THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

ESTIMATING THE PARAMETER k OF THE RAYLEIGH DISTRIBUTION FROM CENSORED SAMPLES

by

Dwight B. Brock

Technical Report No. 29
Department of Statistics THEMIS Contract

April 16, 1969

Research sponsored by the Office of Naval Research
Contract N00014-68-A-0515
Project NR 042-260

DEPARTMENT OF STATISTICS
Southern Methodist University

ESTIMATING THE PARAMETER k OF THE RAYLEIGH DISTRIBUTION FROM CENSORED SAMPLES

A Thesis Presented to the Faculty of the Graduate School

of

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Master of Science

with a

Major in Statistics

by

Dwight Brandon Brock
(B.S., Southern Methodist University, 1967)

April 15, 1969

Brock, Dwight Brandon

B.S., Southern Methodist University, 1967

Estimating the Parameter k of the Rayleigh Distribution from Censored Samples

Adviser: Associate Professor C. H. Kapadia

Master of Science degree conferred May 25, 1969

Thesis completed April 15, 1969

Let g(x) be the ratio of the ordinate and the probability integral for the Rayleigh distribution. That is, $g(x) = f(x)/F(x)$, where $f(x) = (2x/k)\exp\{-x^2/k\}$, $x > 0$, $k > 0$, and $F(x) = \int_0^x f(t)dt$. Tiku's local approximation $g(x) \simeq \alpha + \beta x/\sqrt{k}$ is used to simplify the maximum likelihood equation for estimating k from a doubly censored sample from this population. The solution to the simplified maximum likelihood equation is the estimator for k , which is called k_c . It is much easier to compute than the maximum likelihood estimator, since no iterative procedure is required.

After the solution for k_c is given, equations are developed for its bias and variance. Numerical comparisons are made among k_c and other estimators for k .

ACKNOWLEDGMENTS

I wish to express my appreciation to my adviser Associate Professor C. H. Kapadia for his suggesting the topic of this paper and for his helpful comments during its preparation. I would also like to thank Professors D. B. Owen, J. T. Webster and R. K. Williams for their aid and advice during my studies. Special acknowledgment is due Professor Paul D. Minton for his helping me obtain financial aid through the National Defense Education Act Title IV Fellowship under which this work was done.

I am also indebted to Mrs. Linda White for her excellent job of typing the manuscript, and to my wife Mary for her support during the preparation of the paper.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

In conducting experiments of varying natures, the experimenter is often confronted with censored data, data lacking one or more observations at the extremes. Censoring may be the result of several naturally-occurring or experiment-imposed conditions. For example, suppose the experimenter does not know the exact magnitude of some of the readings, except that they lie at one or both extremes of the sample. Or suppose a measuring instrument fails at a known point in time in such a way that the observations following the instrument failure are unreliable. These are naturally occurring situations which force censoring of the data. Other examples of censoring are ones when a manufacturer curtails sampling in the case of destructive testing and when a biologist waits for the death of a certain number of individuals before he begins making observations. These are examples in which the censoring is controlled by the experiment. Thus for many reasons it is often necessary to analyze censored data.

This paper is concerned with estimating the parameter k of the Rayleigh distribution from censored samples. The Rayleigh distribution, sometimes called the Rayleigh amplitude distribution, arises as a consequence of finding the resultant amplitude of several coplanar random amplitude vectors which are normally distributed, a fact which Siddiqui [6] has pointed out. However, he prefers to limit his discussion to the power distribution, which is the square of the amplitude distribution. In the case of censored data, though, it is easier to work with the amplitude distribution. This problem might occur in the analysis of acoustic data or other data obtained from measurements of amplitudes of electromagnetic waves received through a scattering medium.

The form of the Rayleigh distribution is as follows:

$$f(x) = \begin{cases} (2x/k) \exp\{- x^2/k\} & , \quad 0 < x < \infty \\ 0 & , \quad \text{elsewhere} \end{cases} \tag{1.1}$$

for positive values of k . Since the expected value of x

$$\begin{aligned} E(X) &= \int_0^\infty xf(x)dx \\ &= \frac{1}{2}\sqrt{k\pi} \end{aligned} \tag{1.2}$$

and the variance of X

$$\mathrm{Var}(X) = \int_0^\infty x^2 f(x)dx - \left[\int_0^\infty xf(x)dx\right]^2 = k[1 - \pi/4] \tag{1.3}$$

are both functions of k , the importance of estimating k is obvious. The results given here are similar to those obtained by Tiku [7, 8, 9] for the normal, exponential,and logistic distributions. His method follows the line of reasoning given below.

Let g(x) = f(x)/F(x) , where f(x) is the ordinate in (1.1) and F(x) is the probability integral of X . Then over a small interval, $a \leq x \leq b$, g(x) lies very close to the line $\alpha + \beta x/\sqrt{k}$, where α and β are constants such that

$$\beta = \{g(b) - g(a)\}/(b - a) \quad \text{and} \tag{1.4}$$

$$\alpha = g(a) - a\beta \quad . \tag{1.5}$$

Empirical justification for these statements is given in the Appendix, along with tables of α and β for various sample sizes. The substitution of $\alpha + \beta x/\sqrt{k}$ for $g(x)$ in the likelihood equation results in a solution which is easy to compute, in that it requires no iteration, and the estimator k_c derived from that solution has the desirable properties of the maximum likelihood estimator.

Tiku's [7, 8, 9] results have been seen to compare favorably with the actual maximum likelihood estimate (computed by iteration), the best linear unbiased estimates (computed by least squares as in Lloyd [3]), and with estimates computed by the method of moments. In this paper similar comparisons will be made with population values, the best linear unbiased estimates, and the moment estimates.

In Chapter II the maximum likelihood equation will be set up and solved for k_c , the desired estimator, by making the substitution $\alpha + \beta x/\sqrt{k}$ for $g(x)$ in that equation.

In Chapter III properties of the estimator will be discussed. Included will be a calculation of the expected value of k_c and its variance as well as a discussion of its efficiency and asymptotic properties.

A numerical example is worked in Chapter IV, and comparisons are made among k_c , the moment estimator, the least squares estimator, and the population values.

CHAPTER II

DERIVATION OF THE ESTIMATOR k_c

Let the random variable X be distributed according to the Rayleigh distribution with parameter k . Let $x_1 , x_2 , \cdots , x_n$ be a random sample from the Rayleigh population with the smallest $r_1 = q_1 n$ and the largest $r_2 = q_2 n$ observations being censored, where q_1 and q_2 are fixed. The remaining sample values, arranged in order of magnitude, are

$$Y_{r_1+1} , Y_{r_1+2} , \cdots , Y_{n-r_2-1} , Y_{n-r_2} .$$

The joint density of these order statistics is well known (see, for example Saw [5]) to be

$$f(Z_{r_1+1} , \cdots , Z_{n-r_2}) = \frac{n!}{r_1!r_2!}(2/\sqrt{k})^{n-r_1-r_2}\left[\prod_{i=r_1+1}^{n-r_2} Z_i\right]\exp\left\{-\sum_{i=r_1+1}^{n-r_2} Z_i^2\right\} \tag{2.1}$$

$$\times \left\{F(Z_{r_1+1})\right\}^{r_1}\left\{1 - F(Z_{n-r_2})\right\}^{r_2}$$

where $Z_i = \frac{Y_i}{\sqrt{k}}$, and F is the probability integral of X . Then

$L = \log f$

$$= \log\left(\frac{n!}{r_1!r_2!}\right) + (n - r_1 - r_2)\log(2/\sqrt{k}) + \sum_{i=r_1+1}^{n-r_2} \log(Z_i) - \sum_{i=r_1+1}^{n-r_2} Z_i^2 \tag{2.2}$$

$$+ r_1\log\left\{F(Z_{r_1+1})\right\} - r_2 Z_{n-r_2}^2 .$$

Now, taking the partial derivative of L with respect to k we obtain the following:

$$\frac{\partial L}{\partial k} = (n - r_1 - r_2)(\sqrt{k}/2)(-1/k\sqrt{k}) + \sum_{i=r_1+1}^{n-r_2} \left(\frac{1}{Z_i}\right)\left(-\frac{1}{2}\cdot\frac{Z_i}{k}\right) - \left(-\frac{1}{k}\right)\sum_{i=r_1+1}^{n-r_2} Z_i^2$$

$$- r_1\left[(1/k)Z^2_{r_1+1}\exp\{-Z^2_{r_1+1}\}\right] \Big/ \left[1 - \exp(-Z^2_{r_1+1})\right] + \frac{r_2 Z^2_{n-r_2}}{k} \quad . \tag{2.3}$$

This equation simplifies to

$$\frac{\partial L}{\partial k} = \left[\frac{r_1 + r_2 - n}{k}\right] + \frac{1}{k}\sum_{i=r_1+1}^{n-r_2} Z_i^2 - \left[\frac{r_1 Z_{r_1+1}}{2\sqrt{k}}\right]\left[\frac{f(Z_{r_1+1})}{F(Z_{r_1+1})}\right] + \frac{r_2 Z^2_{n-r_2}}{k} \quad . \tag{2.4}$$

Setting $\frac{\partial L}{\partial k}$ equal to zero in equation (2.4) and solving for k would give the ordinary maximum likelihood estimator $\hat{k}$ for k . However, this is a difficult problem to solve, due to the fact that the term

$$g(Z_{r_1+1}) = \frac{f(Z_{r_1+1})}{F(Z_{r_1+1})} \tag{2.5}$$

is implicit in k, and thus (2.4) cannot be solved exactly except when censoring is on the right only. The only way to solve it as it stands is by some iterative procedure, which could be expensive and time consuming. Instead of this, it is here proposed that $g(Z_{r_1+1})$ be replaced by a linear approximation

$$g(Z_{r_1+1}) \simeq \alpha + \beta Z_{r_1+1} \quad . \tag{2.6}$$

Consider now

$$\frac{\partial L}{\partial k} \simeq \frac{\partial L'}{\partial k} = \left[(r_1 + r_2 - n)/k\right] + \frac{1}{k} \sum_{i=r_1+1}^{n-r_2} Z_i^2$$

$$- \left[\frac{r_1 Z_{r_1+1}}{2\sqrt{k}}\right]\left[\alpha + \beta Z_{r_1+1}\right] + \frac{r_2 Z_{n-r_2}^2}{k} \quad . \tag{2.7}$$

In this equation α and β are such that

$$\beta = \{g(h_2) - g(h_1)\}/(h_2 - h_1) \quad \text{and}$$

$$\alpha = g(h_1) - h_1 \beta \tag{2.8}$$

where the interval (h_1 , h_2) is wide enough to cover Z_{r_1+1} . This is accomplished for large enough $n - r_1 - r_2$ by choosing h_1 so that

$$F(h_1) = q_1 - \sqrt{\frac{1}{n} q_1 (1 - q_1)} \tag{2.9}$$

and h_2 so that

$$F(h_2) = q_1 + \sqrt{\frac{1}{n} q_1 (1 - q_1)} \quad . \tag{2.10}$$

The reasoning behind this choice of interval endpoints is that it is logical to think of Z_{r_1+1} as a point below which $100q_1$ percent of the population $f(Z)$ lies. So, choosing $F(h_1)$ and $F(h_2)$ in this way helps to assure us that the probability is small that Z_{r_1+1} will fall outside the interval. This situation is similar to the problem of setting up control limits for fraction defective in quality control with a quality level of q_1 .

From (2.9) and (2.10) it is seen that as n becomes larger and larger $F(h_1)$ approaches q_1 from the left at the same rate that $F(h_2)$ approaches q_1 from the right. Thus the interval (h_1 , h_2) shrinks to a single point, and α and β can be found by evaluating the derivative of $g(Z)$ at the point h , where h is the solution to

$$F(h) = q_1 \quad . \tag{2.11}$$

Thus we have

$$\beta = \frac{d}{dZ}\Big[g(Z)\Big]_{Z = h}$$

and (2.12)

$$\alpha + \beta h = g(h)$$

giving values to α and β . The degree of accuracy of the linear approximation depends on the width of the interval. As (h_1 , h_2) becomes smaller and smaller, the amount of error between $g(Z)$ and $\alpha + \beta Z$ gets closer and closer to zero. It is a natural and desirable property of (h_1 , h_2), then, that it does decrease in size as n becomes larger.

Returning to equation (2.7), we see that a solution k_c for k can be obtained by setting

$$\frac{\partial L'}{\partial k} = 0$$

and solving for k . After substitution of $Y_i = \sqrt{k}\, Z_i$, equation (2.7) takes the form

$$\Big[(r_1 + r_2 - n)/k\Big] + \frac{1}{k^2}\sum_{i=r_1+1}^{n-r_2} Y_i^2 - \left[\frac{r_1 Y_{r_1+1}}{2k}\right]\left[\alpha + \frac{\beta Y_{r_1+1}}{\sqrt{k}}\right] + \frac{r_2 Y_{n-r_2}^2}{k^2} = 0 \tag{2.13}$$

when it is equated to zero. Multiplication by $2k^2$ to eliminate the fractions yields

$$2k(r_1 + r_2 - n) + 2\sum_{i=r_1+1}^{n-r_2} Y_i^2 - (r_1 Y_{r_1+1})(k\alpha + \sqrt{k}\,\beta Y_{r_1+1}) + 2r_2 Y_{n-r_2}^2 = 0 \ . \quad (2.14)$$

Next, isolating the square root on one side of the equation, squaring and collecting terms gives

$$k^2G^2 + k(2GB - D^2) + B^2 = 0 \ , \qquad (2.15)$$

where

$$G = 2(r_1 + r_2 - n) - r_1 Y_{r_1+1}\alpha$$

$$B = 2\sum_{i=r_1+1}^{n-r_2} Y_i^2 + 2r_2 Y_{n-r_2} \qquad (2.16)$$

$$D = r_1 Y_{r_1+1}^2 \beta \ .$$

The estimator k_c is then the positive root of equation (2.15):

$$k_c = \left\{(D^2 - 2GB) + \sqrt{(2GB - D^2)^2 - 4G^2B^2}\right\} \Big/ (2G^2) \ . \qquad (2.17)$$

For the sake of completeness, and to give a somewhat mathematical justification for the solution in (2.17), an analysis of the discriminant in (2.15) is given. From the definitions given in (2.16), we see that G is always negative, since

$$(r_1 + r_2) < n \quad , \quad Y_i > 0 \text{ for every } i \quad , \quad \text{and } \alpha > 0 \qquad (2.18)$$

due to the nature of the relationship of α to $g(Z)$. Similar reasoning shows that B and D^2 are always positive. Therefore,

$$\begin{aligned}(2GB - D^2)^2 - 4G^2B^2 &= 4G^2B^2 - 4GBD^2 + D^4 - 4G^2B^2 \\ &= D^4 - 4GBD^2\end{aligned} \tag{2.19}$$

will be positive if

$$D^4 > 4GBD^2 \quad . \tag{2.20}$$

But, since $D^2 > 0$, then dividing by D^2 gives us $D^2 > 4GB$, which we know to be true, since B is positive and G is negative.

Thus the solution k_c will always be a real-valued solution. This is a necessary requirement to be fulfilled, since the parameter k will always be a real-valued parameter for the Rayleigh population.

CHAPTER III

PROPERTIES OF THE ESTIMATOR

In the previous chapter an estimator was derived which is closely related to a maximum likelihood estimator, at least in the method by which it is derived. The difference is of course the fact that an approximation was made in the course of actually finding the estimator. It is a natural question, then, to ask how "good" is the estimator k_c ? The usual answers to this question come in the form of expected value, bias, variance and asymptotic properties of k_c , and in comparisons between k_c and other estimators for k . This chapter will be devoted to these discussions, and attempts will be made to answer the above question.

Calculation of the expected value of k_c would be very difficult to accomplish, particularly if we consider equation (2.17), which is the actual value of k_c . Instead it is easier to discuss the approximate conditional bias in the asymptotic case. Following Tiku [7], p. 160, who cites Kendall and Stuart [2], p. 44, this approximate bias becomes

$$B_1 = E\left(\frac{\partial L'}{\partial k}\right)/R^2(k) \quad , \tag{3.1}$$

where

$$R^2(k) = -E\left(\frac{\partial^2 L'}{\partial k^2}\right) \tag{3.2}$$

for large values of $n - r_1 - r_2$. Using equation (2.7), the value of B_1 can be calculated from the following equation:

$$E\left(\frac{\partial L'}{\partial k}\right) = [(r_1 + r_2 - n)/k)] + \frac{1}{k}\sum_{i=r_1+1}^{n-r_2} E(Z_i^2)$$

$$- (r_1\alpha/2\sqrt{k})E(Z_{r_1+1}) - (r_1\beta/2\sqrt{k})E(Z_{r_1+1}) + (r_2/k)E(Z^2_{n-r_2}) \; . \tag{3.3}$$

This calculation, then, requires expected values of order statistics from the Rayleigh distribution, which we proceed to find.

The distribution of the i^th^ order statistic from a random sample of size n from this population, a well-known result (see for example Sarhan and Greenberg [4]), is

$$\Psi(Z_i) = \frac{n!}{(i-1)!(n-i)!}\{F(Z_i)\}^{i-1}\{1 - F(Z_i)\}^{n-i}f(Z_i)$$

$$= \frac{n!}{(i-1)!(n-i)!}\{1 - e^{-Z_i^2}\}^{i-1}\{e^{-Z_i^2}\}^{n-i}\; 2Z_i e^{-Z_i^2} \; . \tag{3.4}$$

Now

$$E(Z_i) = \int_0^\infty Z_i\Psi(Z_i)dZ_i$$

$$= \frac{n!}{(i-1)!(n-i)!}\int_0^\infty \{1 - e^{-Z_i^2}\}^{i-1}\; 2Z_i^2 e^{-(n-i+1)Z_i^2}\, dZ_i \; . \tag{3.5}$$

Expanding the first term of the integrand function by the binomial theorem, we get

$$\left[1 - e^{-Z_i^2}\right]^{i-1} = \sum_{j=0}^{i-1}\binom{i-1}{j}(-1)^j e^{-jZ_i^2} \; . \tag{3.6}$$

Substituting this expression into (3.5) and interchanging the order of integration and summation, we obtain

$$E(Z_i) = \frac{n!}{(i-1)!(n-i)!} \sum_{j=0}^{i-1} \binom{i-1}{j} (-1)^j \int_0^\infty 2Z_i^2 e^{-(n-i+1+j)Z_i^2} \, dZ_i \ . \tag{3.7}$$

Gröbner and Hofreiter [1] give the value of the integral as

$$[\sqrt{\pi}]/2[n - i + 1 + j]^{3/2} \ ,$$

so that

$$E(Z_i) = \frac{n!}{(i-1)!(n-i)!} \sum_{j=0}^{i-1} \binom{i-1}{j} (-1)^j \left[\frac{\sqrt{\pi}}{2(n-i+1+j)^{3/2}} \right] . \tag{3.8}$$

Also

$$E(Z_i^2) = \int_0^\infty Z_i^2 \Psi(Z_i) dZ_i$$

$$= \frac{n!}{(i-1)!(n-i)!} \sum_{j=0}^{i-1} \binom{i-1}{j} (-1)^j \int_0^\infty 2Z_i^3 e^{-(n-i+1+j)Z_i^2} \, dZ_i \ , \tag{3.9}$$

after manipulation similar to that used in deriving (3.8). Referring once again to Gröbner and Hofreiter [1], we find the value of this integral and use it in (3.9) to obtain

$$E(Z_i^2) = \frac{n!}{(i-1)!(n-i)!} \sum_{j=0}^{i-1} \binom{i-1}{j} (-1)^j \frac{1}{(n-i+1+j)^2} \ . \tag{3.10}$$

Next we must calculate $R^2(k) = -E\left(\frac{\partial^2 L'}{\partial k^2}\right)$.

$$\frac{\partial^2 L'}{\partial k^2} = \left(\frac{n - r_1 - r_2}{k^2} \right) - \frac{1}{k^2} \sum_{i=r_1+1}^{n-r_2} Z_i^2 + \frac{r_1 \alpha}{2k} Z_{r_1+1} + \frac{r_1 \beta}{2k\sqrt{k}} Z_{r_1+1}^2 - \frac{2r_1 Z_{n-r_2}^2}{k^2} \ . \tag{3.11}$$

Then, taking expected values and multiplying by minus one,

$$-E\left(\frac{\partial^2 L'}{\partial k^2}\right) = \left(\frac{n - r_1 - r_2}{k^2}\right) - \frac{1}{k^2}\sum_{i=r_1+1}^{n-r_2} E(Z_i^2) + \frac{r_1\alpha}{2k} E(Z_{r_1+1}) + \frac{r_1\beta}{2k\sqrt{k}} E(Z_{r_1+1}^2)$$
$$- \frac{2r_1}{k^2} E(Z_{n-r_2}^2) \quad . \qquad (3.12)$$

Evaluation of the expected values by the use of (3.9) and (3.10) gives the terms necessary to calculate $R^2(k)$ from (3.12). Then the bias B_1 can be calculated from (3.1) as a function of k .

Kendall and Stuart [2] also give an expression for the asymptotic variance of k_c in terms of the likelihood equation, namely,

$$\mathrm{Var}(k_c) = \left[-E\left(\frac{\partial^2 L'}{\partial k^2}\right)\right]^{-1} = \left[R^2(k)\right]^{-1} \qquad (3.13)$$

which can be evaluated from (3.12). This variance and bias, then, can be used to compare the relative efficiency of k_c to other estimators whose variances and biases can be calculated.

Tiku [7] justifies the use of these asymptotic properties for his estimators because of their similarity to maximum likelihood estimators, and because the approximation used to find these estimators becomes more and more accurate as the sample size increases. In the case of the normal distribution, Tiku shows that the efficiencies of his estimators are as good as the usual maximum likelihood estimators and the best linear unbiased estimators. Thus an estimator calculated in this way does have desirable properties, perhaps the most important one being the ease with which it may be calculated. This is certainly true in the case of the Rayleigh distribution, since the actual calculation requires no iterative procedure nor any expected values of order statistics, both of which are computationally tedious procedures.

CHAPTER IV

SAMPLE GENERATION AND NUMERICAL RESULTS

The density (1.1) has the nice property that it can be integrated into a closed form for the cumulative distribution function F(x) . This property and the probability integral transformation are used to obtain a sample of size n from the population (1.1).

First we note that

$$u_i = F(x_i) = \int_0^{x_i} f(T)\,dT = 1 - e^{-\frac{x_i^2}{k}} \tag{4.1}$$

is uniformly distributed between zero and one. Then a sample u_1 , u_2 , ··· , u_n may be generated from the uniform distribution, such a sample being fairly easy to obtain by any of a number of methods, either from tables or by a simple computer program. From (4.1) we obtain

$$e^{-\frac{x_i^2}{k}} = 1 - u_i$$

or

$$\frac{x_i^2}{k} = - \log(1 - u_i) \quad .$$

Hence

$$x_i = \sqrt{k[- \log(1 - u_i)]} \tag{4.2}$$

will be distributed according to the Rayleigh distribution with parameter k . Using this procedure, we have the following values for our sample, size $n = 20$, of which $q_1 n = (.2)(20) = 4$ values have been censored from the left, and $q_2 n = (.1)(20) = 2$ values have been censored from the right. The values, in order of magnitude, are:

.69905, .78341, .79853, .91500, .95984, .97785, 1.00619,
1.13569, 1.17262, 1.20904, 1.27873, 1.37311, 1.41461, 1.58910,

using a population value $k = 2.007$ to generate the sample.

Now, using (2.9) and (2.10), we get the following values for interval endpoints:

$$h_1 = .34229 \quad , \quad h_2 = .58455$$

and, using (2.8) we get the following values of α and β :

$$\beta = -10.88613$$

$$\alpha = \quad 9.23360 \quad .$$

Now, substituting all these values into (2.17) yields

$$k_c = 1.703 \quad .$$

The following table gives values of k for k_c , the method of moments, least squares, and the population value.

Population Value	2.007
Moment Estimate	1.529
Least Squares Estimate	2.410
Approximate Maximum Likelihood (k_c)	1.703

These values show that k_c is as close to the actual value of k as any of the other estimates. While the maximum likelihood estimate has not been included in this discussion, it is clear that k_c is still a close estimate to k . Another advantage to k_c is that it can be used as a good starting point in an actual iterative solution of the maximum likelihood equation, if such a solution is desired. Beginning the iteration with k_c would give faster convergence than if a "guess" were made. For these reasons it is believed that k_c is a valuable tool to be used in working with the Rayleigh distribution.

APPENDIX

AN EMPIRICAL STUDY OF THE LINEAR APPROXIMATION $\alpha + \beta Z$

In Chapter I it was stated that the function

$$g(Z) = \frac{f(Z)}{F(Z)}$$

lies very close to the line

$$\alpha + \beta Z$$

for Z in a small interval (a,b) . In this discussion we give values of Z for which the approximation is close to the actual function values, and we evaluate the error between the actual value g(Z) and the approximation $\alpha + \beta Z$.

The following graph shows g(Z) and several of the straight lines used to approximate it. The absolute values of the maximum errors are also given in the description of each curve, which accompanies the plot itself.

The graph of g(Z) and the straight lines show that for values of Z less than .5, the approximation is not good. This is further verified by evaluation of the derivative of g(Z) at some of these points. When Z is less than .5, the derivative changes rapidly for small changes in Z . For example, the derivative at Z = .3 is approximately -19 . At Z = .4 it is approximately -9 , and at Z = .5 it is -6 . For larger and larger values of Z , the derivative becomes more nearly constant, and the linear approximation, then, becomes better and better.

Following the graph is a table of values of α and β for various sample sizes and different amounts of censoring. Equations (2.8) through (2.12) were used in the computations of α and β which are provided for use in computing k_c for various combinations of sample sizes and censoring.

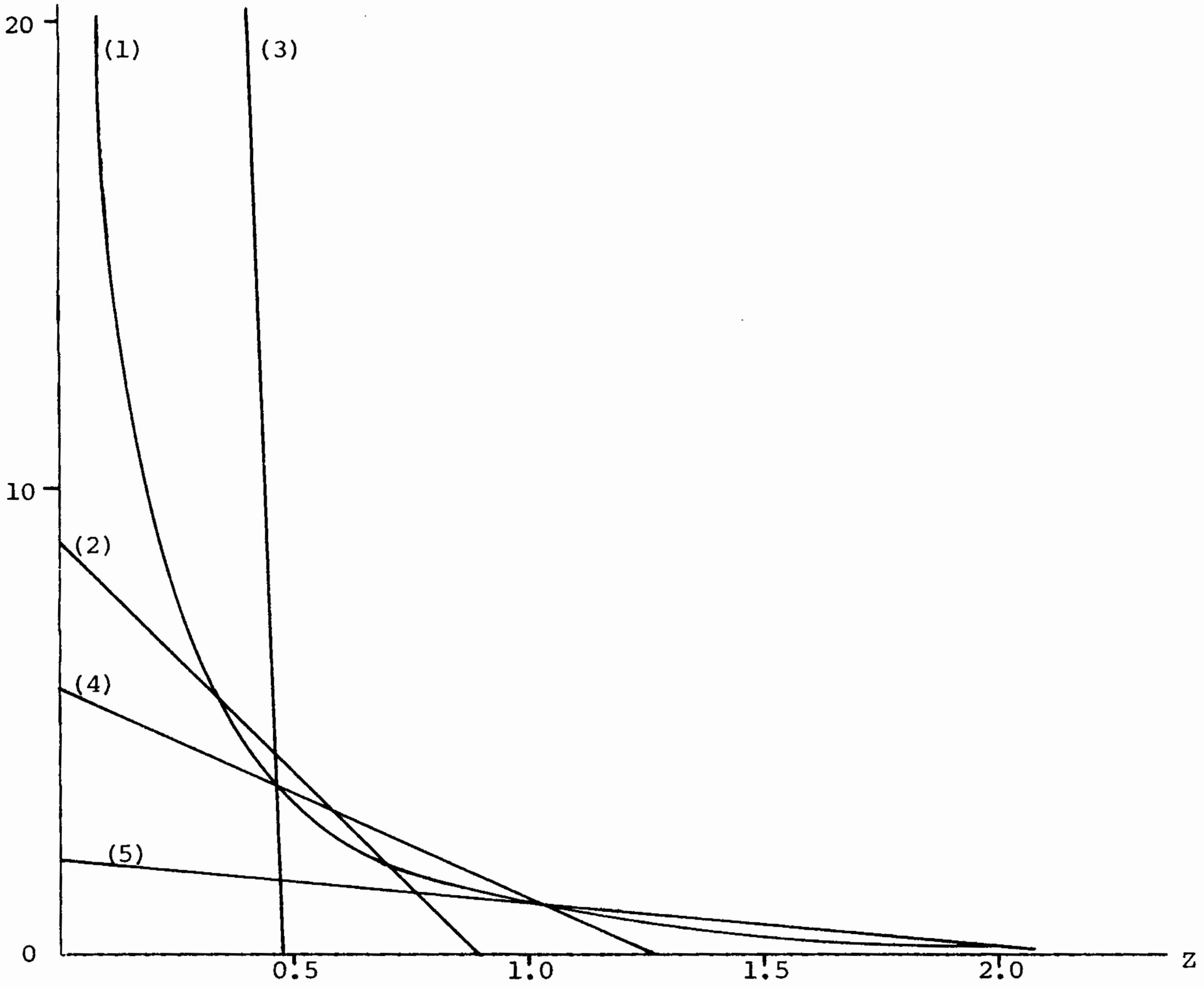

FIGURE 1. A graph comparing g(Z) and several of the straight lines used to approximate it.

(1) $g(Z) = (2Ze^{-Z^2})/(1 - e^{-Z^2})$.

(2) $\alpha + \beta Z$ used in computing sample values in Chapter IV.

(3) $\alpha + \beta Z$ for $Z\varepsilon(0.01, 0.5)$. In this interval the approximation is obviously not good and should not be used.

(4) $\alpha + \beta Z$ for $Z\varepsilon(0.5, 1.0)$. The maximum error in this interval = $|g(Z) - \alpha - \beta Z|$ = 0.364. Thus, the approximation can be used here with reasonably good results.

(5) $\alpha + \beta Z$ for $Z\varepsilon(1.0, 2.0)$. Here the maximum error = 0.270, so the approximation improves with increasing Z.

TABLE 1.

Values of α and β for Various Sample Sizes and Proportions of Censoring.

n	q_1	α	β
∞	.1	12.312	-19.930
	.2	8.433	-9.852
	.3	6.627	-6.430
	.4	5.476	-4.662
	.5	4.617	-3.545
	.6	3.898	-2.739
	.7	3.235	-2.091
	.8	2.552	-1.512
	.9	1.725	-0.915

n	q_1	α	β
20	.1	15.601	-26.518
	.2	9.234	-10.886
	.3	6.985	-6.764
	.4	5.668	-4.790
	.5	4.724	-3.587
	.6	3.954	-2.737
	.7	3.250	-2.060
	.8	2.525	-1.453
	.9	1.613	-0.806

n	q_1	α	β
10	.1	32.204	-60.908
	.2	10.393	-12.406
	.3	7.418	-7.173
	.4	5.883	-4.935
	.5	4.840	-3.633
	.6	4.011	-2.733
	.7	3.264	-2.023
	.8	2.485	-1.380
	.9	1.279	-0.585

THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

POISSON COUNTS OF A MARKOVIAN RENEWAL PROCESS

by

A. M. Kshirsagar

Technical Report No. 30
Department of Statistics THEMIS Contract

April 18, 1969

Research sponsored by the Office of Naval Research
Contract N00014-68-A-0515
Project NR 042-260

DEPARTMENT OF STATISTICS
Southern Methodist University

POISSON COUNTS OF A MARKOVIAN RENEWAL PROCESS

by

A. M. Kshirsagar*
Southern Methodist University
Dallas, Texas 75222

1. INTRODUCTION

Let F and G be two independent renewal processes commencing simultaneously at time $t = 0$. Let $F(x)$ be the distribution function (d.f.) of the intervals between successive renewals of the F-process and let $G(x)$ be the corresponding quantity for the G-process. If the renewals of the G-process occur at times $t = T_1 , T_2 , \cdots , T_r , \cdots$, we define the G-counts of the F-process by $P_{r,k}$ $(r = 1, 2, 3, \cdots ; k = 0, 1, 2, \cdots)$ where $P_{r,k}$ is the probability that the number of renewals of the F-process in the interval (T_{r-1} , T_r) is k . In particular, if the G-process is a Poisson process, $P_{r,k}$ are known as the Poisson counts of the F-renewal process. Kingman [3] has derived the generating function of such Poisson counts. In this paper, this generating function is derived by a different method viz. that of using the distribution of the number of renewals in an arbitrary interval of time $(t_0 , t_0 + t)$, where $t_0 \neq 0$ (see Cox [2], pg. 67). This method is then extended to obtain the Poisson counts of a Markov Renewal Process, as defined by Pyke [6], [7].

Let $T_r - T_{r-1} = x_r$, $T_0 = 0$ $(r = 1, 2, \cdots)$. Then T_{r-1} and x_r are independently distributed and the probability density function (p.d.f.)

*This research was sponsored by the Office of Naval Research, Contract No. N00014-68-A-0515, Project No. NR 042-260.

$p_1(T_{r-1})$ of T_{r-1} is given by (for $r \geq 2$),

$$p_1(T_{r-1})dT_{r-1} = \text{\{Probability that there are } r - 2 \text{ renewals of the G-process in } (0, T_{r-1})\text{\}\{Probability of a renewal in } (T_{r-1}, T_{r-1} + dT_{r-1})\text{\}} \tag{1.1}$$

$$= \left\{G_{r-2}(T_{r-1}) - G_{r-1}(T_{r-1})\right\}\left\{\frac{d}{dT_{r-1}} E\left[N^G(T_{r-1})\right]\right\} dTr-1$$

where $G_r(x)$ is the r-fold convolution of the d.f. $G(x)$ and $N^G(t)$ is the number of renewals of the G-process in the interval $(0,t)$; E stands for the expectation operator. The distribution function of x_r is obviously $G(x_r)$.

Let $A_k(t_0, t)$ be the probability that there are k renewals of the F-process in the interval $(t_0, t_0 + t)$ and let the corresponding probability generating function (p.g.f.) be

$$B(t,\xi,t_0) = \sum_{k=0}^{\infty} \xi^k A_k(t_0, t) . \tag{1.2}$$

When $t_0 = 0$, we denote $B(t,\xi,0)$ by $B(t,\xi)$ only. It can be readily seen that

$$P_{r,k} = \int_0^\infty\int_0^\infty A_k(T_{r-1}, x_r)p_1(T_{r-1})dT_{r-1}dG(x_r) , \text{ for } r \geq 2 , \tag{1.3}$$

and

$$P_{1,k} = \int_0^\infty A_k(0,x_1)dG(x_1) . \tag{1.4}$$

The double generating function (with respect to r and k) of the G-counts of the F-process is, therefore,

$$\psi(\theta,\xi) = \sum_{r=1}^{\infty} \sum_{k=0}^{\infty} \theta^{r}\xi^{k}P_{r,k}$$

$$= \theta\int_0^{\infty} B(x,\xi)\,dG(x) + \sum_{r=2}^{\infty} \theta^{r} \int_0^{\infty}\int_0^{\infty} B(x,\xi,T_{r-1})\,p_1(T_{r-1})\,dT_{r-1}\,dG(x) \; . \tag{1.5}$$

2. POISSON COUNTS OF THE F-PROCESS

The expression (1.5) for $\psi(\theta,\xi)$ can be evaluated explicitly, when the G-process is a Poisson process with parameter λ . For this we define the following Laplace transforms:

$$\int_0^{\infty} e^{-sx}dF(x) = f(s) \; ; \tag{2.1}$$

$$\int_0^{\infty} e^{-sx}B(x,\xi)\,dx = b(s,\xi) \; ; \tag{2.2}$$

$$\int_0^{\infty} e^{-sx}B(x,\xi,t_0)\,dx = b(s,\xi,t_0) \; ; \tag{2.3}$$

$$\int_0^{\infty} e^{-s_0t_0}b(s,\xi,t_0)\,dt_0 = b^{*}(s,\xi,s_0) \; . \tag{2.4}$$

Then it has been proved (see Cox [2], pp. 67, 68, 37) that

$$b(s,\xi) = \frac{1 - f(s)}{s\bigl(1 - \xi f(s)\bigr)} \; ; \tag{2.5}$$

$$b^*(s,\xi,s_0) = \frac{1}{ss_0} - \frac{(1 - \xi)\left\{f(s) - f(s_0)\right\}}{s(s_0 - s)\bigl(1 - \xi f(s)\bigr)\bigl(1 - f(s_0)\bigr)} \quad . \tag{2.6}$$

When the G-process is a Poisson process,

$$G(x) = 1 - e^{-\lambda x} \tag{2.7}$$

and so (1.5) reduces to

$$\psi(\theta,\xi) = \theta\lambda b(\lambda,\xi) + \sum_{r=2}^{\infty} \lambda^2\theta^2 \int_0^{\infty} \frac{(\lambda T_{r-1}\theta)^{r-2}}{(r - 2)!} b(\lambda,\xi,T_{r-1}) e^{-\lambda T_{r-1}} dT_{r-1} \tag{2.8}$$

$$= \theta\lambda b(\lambda,\xi) + \theta^2\lambda^2 b^*\bigl(\lambda,\xi,\lambda(1 - \theta)\bigr) \tag{2.9}$$

$$= \frac{\theta}{1 - \theta} - \frac{\theta(1 - \xi)f\bigl(\lambda(1 - \theta)\bigr)\bigl(1 - f(\lambda)\bigr)}{\bigl(1 - \xi f(\lambda)\bigr)\Bigl(1 - f\bigl(\lambda(1 - \theta)\bigr)\Bigr)} \quad , \tag{2.10}$$

on account of (2.5) and (2.6). This agrees with Kingman's [3] equation (16) on pg. 1220, except for the fact that he takes $\lambda = 1$.

Incidently, this shows that, if two renewal processes F_1 and F_2 have the same Poisson counts, expression (2.10) for F_1 will be identical with a similar expression for F_2 for all values of ξ and θ and this will show that the Laplace transforms of F_1 and F_2 are the same or that the two renewal processes are identical.

3. POISSON COUNTS OF A CUMULATIVE PROCESS

Suppose that a variable w , whose d.f. is $C(w)$ is associated with every renewal of the F-process. A cumulative process, then can be defined as

$$Z(t) = \sum_{i=1}^{N^F(t)} w_i \; . \tag{3.1}$$

where, as defined earlier, $N^F(t)$ is the number of renewals of the F-process in (o,t). The w_i are all identically and independently distributed. Poisson counts of this cumulative process are obtained by considering the distribution of

$$Z_r = \sum w_j \quad , \quad r = 1, 2, \cdots \tag{3.2}$$

where the summation is over the renewals of the F-process that occur in (T_{r-1} , T_r). If there are k such renewals (and the probability of this $P_{r,k}$) , the Laplace transform of the distribution of Z_r , when k is fixed, is $\{c(s)\}^k$, where $c(s)$ is the Laplace transform of $C(w)$. Hence the unconditional Laplace transform of the distribution of Z_r is

$$\sum_{k=0}^{\infty} [c(s)]^k P_{r,k} \quad ; \quad r = 1, 2, \cdots \tag{3.3}$$

The generating function of these Laplace transforms is

$$\sum_{r=1}^{\infty} \sum_{k=0}^{\infty} \theta^r [c(s)]^k P_{r,k} = \psi\big(\theta,c(s)\big) \; . \tag{3.4}$$

where $\psi(\theta,\xi)$ is given by (2.10)

4. POISSON COUNTS OF A MARKOVIAN RENEWAL PROCESS

Markov Renewal Processes (M.R.P.) have been defined and studied extensively by Pyke [6], [7]. We shall use the same notation, as far as possible, as Pyke has used. A M.R.P. records at each time t , the number of times a system visits each of m states $(m < \infty)$, in time t , if the transitions from state to state are according to a Markov chain and if the time required for each successive move is a random variable, the d.f. of which depends on the two states, between which the move is made. Let $[p_{ij}]$ be the transition probability matrix of the Markov chain, and let $F_{ij}(x)$ be the d.f. of the time taken to make a transition from state i to state j . We set $Q_{ij}(x) = p_{ij}F_{ij}(x)$, and define the Laplace-Stieltjes transform (L. - S.T.) as

$$q_{ij}(s) = \int_0^\infty e^{-sx} dQ_{ij}(x) \quad ; \quad i, j = 1, 2, \cdots , m \quad . \tag{4.1}$$

Let q(s) be the $m \times m$ matrix of the $q_{ij}(s)$. We assume the system to be in state i initially and denote this by $Z_0 = i$; Z_t denotes the state of the system at time t . Further, we define the following quantities:

$$N_j(t) = \begin{array}{l}\text{number of times the system visits state } j \text{ in} \\ \text{the interval } (0,t) \quad j = 1, 2, \cdots , m \quad .\end{array} \tag{4.2}$$

$$N_j(t_0 , t) = \begin{array}{l}\text{number of times the system visits state } j \text{ in} \\ \text{the interval } (t_0 , t_0 + t) \ , \ t_0 \neq 0 \quad .\end{array} \tag{4.3}$$

$$\xi = \text{a diagonal matrix with elements } \xi_1 , \cdots , \xi_m \ . \tag{4.4}$$

$$\underline{e} = \text{a column vector } (m \times 1), \text{ with unit elements.} \tag{4.5}$$

$$B_i(t,\xi,t_0) = \sum_{n_1,\cdots,n_m=0}^{\infty} \xi_1^{n_1}\xi_2^{n_2}\cdots\xi_m^{n_m} \, \text{Prob}\{\mathbf{N}_j(t_0 , t) = n_j \ ; \ j=1,2,\cdots,m | Z_0=i\}, \tag{4.6}$$

$$B_i(t,\xi) = \sum_{n_1,\cdots,n_m=0}^{\infty} \xi_1^{n_1} \cdots \xi_m^{n_m} \text{Prob}\{N_j(t) = n_j | Z_0 = i\},\ j = 1, 2, \cdots, m \tag{4.7}$$

$$\begin{aligned} b_i(s,\xi,t_o) &= \text{L. - S.T. of } B_i(t,\xi,t_0) \text{ with respect to } t \\ &= \int_0^\infty e^{-st} d_t B_i(t,\xi,t_0) dt_0 \quad , \end{aligned} \tag{4.8}$$

$$b_i^*(s,\xi,s_0) = \int_0^\infty e^{-s_0 t_0} b_i(s,\xi,t_0) dt_0 \quad , \tag{4.9}$$

$$b_i(s,\xi) = \int_0^\infty e^{-st} d_t B_i(t,\xi) \quad . \tag{4.10}$$

$\underline{b}^*(s,\xi,s_0)$ is the column vector of $b_i^*(s,\xi,s_0)$; $i = 1, 2, \cdots, m$. (4.11)

$\underline{b}(s,\xi)$ is the column vector of $b_i(s,\xi)$; $i = 1, 2, \cdots, m$. (4.12)

The author has shown elsewhere [4] that

$$b^*(s,\xi,s_0) = \frac{1}{s_0}\underline{e} + \frac{1}{s_0 - s}\Big(I - q(s_0)\Big)^{-1}\Big(q(s) - q(s_0)\Big)\Big(\xi\underline{b}(s,\xi) - \underline{e}\Big) \tag{4.13}$$

and

$$\underline{b}(s,\xi) = \Big(I - q(s)\xi\Big)^{-1}\Big(I - q(s)\ \underline{e}\Big) . \tag{4.14}$$

The last result has been proved in [5]; see also [1].

Poisson counts of the M.R.P. can now be defined by

$$\begin{aligned} P_{ir}(n_1, n_2, \cdots, n_m) &= \text{Prob}[N_j(T_{r-1}, T_r) = n_j | Z_0 = i]\ j = 1, 2, \cdots, m \\ i &= 1, 2, \cdots, m \\ r &= 1, 2, \cdots\cdots\cdot \end{aligned} \tag{4.15}$$

This is the joint distribution of the number of visits of the M.R.P. to the various states, in the random interval $(T_{r-1}\ ,\ T_r)$ defined for the Poisson process. Proceeding exactly in the same way, as in Section 2, the generating function of these counts is

$$\psi_i(\theta,\xi) = \sum_{r=1}^{\infty} \theta^r \sum_{n_1,\cdots,n_m=0}^{\infty} \xi_1^{n_1}\xi_2^{n_2} \cdots \xi_m^{n_m}\, P_{ir}(n_1\ ,\ n_2\ ,\ \cdots\ ,\ n_m)$$

$$= \theta^2\lambda^2 \int_0^{\infty}\int_0^{\infty} B_i(x,\xi,T_{r-1})e^{-\lambda(1-\theta)T_{r-1}}\ e^{-\lambda x}\ dT_{r-1}dx$$

$$+ \theta\lambda\int_0^{\infty} B_i(x,\xi)e^{-\lambda x}dx \tag{4.16}$$

$$= (\theta\lambda)\ \frac{b_i(\lambda,\xi)}{\lambda} + (\theta^2\lambda^2) \int_0^{\infty} \frac{b_i(\lambda,\xi,T_{r-1})}{\lambda}\ e^{-\lambda(1-\theta)T_{r-1}}\ dT_{r-1}$$

$$= \theta b_i(\lambda,\xi) + \theta^2\lambda b_i^*\Big(\lambda,\xi,\lambda(1 - \theta)\Big)$$

We shall denote the column vector of the m generating functions $\psi_i(\theta,\xi)$, $(i = 1,\ 2,\ \cdots\ ,\ m)$ by $\underline{\psi}(\theta,\xi)$ and on account of (4.16) and (4.13), (4.14), it can be expressed as

$$\frac{\theta^2}{1 - \theta}\underline{e} + \theta\Big(I - q(\lambda)\xi\Big)^{-1}\Big(I - q(\lambda)\Big)\underline{e}$$
$$- \theta\Big(I - q(\mu)\Big)^{-1}\Big(q(\lambda) - q(\mu)\Big)\xi\Big(I - q(\lambda)\xi\Big)^{-1}(I - \xi^{-1})\underline{e} \tag{4.17}$$

where $\mu = \lambda(1 - \theta)$.

5. CUMULATIVE PROCESS

Corresponding to each state visited, let there be a random variable, whose d.f. depends on the state and assume that these variables are independently distributed. Let w be the variable and its d.f. be $C_j(w)$, corresponding to the state j , $(j = 1, 2, \cdots , m)$. Then this w defines a cumulative process associated with the M.R.P. and is defined by

$$Z(t) = \sum_{\substack{\text{all transitions of} \\ \text{the M.R.P. in } (0,t)}} w \tag{5.1}$$

Poisson counts of this cumulative process can be defined by

$$Z_r = \sum w \quad , \quad r = 1, 2, \cdots \cdot \tag{5.2}$$

where the summation is over all transitions in the random interval (T_{r-1} , T_r) . If the system visits state j , n_j times in (T_{r-1} , T_r) , $(j = 1, \cdots , m)$, then conditional on $n_1 , \cdots , n_m$, the L. - S.T. of Z_r is

$$\prod_{j=1}^{m} \{c_j(s)\}^{n_j} \tag{5.3}$$

where $c_j(s)$ is the L. - S.T. of $C_j(w)$, and hence the unconditional L. - S.T. is

$$\sum_{n_1,\cdots,n_m=0}^{\infty} P_{ir}(n_1 , \cdots , n_m) \prod_{j=1}^{m} \{c_j(s)\}^{n_j} \quad , \qquad \text{if } Z_0 = i \text{ .} \quad r = 1, 2, \cdots , m \text{ .} \tag{5.4}$$

The generating function of these L. - S.T.'s is, therefore

$$\sum_{r=1}^{m} \theta^r \sum_{n_1,\cdots,n_m=0}^{\infty} P_{ir}(n_1 , \cdots , n_m) \prod_{j=1}^{m} \{c_j(s)\}^{n_j} = \psi_i\Big(\theta, c(s)\Big) , \qquad (5.5)$$

where ψ_i is given by (4.16) and (4.17) and c(s) is the diagonal matrix of the m elements $c_j(s)$; j = 1, 2, $\cdots$, m .

REFERENCES

[1] Cinlar, E. (1968). Some joint distributions for Markov renewal processes. J. Austr. Stat. Soc., 10, 8-20.

[2] Cox, D. R. (1962). Renewal Theory. John Wiley and Sons, Inc., New York.

[3] Kingman, J. F. C. (1963). Poisson counts for random sequences of events. Ann. Math. Statist., 34, 1217-1232.

[4] Kshirsagar, A. M. and Gupta, Y. P. (1969). Distribution of the number of Markovian renewals in an arbitrary interval. To appear in J. Austr. Stat. Soc.

[5] Kshirsagar, A. M. and Gupta, Y. P. (1969). Some results in Markov renewal processes. To appear in Cal. Stat. Assoc. Bull.

[6] Pyke, R. (1961). Markov renewal processes: definitions and preliminary properties. Ann. Math. Statist., 32, 1231-1242.

[7] Pyke, R. (1961). Markov renewal processes with finitely many states. Ann. Math. Statist., 32, 1243-1259.