periodogram values, *PERIO*, $n' = (n/2) - 1$ of them in Case 1, and $n' = (n/2) - 1/2$ of them in Case 2, are put back into the whole program and processed by subroutines TREND and EXPO with $T = 0$, $n = n'$ and

$x_1' = PERIO$ for $I = 1$

$x_2' = PERIO$ for $I = 2$

In TREND we take $K = [n'/18]$.

The statistics computed in EXPO give direct tests of independence, and the output of TREND is used for tests based on the homogeneity of the variance statistic. The tests based on the periodogram are, in general, more exact and probably more powerful than tests based on the estimated serial correlation coefficients. However, owing to the approximation involved, tests based on the periodogram always measure to some degree the departure from normality of the marginal distribution of intervals. This departure is always present, since we deal with positive random variables. The large amount of computation involved in these tests should be noted; they would never be used if there were a strong indication of independence between intervals from tests based on the estimated serial correlation coefficients.

variance-time curve

Subroutine VART. This subroutine computes an estimate of the variance-time curve $V(t)$ for the series of events, i.e., the variance of the number of events in an interval of length t, considered as a function of t,

$$V(t) = \text{var}\ \{N_t\}.$$

The computed estimate is for values of t equal to Δ, 2Δ, 4Δ, 8Δ, 12Δ, $\cdots$. These intervals increase in steps of four after the initial steps until the interval becomes greater than $T/4$ (or $t_{(n)}/4$ if $T = 0$).

Another type of estimate of the variance-time curve, described later in this paper, is computed in subroutine COV. The estimate computed by VART is essentially the standard variance estimate combined with a moving average procedure to give greater precision to the estimate. To obtain the greatest precision from this procedure, a rough guide is to choose Δ so that no interval of length Δ in the series contains more than two or three events.

The computational procedure is as follows. Let n_j be the number of $t_{(i)}$'s, $i = 1, 2, \cdots, n$, falling in the interval $((j-1)\Delta, j\Delta]$, i.e., add one to n_j for each $t_{(i)}$ for which $(j-1)\Delta < t_{(i)} \leq j\Delta$.

Table 2 Output of subroutine VART

Δ	2Δ	4Δ	8Δ	12Δ
n_1	n_1+n_2	$n_1+\cdots+n_4$	$n_1+\cdots+n_8$	$n_1+\cdots+n_{12}$
n_2	n_2+n_3	$n_2+\cdots+n_5$	$n_2+\cdots+n_9$	$n_2+\cdots+n_{13}$
n_3	n_3+n_4	$n_3+\cdots+n_6$	$n_3+\cdots+n_{10}$	$n_3+\cdots+n_{14}$
—	—	—	—	—
—	—	—	—	—
—	—	—	—	—
n_{s-1}	$n_{s-1}+n_s$	—	—	—
n_s	—	—	—	—

Here j runs from 1 to s, where s is the largest interger such that $s\Delta \leq T$. Subroutine VART computes and lists the quantities shown in Table 2.

For each column in Table 2, the following quantities are computed and listed:

- Number of entries in the column, A
- Sum of the quantities in the column, $SIGM1(K)$
- Mean for the column, $AMEAN(K) = SIGM1(K)/A$
- Sum of the squares of the entries in the column, $SIGM2(K)$
- Corrected sum of squares, $SIGC0(K) = SIGM2(K) - AMEAN(K) \times SIGM1(K)$
- Multiplier, $AK(K) = 3A/[3A(A - r) + r^2 - 1]$, where r is the multiplier of Δ for the column
- Estimate of the variance of the number of counts in an interval of length $r\Delta$; $AVAR(K) = AK(K) \times SIGC0(K)$
- Normalized variance estimate, $AVAR(K)/AMEAN(K)$
- Sum of the products of successive non-overlapping entries in the column, $SIGMA(K)$

In the case of the last computed quantity, we obtain for 4Δ for example

$$SIGMA\ (K) = (n_1 + \cdots + n_4)(n_5 + \cdots + n_8) + (n_5 + \cdots + n_8)(n_9 + \cdots + n_{12}) + \cdots$$

The first six quantities computed for each column give the successive steps in the standard calculation of a variance estimate, except for the multiplier $AK(K)$. This multiplier is used to compensate for the overlapping of intervals and gives an unbiased estimate for Poisson processes. For the first column the multiplier is equal to $1/(A - 1)$. The seventh column is an estimate of the square of the coefficient of variation of the number of counts in the interval. For a Poisson process, the coefficient of variation has the theoretical value of one for all intervals. The ninth computation may be used for studies of the correlational properties of lag one of the counting process for various interval lengths.

The columns in Table 2, e.g., $n_1, n_2, n_3, \cdots$, may be suppressed and only the nine resultant quantities printed out. However, the contents of the successive columns are useful in looking for and testing cyclic trends in the series.

covariance of counts

Subroutine COV. This subroutine computes estimates of serial covariance and correlation coefficients of various lags for the counts in intervals of length $\Delta, 2\Delta, 4\Delta, \cdots$. Also a variance-time curve estimate is constructed from these estimated correlation coefficients.

Using the quantities in the first column of Table 2 from subroutine VART, i.e., $n_1, n_2, \cdots, n_A$, COV first computes the following quantities for all integer $j(J)$ greater than one and less than $A/2$.

$$CJ = \sum_{i=1}^{A-j} n_i n_{i+j}$$

$$BCJ1 = \frac{CJ}{(A-j)} - \frac{1}{(A-j)^2}\left(\sum_{i=1}^{A-j} n_i\right)\left(\sum_{i=1}^{A-j} n_{i+j}\right)$$

The quantity $BCJ1$ is the standard estimate of a serial covariance of lag $j(J)$. The 1 in $BCJ1$ signifies that the covariances are for counts in intervals of length Δ. The corresponding estimate of the serial correlation coefficient of lag $j(J)$ is formed by dividing $BCJ1$ by the estimated variance for the interval Δ, i.e., the seventh quantity computed for the first column in Table 2 by subroutine VART, and designated $V1$:

$$BRHOJ1 = BCJ1/V1$$

For testing purposes, subroutine COV also computes and lists

$$BRHOJ1\ (A-j)^{1/2}$$

An estimate of serial covariances of lag 1 for intervals of length $L\Delta$ $(l\Delta)$ is formed, by analogy with standard probability relations, as

$$BC1L = \sum_{j=1}^{l} jBCJ1 + \sum_{j=l+1}^{2l} (2l-j)BCJ1$$

An alternative estimate of the variance-time curve is formed for intervals of length $L\Delta(l\Delta)$, for integer L greater than one and less than $A/2$

$$BV1 = V1 \qquad l = 1$$

$$BVL = lV1 + 2\sum_{s=1}^{l-1}\sum_{J=1}^{s} BCJ1$$

The last computations performed are of

$$BRHOIL = BC1L/BVL$$

and

$$DELBVL = BVL - BV(L-1) \qquad BVO = 0$$

The latter quantity is useful in determining the range of values over which the estimated variance-time curve is increasing approximately linearly.

The above quantities are listed by subroutine COV in successive columns and in the order

J or $L, CJ, BCJ1, BRHOJ1, BRHOJ1(A - j)^{1/2}$, $BC1L, BRHOIL, BVL$, and $DELBVL$

covariance density

Subroutine DENS. The second-order joint properties of the counting process N_t can also be investigated by means of a *covariance density* $\gamma_+(t)$ that is related to the variance-time curve by the equation

$$\gamma_+(t) = \frac{V''(t)}{2} = m[m_f(t) - m]$$

In this equation $m = 1/[E(X)]$. The function $m_f(t)$ is known as the *intensity function* or as the *renewal density* in renewal theory. The intensity function is the derivative of the expected number of events in a stationary series observed for an interval of length t starting from an arbitrarily selected event. An estimate of the quantity $m_f(t)$ is calculated in this subroutine. It has the constant value m for a Poisson process.

As with all estimates of density and intensity functions, smoothing is required. Let δ be the smallest interval over which the smoothing is required. The interval δ is specified as an input parameter and should not be much less than $E(X)$. Another input parameter needed is L, where $L\delta$ is the range of t over which it is desired to estimate $m_f(t)$. Subroutine DENS computes the estimates MFJ, for $J = 1, \cdots, L$, where MFJ is the number of sums

$$\sum_{i=p}^{q} x_i$$

for $p \leq q \leq n$, lying in the interval $[(J - 1)\delta, J\delta]$.

The subroutine prints out the MFJ in a column, and in succeeding columns prints out the sum of each successive pair of MFJ's, the sum of each successive set of three MFJ's and the sum of each successive set of four MFJ's. The combined quantities are used in computing estimates of $m_f(t)$ smoothed over intervals of lengths $\delta' = 2\delta, 3\delta, 4\delta$.

When n (or T) is large, a well-behaved, smoothed estimate of $m_f(t)$ is proportional to the MFJ's. Otherwise, a correction for bias is needed and must be hand-computed. Such an estimate of $m_f(t)$, smoothed over an interval δ, is as follows:

$$\bar{m}_f(J\delta + \tfrac{1}{2}\delta) = \frac{T \times MFJ}{n(T - JT - \frac{1}{2}T)\delta} \qquad J = 1, \cdots, L$$

spectrum of counts

Subroutine BART. A third characterization of the second-order joint properties of the counting process N_t is the spectrum of the counting process $g_+(\omega)$. This is the Fourier transform of $\gamma_+(t)$, and may be written as

$$g_+(\omega) = \frac{m}{\pi} + \frac{1}{\pi}\int_0^\infty (e^{-i\omega\tau} + e^{i\omega\tau})\gamma_+(\tau)\, d\tau \qquad \omega \geq 0$$

This function is not periodic, as is the spectral density $f_+(\omega)$ of intervals. Thus, one problem in using $g_+(\omega)$ is to determine the range of interest of ω. The distributional theory of estimates of $g_+(\omega)$ is much simpler than that for the estimates of the variance-time curve and the intensity function. For this reason, it is in many cases the preferable characterization of the second-order joint properties of N_t for use in a statistical analysis. For a Poisson process with parameter λ

$$g_+(\omega) = \frac{m}{\pi} = \frac{\lambda}{\pi} \qquad \omega \geq 0$$

An estimate of $g_+(\omega)$ that has suitable distribution properties is

$$\tilde{g}_+(\omega) = \frac{1}{\pi}\left\{\frac{n}{T} + \frac{2}{T}\sum_{s=1}^{n-1}\sum_{j=1}^{n-s}\cos\{\omega[t_{(s+j)} - t_{(j)}]\}\right\}$$

For observations from a Poisson process, it can be shown that as $T \to \infty$, $\tilde{g}_+(\omega)$ has an exponential distribution with parameter such that $E[\tilde{g}_+(\omega)] = \gamma/\pi$ if $\omega T/(2\pi) = 1, 2, \cdots$. In other words, $\tilde{g}_+(\omega)$ is an unbiased but not consistent estimator of $g_+(\omega)$. For finite T, the exponential distribution is approximately correct. The exact moments are

$$E[\tilde{g}_+(\omega)] = \frac{\lambda}{\pi} \qquad \frac{T\omega}{2\pi} = 1, 2, \cdots$$

$$\text{var}\,[\tilde{g}_+(\omega)] = \frac{\lambda^2}{\pi^2}\left(1 + \frac{1}{\lambda T}\right) \qquad \frac{T\omega}{2\pi} = 1, 2, \cdots$$

$$\text{corr}\,[\tilde{g}_+(\omega_1), \tilde{g}_+(\omega_2)] = \frac{1}{1 + \lambda T} \qquad \begin{cases} \dfrac{T\omega_1}{2\pi} = 1, 2, \cdots \\ \dfrac{T\omega_2}{2\pi} = 1, 2, \cdots \end{cases} \qquad \omega_1 \neq \omega_2$$

It is convenient for statistical analysis to estimate a normalized spectrum, $\pi g_+(\omega)/m$, having the value one for all ω for a Poisson process. The normalized spectrum is the function whose estimate is computed in subroutine BART. The estimate, $I(J)$, is computed as follows:

$$n = \frac{t_{(n)} - t_{(1)}}{n - 1}$$

The subroutine computes

$$A(J) = \sum_{i=2}^{n} \cos\left\{jB\,\frac{[t_{(i)} - t_{(1)}]}{\eta}\right\}$$

$$B(J) = \sum_{i=2}^{n} \sin\left\{jB\,\frac{[t_{(i)} - t_{(1)}]}{\eta}\right\}$$

and

$$I(J) = \frac{2}{(n-1)}\{[A(J)]^2 + [B(J)]^2\}$$

for $j = J = 1, 2, \cdots, P$. Here P and B are input parameters. If B is taken to be $2\pi/(n-1)$ as is normally the case, then j is related to ω by

$$j = \frac{\omega T}{2\pi}$$

Also the integral values of j used in the subroutine give values of ω for which the estimate is well behaved. For observations from a Poisson process and for j with integral and non-zero values, it can be shown that

$$E[I(J)] \approx 1$$

$$\text{var } [I(J)] \approx 1 + \frac{1}{n}$$

$$\text{corr } [I(J_1), I(J_2)] \approx 1 + \frac{1}{n} \qquad J_1 \neq J_2$$

The input parameter P should be greater than n. Then, in most cases, all the salient features of the spectrum will be shown by computations of $I(J)$ for J up to $P = 2n$.

The estimate $I(J)$ has to be smoothed to obtain a consistent estimate of $g_+(\omega)$. A uniform weight scheme is usually adequate. For this reason, the program prints out in adjacent columns J, $I(J)$, the sum of successive sets of two $I(J)$'s, the sum of successive sets of three $I(J)$'s, the sum of successive sets of four $I(J)$'s, and the sum of successive sets of five $I(J)$'s.

An example of an analysis

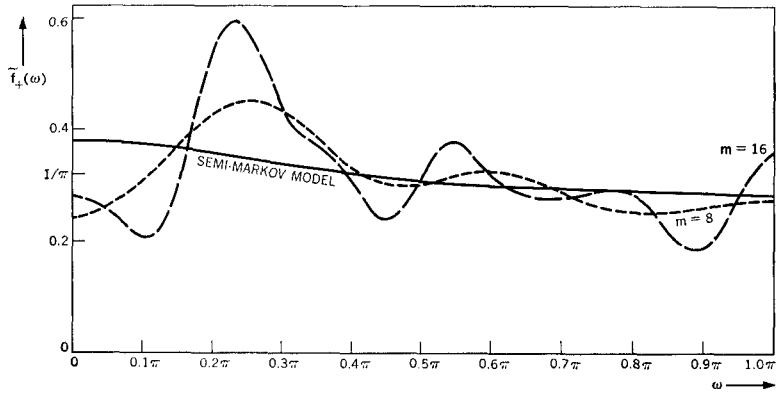
As a demonstration of the preceding methods, consider now the analysis of data given by Bartlett. The events are times at which automobiles passed a point on a road in Sweden. Statistical analysis of traffic data is of interest because of the current activity in traffic queuing and interference problems. Analytical solutions to these problems use models having specific assumptions concerning the probability structure of the sequence of times at which cars pass a point on a road. It is common, for instance, to assume that this sequence constitutes a Poisson process, and the solutions obtained depend on the validity of this assumption. One would clearly prefer to make assumptions that are empirically valid under given circumstances. The SASE program is designed to help obtain such an empirical validation.

The data consist of $n = 128$ events recorded during an interval T of length 2039.3 seconds. The following computations on this sample were made using subroutine INTER:

Estimated mean $\tilde{\mu} = 15.81$ seconds
Estimated coefficient of variation $C = \tilde{\sigma}/\tilde{\mu} = 1.50$
Estimated coefficient of skewness $\tilde{\gamma}_1 = 2.54$

A plot of the cumulative number of events against time shows a tendency for the events to cluster. The clustering shows up

Figure 2 Estimated and fitted spectra of intervals for traffic data



inconsistent with a renewal model. For instance, $\tilde{\rho}_1 = +0.092$, and multiplying by $(n-1)^{1/2}$ we obtain

$$\tilde{\rho}_1(n-1)^{1/2} = +1.04$$

This statistic is tested as a normal random variable with mean zero and standard deviation one. The upper two-sided five percent point of this distribution is 1.96, so that the value $+1.04$ is not significantly large. Testing the values *PERIO* computed in subroutine SPEC as a Poisson process by putting them through EXPO, we obtain the values

$$KS = 1.11 \qquad WN2 = 2.01$$

These values correspond to levels of seventeen and nine percent respectively, so that again there is no rejection of the renewal hypothesis.

Figure 2 shows, as dashed curves, two smoothed spectral estimates obtained from subroutine SPEC with lag windows λ_i in which the parameter m is 8 and 16. The curve with the high peak is the estimate with $m = 16$. Since the theoretical spectrum for a renewal process is constant with value $1/\pi$, the tests of the renewal hypothesis indicate that all of the deviation from $1/\pi$ of the estimated spectra is due to sample fluctuations. In particular, there is no evidence for a peak in the underlying spectrum in the vicinity of $\omega = 0.23\pi$.

Despite the acceptance of the renewal model, it is instructive to try to fit the two-state semi-Markov model to the data. The density of the time-between-events in this model is

$$f_X(x) = \pi_1 f_1(x) + \pi_2 f_2(x)$$

where $\pi_1 = (1-\alpha_2)/(2-\alpha_1-\alpha_2)$ and $\pi_2 = 1-\pi_1$. The previous discussion of models suggests taking $f_1(x)$ to be exponential, and $f_2(x)$ to be a gamma distribution with $a > 1$. Thus

$$f_X(x) = \pi_1\left(\frac{1}{\mu_1}\right)e^{-x/\mu_1} + \pi_2\left(\frac{a}{\mu_2}\right)^a \frac{x^{a-1}e^{-ax/\mu_2}}{\Gamma(a)}$$

The serial correlation coefficients for this model are given by

$$\text{corr}\,(X_i, X_{i+j}) = \frac{(\mu_1 - \mu_2)^2\pi_1\pi_2}{\pi_1\sigma_1^2 + \pi_2\sigma_2^2 + \pi_1\pi_2(\mu_1 - \mu_2)^2}(\alpha_1 + \alpha_2 - 1)^j$$

$$= c\beta^j$$

The corresponding spectral density is

$$f + (\omega) = \frac{1}{\pi}\left[\frac{1 + (1 - 2c)\beta^2 - 2\beta(1 - c)\cos\omega}{1 + \beta^2 - 2\beta\cos\omega}\right]$$

A convenient way to obtain initial estimates of the parameters μ_1, μ_2, a, α_1 and α_2 is by the method of moments. We equate $E(X)$, $E(X^2)$, $E(X^3)$ and $\text{cov}(X_i, X_{i+1})$ to their sample values. Another equation could be used, but it is simpler to solve the four equations for different values of a. It is found that with $a = 3$ the estimates

$$\tilde{\mu}_1 = 27.15 \qquad \tilde{\mu}_2 = 2.91 \qquad \tilde{\alpha}_1 = 0.652 \qquad \tilde{\alpha}_2 = 0.607$$

give a good fit to the estimated marginal distribution of intervals for the data, as obtained from subroutine INTER. The logarithm of the estimated survivor function of intervals between events is shown as a series of crosses in Figure 3. The recorded number of intervals with values between given limits are shown parallel to the time axis. The theoretical log survivor function for a semi-Markov model is shown as the solid line. The first type of interval for the fitted log survivor function has an exponential distribution function, whereas the second type of interval has a gamma distribution function with index $a = 3$. There is evidently a good fit of the data to the model survivor function.

Figure 3 Empirical and fitted log survivor functions for traffic data

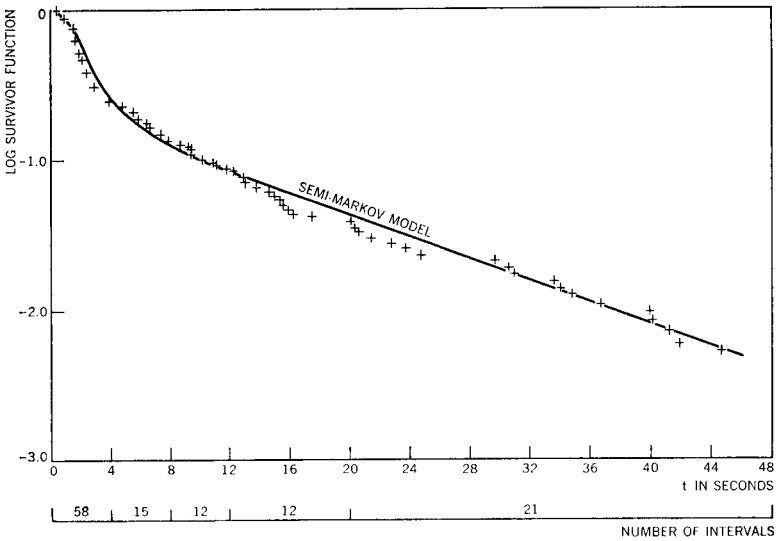
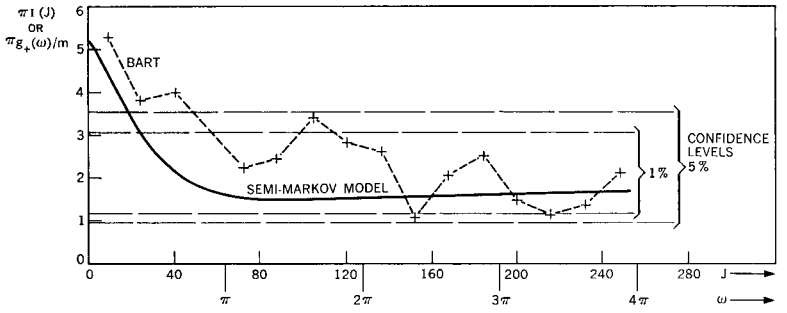


Figure 4 Estimate of normalized spectrum of counts for traffic data



The solid line in Figure 3 is the spectrum of intervals, $f_+(\omega)$, for the semi-Markov model, using the estimated parameter values. Note that it is close to the constant spectrum for a renewal model. In fact, while one would not use any but a renewal model on the basis of this data, the semi-Markov model fits the data well. The main import of this example might be that rather long series are needed to differentiate among models.

The estimated spectrum of counts (using 16-point uniform weighting) is shown in Figure 4. The bands are five and one percent confidence levels for individual estimates under the assumption that the series is a Poisson process. The fact that the initial values of the estimated spectrum fall outside these lines gives another indication of the departure from the Poisson process. The solid line in Figure 4 is the fitted spectrum for the semi-Markov model; the fitted spectrum for a renewal model is virtually indistinguishable from it.

ACKNOWLEDGMENTS

The initial programming of this problem was done by Mr. A. deKorvin and Mr. C. F. Corley of the IBM Systems Development Division. The Bay Area Scientific Computing Center of the Service Bureau Corporation, Palo Alto, California, provided the assistance of Mrs. R. S. McDuffie on parts of the program. The later version of the program, SASE II, was written by Mr. T. C. Kelly and Dr. J. M. Miller of the IBM Research Division.

CITED REFERENCES AND FOOTNOTES

1. P. A. W. Lewis, "A branching Poisson process model for the analysis of computer failure patterns," *Journal of the Royal Statistical Society, Series B* **26**, 398–456 (1964).
2. G. P. Moore, D. H. Perkel, and J. P. Segundo, "Statistical analysis and functional interpretation of neural spike data," *Annual Review of Physiology* **28**, 493–522 (1966).
3. J. M. Berger and B. Mandelbrot, "A new model of error clustering on telephone circuits," *IBM Journal of Research and Development* **7**, No. 3, 224–236 (1963).

4. P. A. W. Lewis and D. R. Cox, "A statistical analysis of telephone circuit error data," *IEEE Transactions on Communication Technology* **COM-14**, No. 4, 382–389 (1966).
5. SASE I, an experimental program written for the IBM 7094, is available through the author. SASE II, an improved version also written for the IBM 7094, has been submitted to SHARE. A version for the SYSTEM/360 has also been submitted to SHARE.
6. D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events*, Methuen & Co., Ltd., London: John Wiley & Sons, Inc., New York (1966).
7. P. A. W. Lewis, "Some results on tests for Poisson processes," *Biometrika* **52**, 67–77 (1965).
8. M. S. Bartlett, "The spectral analysis of point processes," *Journal of the Royal Statistical Society, Series B* **25**, 264–296 (1963).
9. J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation* **19**, No. 90, 297–301 (April 1965).