INPUT MODELING: ESTIMATION USING EVENT COUNT DATA

Ronald Dattero

Texas A & M University
College Station, Texas

Bruce W. Schmeiser

Purdue University
West Lafayette, Indiana

Abstract

In modeling a system, identifying the times at which the state of the system changes is important. In particular, the probability distributions of the interevent times (time lengths between consecutive events) of each process (such as interarrival times) are required in a simulation model. Often the only available data for a particular process is the number of events per unit time (count data) rather than the event times. This paper surveys estimation of the interevent time distribution from count data.

INTRODUCTION

Often information characterizing the interevent times of a discrete event simulation model may be unavailable or difficult to obtain. On the other hand, the number of events per unit time are simple and economical to collect. Furthermore, the only information available in many situations consists of count data. Two common examples are accident occurrence and traffic flow. In the first example, accident statistics are given as a specific number per hour, per day, or per year. In the second example, the time between consecutive vehicles passing a given point is more difficult to measure than count data. In addition, the statistic which is usually given by traffic authorities is the number of vehicles passing a given point in a fixed time period.

Count data, however, provides less information about the process than event time data. That is, there is more information in the statement that n events occurred at times $t_1,\ldots,t_n$, respectively, than in the statement that n events occurred by epoch t. On the other hand, the actual counting process and the associated count data process agree at the points when the counts are recorded. Furthermore, as the sampling interval lengths decrease to zero, the associated count data process approaches the actual counting process in terms of the information provided by the data.

In the last few years, there has been much research in stochastic modeling using point processes by a varied group of researchers (see (1, 2, 3, 4)). Despite this extensive and diverse research effort, little attention has been given to estimation techniques based on event count data rather than event time data. In the last two years, however, some new estimation techniques have been developed. This paper surveys estimation based on event count data.

General Framework

Fitting a model from count data can be described by the following four step approach:

1. identifying an appropriate model for the process;
2. designing the method of data acquisition;
3. estimating the parameters of the model from the data; and
4. testing the adequacy of the model.

As a first step in detailing the four step approach, we developed a flowchart (Figure 1) that represents a logical decision making approach to model selection, data collection, model fitting, and model testing. Decision boxes with an "M" above them are modeler decisions; for instance, does the modeler believe the process is stationary.

Proceedings of the 1982
Winter Simulation Conference
Highland * Chao * Madrigal, Editors

82CH1844-0/82/0000-0299 $00.75 © 1982 IEEE

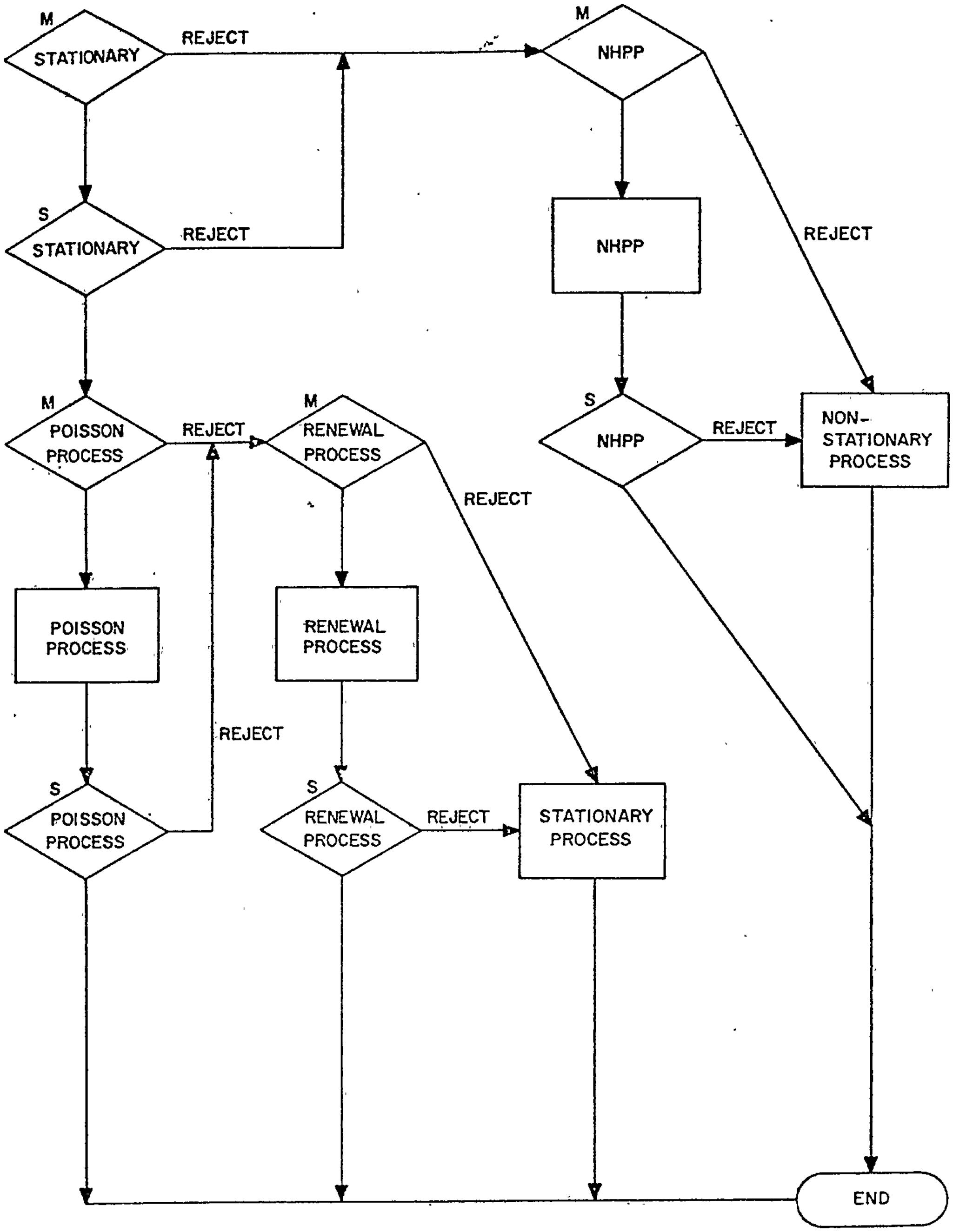

Figure 1. Modeling Approach

Rather than a yes/no response, the modeler decides to not reject/reject the assumption. Decision boxes with an "S" above them are statistical tests of hypothesis. Process boxes perform the data collection and model fitting task (NHPP denotes Non-Homogeneous Poisson Process). This flowchart will be detailed in the remainder of the paper.

Statistical Tests for Stationarity

Statistical tests for stationarity are rather limited when the data is count data. Most of the tests are for trend. In this section, three tests for trend are discussed that are applicable to count data. To employ these tests, the counting intervals must be of equal length τ.

The Cox-Stuart (5) test for trend is a modification of the sign test. For each observed replication of the process, one of the early observations (number of events in a particular interval), the j^{th} observation, is paired with one of the later observations, the $(j+c)^{th}$ observation where $j=1,\ldots,\nu-c$, $c = \text{integer } [(\nu+1)/2]$, and ν is the number of observations (intervals) per replication. The test statistic is based upon the number of times the earlier observation is greater than the later observation. This test, however, assumes that the observations are independent.

Alternative tests for trend involve Kendall's and Spearman's rank correlation coefficients. The (T_i, A_i) pairs, where $T_i = i\tau$ and the number of events in the i^{th} interval is denoted by A_i, are a sample of bivariate observations in which each variable can be ordered. Employing Kendall's rank correlation coefficient as a nonparametric test for trend is discussed fully in (6). Rank correlation methods are discussed thoroughly in (7).

Fitting the Poisson Process

One of the most widely used assumptions is that a process can be modeled quite well as a Poisson process. This assumption arises because many analytic models, such as queueing network models, need this assumption. Of equal importance, numerous processes exhibit behavior that can be modelled quite well by the Poisson process (8). In particular, the Poisson process is appropriate when modeling light traffic flow on a road at a point a large distance beyond the road entry point with unrestricted overtaking (9, 10, 11, 12). An extensive list of other applications is given in (13).

To estimate the unknown parameter λ of a Poisson process from event count data, the unique minimum variance unbiased estimator of λ, total number of events divided by the total period of observation, is usually employed. That is, if we have count data for disjoint intervals of length $t_1, t_2, \ldots$, the statistic is $\hat{\lambda} = \Sigma n_i / \Sigma t_i$ where n_i is the number of events observed in the i^{th} interval. The variance of this estimator is $\lambda / \Sigma t_i$.

Statistically Testing the Poisson Assumption

A standard test for the hypothesis that the n_i's are Poisson observations is the dispersion test for homogeneity based on the statistic:

$$d = \sum_{i=1}^{k} \frac{(n_i - \bar{n}_i)^2}{\bar{n}_i}$$

where

$$\bar{n}_i = \lambda t_i \quad .$$

A chi-squared distribution with k-1 degrees of freedom is a good approximation to the null-hypothesis distribution of d (14).

If $t = t_1 = \ldots = t_k$, alternative tests exist. In this case, the data consists of k independent observations of a Poisson (λt) random variable. The statistic d, when divided by k-1, is the ratio of the estimated variance to the estimated mean, which is an estimate of the index of dispersion. The alternative test is a comparison to determine whether the estimated index of dispersion differs from its value of unity under the null hypothesis of a Poisson process (1). Another test is the chi-square goodness-of-fit test with the null hypothesis being the sample was drawn from a Poisson population.

Fitting a Renewal Process

If the Poisson process is an inadequate model, a renewal process might be appropriate. Depending upon the type of data (number of replications, length of the observation period per replication, length of the sampling intervals), different approaches can be taken. These different approaches will be outlined in this section.

When the data consists (primarily) of a sequence of zeros and ones, an approach developed by Kimbler (15) or Dattero (16) can be taken.

In his doctoral dissertation, Kimbler (15) developed an estimator for the interevent time distribution of a renewal process from event count data. First the count data record of the process is used as an estimate of the renewal function for an ordinary renewal process. Next the estimated renewal function is approximated by a polynomial of degree p, $\hat{R}(t)$, by solving a set of simultaneous linear equations consisting of:

1. $\hat{R}(0) = 1$ (an event is counted at zero);
2. $\hat{R}(T)$ equals the number of events recorded until epoch T where T is the length of the observation period; and
3. first-order differential equations at selected time epochs.

The polynomial approximation coefficients are then used to form the LaPlace transform of $\hat{F}^*(s)$, the estimate of $F^*(s)$, which is inverted by Heaviside methods to yield $\hat{F}(t)$, the estimate of $F(t)$.

In Kimbler's implementation, three types of numerical inconsistencies can occur. They are easily corrected as follows. If any of the approximations of the first derivative of the renewal function are negative, they are set to zero. If the piece-wise distribution function decreases at any point t_0 $(0 < t_0 \leq T)$, $\hat{F}(t_0)$ is set to $\hat{F}(t_1) + \epsilon$ where t_1 is the last point where the distribution function is non-decreasing and ϵ is a small constant. If $\hat{F}(T) > 1$, the entire distribution function is normalized by dividing it by $\hat{F}(T)$.

Dattero (16) developed an estimation procedure based on the following relationship between the forward recurrence time density function, $g(x)$, and the interevent time distribution, $F(x)$: $F(x)=1- \mu g(x)$. The basic idea is to estimate μ and $g(x)$ from only event count data. The mean interevent time, μ , is estimated by the length of the observation period divided by the number of events observed. To get an estimate of $g(x)$, the associated survivor function is estimated at, $\tau, 2\tau, \ldots, K\tau$. Then an estimate of $g(x)$ is acquired by fitting a second degree Newton divided-difference polynomial that passes through three points and taking the derivative of the polynomial approximation at $x= \tau, \ldots, (K-1)\tau$. An estimate of $F(x)$, $\tilde{F}(x)$, (at $x= \tau, \ldots, (K-1)\tau$) is acquired from the basic relationship with the additional constraint that $\tilde{F}(x)$ be a monotone non-decreasing function between 0 and 1 inclusively. This function is then linearly interpolated. Note that alternative methods of implementation are possible (see (16)).

The major short-coming of this approach is the bias of $\tilde{F}(x)$. Most of the bias usually occurs during the (linear) interpolation stage. While this interpolation scheme may be simple and provide a good approximation for some distributions, it is probably not the shape of the true interevent time distribution. As τ decreases, however, this approximation improves and the bias usually decreases.

Through Monte Carlo experimentation, the two approaches were compared (see (16)). The results strongly indicate that Dattero's approach performs better than Kimbler's. The major reason for the better performance is that Dattero's approach uses more of the information provided by the count data.

When the length of the observation period is long and the number of replications is large, limit theorem results can be employed. The approach consists of estimating the moments of the counting distribution for large t (time displacement from origin under synchronous counting) and then employing Smith's (17) result which relates the moments of the counting distribution to the moments of the interevent time distribution for a renewal process. It should be noted that n+1 moments of the interevent distribution are required to compute n moments of the counting distribution; however, only n moments are required when using only the asymptotic terms. Using only the asymptotic terms, the method proceeds as follows. Suppose for some specified large t, estimates of the first three cumulants of the counting process at t ($\hat{\kappa}_1(t)$, $\hat{\kappa}_2(t)$, and $\hat{\kappa}_3(t)$) are supplied. This produces the following estimators for the first three cumulants of the interevent distribution ($\hat{\kappa}_1$, $\hat{\kappa}_2$, and $\hat{\kappa}_3$):

$$\hat{\kappa}_1 = \frac{t}{\hat{\kappa}_1(t)}$$

$$\hat{\kappa}_2 = \frac{t_2 \quad \hat{\kappa}_2(t)}{[\ \hat{\kappa}_1(t)]^3}$$

and

$$\hat{\kappa}_3 = \frac{t^3}{[\ \hat{\kappa}_1(t)]^4} \left(\frac{3[\ \hat{\kappa}_2(t)]^2}{[\ \hat{\kappa}_1(t)]} - \hat{\kappa}_3(t) \right) .$$

In some cases, using only the α_i's provides sufficient accuracy. When t is not quite large enough, the β terms should also be employed. In these cases, by assuming a specific distributional form, one parametric equation can be added by using a parametric form for the $(n+1)^{st}$ moment of the interevent distribution (in terms of the first n moments).

A form of the interevent time distribution must be assumed in order to provide a constructive definition of the process. One general approach would assume some specific general distribution family. Here the lower order moments would be used to fit the parameters of the distribution while the higher order moments would be used to check the distribution assumption.

When neither of the two renewal process modeling situations (count data consisting (primarily) of a sequence of zeros and ones or the length of the observation period being long and the number of replications is large) occur, various ad hoc approaches can be taken. One basic approach makes assumptions that turns the count data into event time data and then tests the assumptions. One simple assumption is that the first event in the i^{th} interval occurred at $1/(a_i+1)$ where $a_i > 0$ is the number of events occurring during the i^{th} interval. From this "data", an estimate of the forward recurrence time distribution can be constructed. The most important step in researching this and other ad hoc approaches is testing the adequacy of the model. In addition, the sources and measures of the modeling errors must be examined thoroughly.

Statistically Testing the Renewal Process Assumption

As far as statistically testing the renewal process assumption, no tests are available based on count data. Development of these tests would be a valuable research contribution.

Stationary Point Processes Estimation Procedures

One generalization of a renewal process is a stationary point process. The basic relationship used by Dattero (14), $F(x)=1-\mu\ g(x)$, holds not only for renewal processes but also for the marginal distribution of X_i, $F(x)$, in the stationary sequence $\{X_i\}$ for a stationary point process. In addition, the covariance structure of the X_i's are defined by the relationship:

$$\int_0^\infty q(j;t)dt = E(X) + \frac{R(j)}{E(X)} \qquad (j=1,2,...)$$

where $q(j;t) = P\{N(t)=j\}$ and $R(j)$ is the lag j covariance (18). From count data, $q(j;t)$ can be estimated for $t= \tau, 2\tau, ..., K\tau$ by an approach outlined in (16). Given these estimates, the integral can be numerically integrated. Despite having estimates of $F(x)$ and the covariance structure, providing a constructive definition of a counting process with these properties is difficult since little work has been done on autoregressive point processes. Some recent work (19, 20, 21, 22, 23, 24, 25, 26) holds some promise for this area.

Use of the Non-homogeneous Poisson Process

If stationarity is rejected, numerous point process models are still available; however, little published work on fitting and testing these models from count data exists. Use of the non-homogeneous Poisson process though, seems promising.

A non-homogeneous Poisson process has independent increments, orderliness, and the number of events in an interval is Poisson distributed with the mean depending on the location of the interval. An estimator for the mean number of events for the j^{th} interval is:

$$\hat{\lambda}_j = \sum_{i=1}^{M} \frac{a_j(i)}{M} \qquad j=1,...,\nu$$

where $a_j(i)$ is the number of events during the j^{th} interval for the i^{th} replication, M is the number of replications of the process observed, and ν is the number of intervals. To get the estimated rate of the process during the j^{th} interval, the user supplies a function $r_j(t)$ that describes his beliefs on how the process behaves over the interval. Then the estimated rate of the process for the j^{th} interval is:

$$\hat{\lambda}_j\ r_j(t)/\int r_j(t)dt$$

where the range of the integral is the j^{th} interval.

A standard test for the hypothesis that the $a_j(i)$'s (j fixed) are Poisson observations is the dispersion test for homogeneity. Another test is the chi-square goodness-of-fit test with the null hypothesis being the sample was drawn from a Poisson population. In addition, the $a_j(i)$ and $a_k(i)$ ($j\neq k$) observations can be tested for independence (using Kendall's or Spearman's rank correlation coefficient).

BIBLIOGRAPHY

1. Cox, D.R., and Lewis, P.A.W., The Statistical Analysis of Series of Events, Metheun, London, 1966.

2. Lewis, P.A.W. (ed.), Stochastic Point Processes: Statistical Analysis, Theory, and Applications, John Wiley & Sons, New York, 1972.

3. Brillinger, D.R., "Comparative Aspects of the Study of Ordinary Time Series and of Point Processes", Developments in Statistics 1, (P.R. Krishnaiah, Ed.), Academic Press, New York, 1978.

4. Cox, D.R., and Isham, V., Point Processes, Chapman and Hall, London, 1980.

5. Cox, D.R., and Stuart, A., "Some Quick Tests for Trend in Location and Dispersion", Biometrika, 1955.

6. Ferguson, G.A., Nonparametric Trend Analysis, McGill University Press, Montreal, 1965.

7. Kendall, M.G., Rank Correlation Methods (4th edition), Griffin, London, 1970.

8. Kleinrock, L., Queueing Systems Volume 1: Theory, John Wiley & Sons, New York, 1975.

9. Weiss, G.H., and Herman, R., "Statistical Properties of Low Density Traffic", Quarterly of Applied Mathematics, 1962.

10. Breiman, L., "The Poisson Tendency in Traffic Distribution", Annals of Mathematical Statistics, 1963.

Count Data (Continued)

11. Thedeen, T., "A Note on the Poisson Tendency in Traffic", Annals of Mathematical Statistics, 1964.

12. Brown, M., "Low Density Traffic Streams", Advanced in Applied Probability, 1972.

13. Haight, F.A., Handbook of the Poisson Distribution, John Wiley & Sons, New York, 1967.

14. Rao, C.R., and Chakravarti, I.M., "Some Small Sample Tests of Significance for a Poisson Distribution", Biometrics, 1956.

15. Kimbler, D.L., "Approximation of the Interevent Time Distribution Using Empirical Event Count Data", Ph.D. Thesis, Virginia Polytechnic Institute and State University (Industrial Engineering and Operations Research), 1980.

16. Dattero, R., "Stochastic Models from Event Count Data", Ph.D. Thesis, Purdue University, West Lafayette, Indiana, 1982.

17. Smith, W.L., "On the Cumulants of Renewal Processes", Biometrika, 1959.

18. McFadden, J.A., "On the Lengths of Intervals in a Stationary Point Process", Journal of the Royal Statistical Society B, 1962.

19. Lawrence, A.J., "Some Models for Stationary Series of Univariate Events", Stochastic Point Processes: Statistical Analysis, Theory, and Applications, (P.A.W. Lewis, Editor), John Wiley & Sons, New York, 1972.

20. Jacobs, P.A., and Lewis, P.A.W., "A Mixed Autoregressive-Moving Average Exponential Sequence and Point Process (EARMA 1,1)", Advances in Applied Probability, 1977.

21. Lawrance, A.J., and Lewis, P.A.W., "An Exponential Moving-Average Sequence and Point Process (EMA 1)", Journal of Applied Probability, 1977.

22. Lawrance, A.J., and Lewis, P.A.W., "An Exponential Autoregressive-Moving Average Process EARMA (p,q): Definition and Correlation Properties", Technical Report, Department of Operations Research, Naval Postgraduate School, 1978.

23. Lawrance, A.J., and Lewis, P.A.W., "Simulation of Some Autoregressive Markovian Sequences of Positive Random Variables", Proceedings of the Winter Simulation Conference, 1979.

24. Gaver, D.P., and Lewis, P.A.W., "First Order Autoregressive Gamma Sequences and Point Processes", Advances in Applied Probability, 1980.

25. Lewis, P.A.W., "Simple Multivariate Time Series for Simulations of Complex Systems", Proceedings of the Winter Simulation Conference, 1981.

26. Schmeiser, B.W., and Lal, R., "Bivariate Gamma Random Vectors", Operations Research, 1982.

ACKNOWLEDGMENT

This paper was supported by the Statistics and Probability Program of the Office of Naval Research via Contract N00014-79-C-1082.