NOTICE

THIS DOCUMENT HAS BEEN REPRODUCED FROM MICROFICHE. ALTHOUGH IT IS RECOGNIZED THAT CERTAIN PORTIONS ARE ILLEGIBLE, IT IS BEING RELEASED IN THE INTEREST OF MAKING AVAILABLE AS MUCH INFORMATION AS POSSIBLE
STATISTICAL ANALYSIS
OF MULTIVARIATE
ATMOSPHERIC VARIABLES

Final Technical Report

by

Jack D. Tubbs
Principal Investigator

Department of Mathematics
University of Arkansas
Fayetteville, Arkansas 72701

March 31, 1979

Prepared for the
NATIONAL AERONAUTICS
AND SPACE ADMINISTRATION
George C. Marshall Space Flight Center
Marshall Space Flight Center, AL 35812

Under Contract No. NAS 8-31550
Control No. 505-8-10-ES-6-004-300-2150
Final Report

STATISTICAL ANALYSIS OF MULTIVARIATE ATMOSPHERIC VARIABLES

Jack D. Tubbs
Principal Investigator

Contract # NAS 8-31550
Control # 505-8-10-ES-6-004-300-2150

March 31, 1979
STATISTICAL ANALYSIS OF MULTIVARIATE
ATMOSPHERIC VARIABLES

FINAL REPORT
ACKNOWLEDGMENTS

Research work contained in this final report was performed for the George C. Marshall Space Flight Center of the National Aeronautics and Space Administration for the period commencing October 31, 1975 and ending March 31, 1979. Dr. M. Carter was the initial principal investigator and was responsible for the first three reports. Dr. James Dunn and graduate students Ms. Debra Waits, Mr. Bradley Skarpness, Mr. Gary Spencer, and Mr. Chung Jin Lee also contributed to some of the reports.

Jack D. Tubbs
Principal Investigator
PREFACE

The work presented in this final report was in response to the following three topics as suggested in the contract's scope of work.

1) Investigation of possible multivariate extensions of existing univariate distributions which have been used for modeling meteorological phenomenon.

2) Development of Goodness-of-fit tests, in particular for non-Gaussian distributions.

3) Investigation of the effect of correlated observations on statistical inference.

Reports 1-4 are concerned with some aspects of topic #1. Report 1 contains an estimation procedure for several discrete multivariate distributions. Report 2 contains a procedure for computing cloud cover frequencies in the bivariate case. This procedure can be used to compute probabilities for cloud frequencies for either two geographical locations or for the same location at different times. Report 3 contains the procedure and corresponding computer code for calculating conditional bivariate normal parameters. This report was requested by the COR. Report 4 contains a procedure for transforming multivariate non-Gaussian distributions into a nearly Gaussian distribution.

Reports 5 and 6 are concerned with topic #2. Report 5 contains a goodness-of-fit test for the extreme value distribution which is used in many meteorological applications. Report 6 contains a goodness-of-fit test for several continuous distributions.
Report 7 is concerned with the problem given in topic #3. In this report, the effect of autocorrelated observations on confidence regions is investigated.

Report 8 contains a computer code for generating both random and non-random observations for specified distributions. This program was used to generate the samples for the Monte Carlo simulation needed in the other reports.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ACKNOWLEDGMENTS</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREFACE</td>
<td>11</td>
</tr>
<tr>
<td>REPORTS</td>
<td></td>
</tr>
<tr>
<td>1. ESTIMATION IN DISCRETE MULTIVARIATE DISTRIBUTIONS</td>
<td>1</td>
</tr>
<tr>
<td>Debra A. Waits</td>
<td></td>
</tr>
<tr>
<td>2. A PROCEDURE TO PREDICT CLOUD COVER FREQUENCIES IN THE BIVARIATE CASE</td>
<td>17</td>
</tr>
<tr>
<td>Debra A. Waits &amp; Michael C. Carter</td>
<td></td>
</tr>
<tr>
<td>3. A PROGRAM TO COMPUTE CONDITIONAL BIVARIATE NORMAL PARAMETERS</td>
<td>58</td>
</tr>
<tr>
<td>Michael C. Carter</td>
<td></td>
</tr>
<tr>
<td>4. TRANSFORMATION OF NON-NORMAL MULTIVARIATE DATA TO NEAR-NORMAL</td>
<td>65</td>
</tr>
<tr>
<td>James Dunn &amp; Jack Tubbs</td>
<td></td>
</tr>
<tr>
<td>5. TEST OF FIT FOR THE EXTREME VALUE DISTRIBUTION BASED UPON THE GENERALIZED MINIMUM CHI-SQUARE</td>
<td>79</td>
</tr>
<tr>
<td>Chung Jin Lee &amp; Jack Tubbs</td>
<td></td>
</tr>
<tr>
<td>6. TEST OF FIT FOR CONTINUOUS DISTRIBUTIONS BASED UPON THE GENERALIZED MINIMUM CHI-SQUARE</td>
<td>87</td>
</tr>
<tr>
<td>Jack Tubbs &amp; Gary Spencer</td>
<td></td>
</tr>
<tr>
<td>7. EFFECT OF CORRELATED OBSERVATIONS ON CONFIDENCE SETS BASED UPON CHI-SQUARE STATISTICS</td>
<td>123</td>
</tr>
<tr>
<td>Jack Tubbs</td>
<td></td>
</tr>
<tr>
<td>8. GENERATION OF RANDOM VARIATES FROM SPECIFIED DISTRIBUTIONS</td>
<td>134</td>
</tr>
<tr>
<td>Jack Tubbs &amp; Gary Spencer</td>
<td></td>
</tr>
</tbody>
</table>
Summary

Procedures for estimating the parameters of three discrete multivariate distributions, the Multinomial, Negative Multinomial, and the multivariate Poisson distribution, are given along with approximate variances for the parameter estimates.

I. INTRODUCTION

This paper is concerned with the problems associated with the estimation of parameters for three discrete multivariate distributions, the multinomial, negative multinomial, and the multivariate Poisson, which are the multivariate extensions of three common univariate discrete distributions, the binomial, the negative binomial, and the Poisson distribution. The distributions are introduced in Section 2. A detailed explanation of the estimation procedures along with approximate bounds for the variances of the estimates are given in Section 3. An example is presented in Section 4 which is intended to demonstrate the use of the estimation procedures. A listing and card input description of the computer program is given in the Appendix.
II. DISTRIBUTIONS

Johnson and Kotz (1969, Ch. 11) provides a detailed discussion of the functions described below.

2.1 Multinomial Distribution

The simplest of the three distributions both in structure and theory is the multinomial distribution. Let $E_1, E_2, \ldots, E_k$ be possible events which can occur from a series of independent trials. If $E_j$ has probability $P_j$ of occurring and $n_j$ is the number of times $E_j$ occurs in the $N$ trials where $\sum_{j=1}^{k} n_j = N$, then the joint distribution of the random variables $n_1, n_2, \ldots, n_k$ is the multinomial distribution with parameters $N, P_1, P_2, \ldots, P_k$. The distribution is defined by

$$P(n_1, n_2, \ldots, n_k) = \frac{N!}{\prod_{j=1}^{k} n_j!} \left( \prod_{j=1}^{k} P_j^{n_j} \right) (Q \leq n_j \leq N) \quad (1)$$

2.2 Negative Multinomial

Just as the multinomial distribution is a natural extension of the binomial distribution, the multivariate negative binomial distribution is a natural extension of the negative binomial distribution. Hence, the probability generating function for the multivariate negative binomial is defined by

$$(Q - \sum_{i=1}^{k} P_i t_i)^{-N} \quad (2)$$
with $P_i > 0$ for all $i=1,\ldots,k$; $N > 0$, and $\sum_{i=1}^{k} P_i = 1$.

From formula (2) we have the following distribution function

$$P(n_1,n_2,\ldots,n_k) = \frac{\binom{N+r}{\sum_{i=1}^{k} n_i}}{\prod_{i=1}^{k} n_i!} Q^{-N} \prod_{i=1}^{k} (P_i/Q)^{n_i} \quad (3)$$

where $n_i > 0$, $i=1,\ldots,k$.

This is called the negative multinomial (or multivariate negative binomial) distribution with parameters $N,P_1,P_2,\ldots,P_k$, where $N$ is a non-negative integer. A special form of this distribution is a compound Poisson distribution which can be further simplified to a bivariate form as described by Bates and Neyman (1952).

### 2.3 Multivariate Poisson

Consider a sequence of $k$ variables $x_1,x_2,\ldots,x_k$ such that each one is a combination of two independent univariate Poisson variables where one of the Poisson variables is present in all $k$ variables. That is,

$$x_1 = u+v_1, x_2 = u+v_2, \ldots, x_k = u+v_i$$

and $u,v_1,v_2,\ldots,v_k$ are independent univariate Poisson variables with expected values $\xi, \theta_1, \theta_2, \ldots, \theta_k$ respectively. The joint distribution of $x_1,x_2,\ldots,x_k$ is

$$P(x_1,\ldots,x_k) = \exp(-\xi-\theta_1-\ldots-\theta_k) \sum_{j=0}^{\infty} \frac{x_1-j}{\theta_1} \frac{\theta_1}{(x_1-j)!} \frac{x_2-j}{\theta_2} \frac{\theta_2}{(x_2-j)!} \cdots \frac{x_k-j}{\theta_k} \frac{\theta_k}{(x_k-j)!} \quad (4)$$
where \( m = \min(x_1, x_2, \ldots, x_k) \). This is called the multivariate Poisson distribution with parameters \( \xi, \theta_1, \theta_2, \ldots, \theta_k \).

III. ESTIMATION

In section 3.1 the techniques used to estimate the parameters of the three distributions are described. The subsequent section is concerned with the variances of the estimates for the multinomial and negative multinomial distributions. A computer program was written to perform the needed computations.

3.1 Parameter Estimation

The maximum likelihood estimates of \( P_1, P_2, \ldots, P_k \) for the multinomial distribution are the relative frequencies

\[
\hat{P}_j = \frac{n_j}{N} \quad (j = 1, \ldots, k)
\]

(5)

where \( n_j \) is the observed frequency of \( E_j \) given \( N \) independent trials.

The method of moments is the most convenient approach for estimating the parameters of the negative multinomial distribution. The moment generating function of a \( k \) variate negative multinomial distribution is

\[
m(t_1, \ldots, t_k) = (Q - \sum_{i=1}^{k} P_i e^{t_i})^{-N}.
\]

Thus we obtain the following moments

\[
E(n_j) = \left. \frac{\partial m(t_1, \ldots, t_k)}{\partial t_j} \right|_{t=0} = NP_j \quad \text{for} \quad j = 1, \ldots, k
\]
\[ E(n_i n_j) = \frac{a^2 m(t_1, \ldots, t_k)}{t_j t_i} \quad \text{at } t=0 \]

\[ = N(N+1)P_i P_j \]
\[ = N^2 P_i P_j + N P_i P_j \]
\[ = E(n_i) E(n_j) + \frac{E(n_i) E(n_j)}{N} \]

giving

\[ N = \frac{E(n_i) E(n_j)}{E(n_i n_j) - E(n_i) E(n_j)} \quad (6) \]

and

\[ P_j = \frac{E(n_j)}{N}. \quad (7) \]

Equating raw estimates to moments to obtain an estimate for \( N \), we have

\[ \hat{N} = \frac{\bar{n}_i \bar{n}_j}{\bar{n}_i \bar{n}_j - \bar{n}_i \bar{n}_j} \quad \text{and} \quad \hat{P}_j = \frac{\bar{n}_j}{\hat{N}} \quad \text{for } i, j=1, \ldots, k \text{ and } i \neq j \text{ where} \]

\[ \bar{n}_L = \frac{\sum_{i=1}^{k} n_i L_i}{n}; \quad n_L^{n_k} = \frac{\sum_{i=1}^{k} n_i L_i}{n} \quad \text{given } n \text{ observations.} \]

The accompanying computer program utilizes this method of moments in two ways. There are \( k(k-1)/2 \) possible estimates of \( N \) by this method where \( k \) is the number of parameters. Similarly there are \( k(k-1)/2 \) possible values of \( \bar{n}_i \bar{n}_j \) as well as \( \bar{n}_i \bar{n}_j \). The program first averages the \( k(k-1)/2 \) values of \( \bar{n}_i \bar{n}_j \) and \( \bar{n}_i \bar{n}_j \) and then outputs an estimate of \( N \) based on these averages. The second approach calculates the \( k(k-1)/2 \) estimates of \( N \) and prints out the average estimate of \( N \). The parameters \( P_i, i=1, \ldots, k \) is also estimate: twice corresponding to the two estimates of \( N \).
The method of moments is also used in estimating the parameters of a multivariate Poisson. The moment generating function is given by

\[ m(t_1, \ldots, t_k) = \exp \left[ -\xi - \sum_{i=1}^{k} t_i \left( 1 - \exp(-t_i) \right) - \sum_{i=1}^{k} e_i \left( 1 - e_i \right) \right] \]  

It follows that

\[ \frac{\partial m(t_1, \ldots, t_k)}{\partial t_i} \bigg|_{t=0} = E(x_i) = \xi + \theta_i \]  

\[ \frac{\partial^2 m(t_1, \ldots, t_k)}{\partial t_i \partial t_j} \bigg|_{t=0} = E(x_i x_j) = (\xi + \theta_i)(\xi + \theta_j) + \xi \]  

Therefore

\[ \xi = E \left[ x_i x_j \right] - E \left[ x_i \right] E \left[ x_j \right] - \xi \]  

Substituting raw estimates for expected values we have

\[ \hat{\xi} = \bar{x_i x_j} - \bar{x_i} \bar{x_j} \]  

\[ \bar{x_i} = \frac{\sum_{i=1}^{n} x_{i_1}}{n} \]  

\[ \bar{x_i x_j} = \frac{\sum_{i=1}^{n} x_{i_1} x_{j_1}}{n} \]  

Since \( \hat{\theta_i} = E(x_i) - \xi \), a method of moments estimate for \( \theta_i \) is \( (\bar{x_i} - \hat{\xi}) \). Again the accompanying computer program uses two approaches to estimate \( \xi \) via the method of moments. First the program averages all possible values for \( \bar{x_i x_j} \) and \( \bar{x_i} \bar{x_j} \) and estimates \( \xi \) based on these two averages. Next the program averages the \( k(k-1)/2 \) possible estimates of \( \xi \) and outputs
this average as a workable estimate of \( \xi \). The parameters of \( \theta_i \), \( i=1, \ldots, k \) are estimated twice to correspond to the two estimates considered for \( \xi \).

### 3.2 Variances of Parameter Estimates

The exact variance of the estimates for the multinomial parameters can be easily derived. Consider

\[
\text{var} (\hat{\pi}_j) = \text{var} \left( \frac{n_j}{N} \right).
\]

\[
= E\left( \frac{n_j}{N} \right)^2 - \left( E\left( \frac{n_j}{N} \right) \right)^2
\]

\[
= \frac{1}{N^2} (N^2 \hat{p}_j^2 + N \hat{p}_j q_j) - \hat{p}_j^2
\]

\[
= \frac{\hat{p}_j q_j}{N} = \frac{\hat{p}_j (1 - \hat{p}_j)}{N},
\]

hence an approximate variance for \( \hat{\pi}_j \) is \( \hat{\pi}_j (1 - \hat{\pi}_j) / N \).

In order to place approximate bounds on the variances of the negative multinomial parameter estimates, consider Fisher's Information Matrix for the maximum likelihood parameter estimates which is defined as

\[
V(a_1, a_2, \ldots, a_k) = \left( E \left[ - \frac{\partial^2 \log L}{\partial a_i \partial a_j} \right] \right)^{-1}
\]

where \( a_i \) and \( a_j \) are parameters and \( L \) is the likelihood function. Kendall and Stuart have shown that this matrix is the asymptotic variance-covariance matrix for the maximum likelihood parameter estimates. From equation (3), we have the following

\[
L = \prod_{j=1}^{k} \prod_{i=1}^{n_{ij}} \left( n_{ij}! \right) (r(N))^n
\]

\[
\prod_{j=1}^{k} \prod_{i=1}^{n_j} \prod_{i=1}^{n_{ij}} \left( n_{ij}! \right) (r(N))^n
\]

\[
Q^{-n_j} \prod_{i=1}^{n} \prod_{j=1}^{k} \prod_{i=1}^{n_{ij}} \prod_{i=1}^{n_{ij}} (r(N))^n
\]

\[
\frac{\hat{p}_i (P_i/Q_j)}{N} \text{ for } j=1, \ldots, k
\]
\[ \ln L = \sum_{j=1}^{n} \ln r(N+S_j) - \ln \left( \prod_{j=1}^{n} \prod_{i=1}^{k} n_{ij}! \right) - n \ln r(N) \]  \hspace{1cm} (15) \\

\[ -N n \ln Q + \sum_{i=1}^{k} (n \sum_{i=1}^{n} n_{ij})(\ln p_i - \ln Q) \]

where \( S_j = \sum_{i=1}^{n} n_{ij} \), \( S_1 = \sum_{j=1}^{n} n_{ij} \), \( n \) is the number of samples taken and \( n_{ij} \) is the number of times \( E_i \) is satisfied on the \( j \)th sample.

\[ \frac{\partial \ln L}{\partial N} = \sum_{j=1}^{n} \frac{E(S_j) - 1}{\sum_{k=0}^{N+K}} - n \ln \hat{Q} \]  \hspace{1cm} (16) \\

\[ \frac{\partial^2 \ln L}{\partial N^2} = \sum_{j=1}^{n} \frac{E(S_j) - 1}{\sum_{k=0}^{N+K}} = \sum_{k=1}^{\infty} - \frac{(N+k-1)^{-2} E(F_j)}{(N+k)^2} \]  \hspace{1cm} (17) \\

where \( F_j \) is the number of \( S_j \)'s greater than or equal to \( j \),

\[ \frac{\partial \ln L}{\partial p_i} = \frac{-n}{1 + \sum_{i=1}^{k} \hat{p}_i} \]  \hspace{1cm} (18) \\

\[ \frac{\partial \ln L}{\partial p_i} = \frac{-Nn - \sum_{j=1}^{k} \hat{p}_i E(S_{i,j}^1)}{1 + \sum_{j=1}^{k} \hat{p}_j} + \frac{E(S_{i,j}^1)}{p_i} \]  \hspace{1cm} (19) \\

\[ \frac{\partial^2 \ln L}{\partial p_i^2} = \frac{Nn + \sum_{j=1}^{k} E(S_{i,j}^1)}{(1 + \sum_{j=1}^{k} \hat{p}_j)^2} - \frac{E(S_{i,j}^1)}{p_i^2} \]  \hspace{1cm} (20)
\[
\frac{\hat{\mu}_{ij}}{\hat{\mu}_{j}} = \frac{\hat{N} \mu_{ii} + \sum_{k \neq i} \mu_{ik}}{1 + \sum_{k \neq i} \mu_{ik}^2} \quad \text{for } i \neq j
\]

Using the two sets of estimates for \(N, \hat{P}_1, \ldots, \hat{P}_k\) and numerical values for \(E(S_{ij}), E(S_{ij}^{-1}), \text{ and } E(P_{ij})\), we can obtain approximate bounds for \(V(N, \hat{P}_1, \ldots, \hat{P}_k)\).

IV. AN EXAMPLE

Negative multinomial data were obtained from Arbous and Kerrich (1951, p. 424) to illustrate the output from the computer program. The results are found in Table 1. Notice that in the binomial case both estimates of \(N\) are the same since there are only two variables. For this same reason, only one Fisher's information matrix is produced. If more than two variables were considered, we would have obtained two different estimates for \(N\) and the information matrix. From the two distinct variances obtained from these matrices one could obtain the boundary points of the interval about the variance of the parameter estimates.
TABLE 1.

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>The moment estimate of $N$ obtained by averaging the raw moments first is</td>
<td>3.350</td>
</tr>
<tr>
<td>The corresponding probabilities associated with the respective variables are</td>
<td>0.295 0.385</td>
</tr>
<tr>
<td>The moment estimate of $N$ obtained by averaging all possible moment estimates</td>
<td>3.350</td>
</tr>
<tr>
<td>The corresponding probabilities associated with the respective variables are</td>
<td>0.295 0.385</td>
</tr>
<tr>
<td>Fisher's information matrix using the minimum estimate of $N$</td>
<td>1.207  -0.108  0.010  -0.140  0.012  0.017</td>
</tr>
</tbody>
</table>

REFERENCES


APPENDIX

CARD INPUT DESCRIPTION

Card 1

Cols.

3  1 if a multivariate poisson distribution is to be analyzed
2 if a multinomial distribution is to be analyzed
Any other number in this column indicates that the negative multinominal distribution is to be analyzed.

FOR THE MULTIVARIATE POISSON AND NEGATIVE MULTINOMIAL DISTRIBUTIONS

Card 2

1-3 contains the number of variables
4-7 contains the number of observations
7-77 contains 7 pieces of data in consecutive 10-column spaces

Card 3*

1-70 contains 7 pieces of data in consecutive 10-column spaces

FOR THE MULTINOMIAL DISTRIBUTION

Card 2

1-3 contains the number of events
4-74 contains 7 pieces of data in consecutive 10-column spaces

Card 3*

1-70 contains 7 pieces of data in consecutive 10-column spaces
APPENDIX

IMPLICIT REAL*5 (A-H, O-Z)
REAL*8 EL(2), E(I), P(I), AN(I), SUM
CL(1), 15, INF, 03, 14, 13, 35, 12, 12
DO 10 I = 1, K
   DO 10 J = 1, K
10   SUM = SUM + AN(I, J)

20  READ(5,*) ENK = 15) IM
   IF (HK.EQ.2) GO TO 30
   READ(5,*) K = M(I) (NJ(I, J), J = 1, K, I = 1, M)

1 FORMAT(213,17F, J = 1, K)

K = K + 1
   READ(15,*) K = M
   EN = 00
   DO 9 J = 1, M
9      WRITE(b,27) P(I)
   IF (HN.LE.0) GO TO 30
   DO 7 I = 1, K
5      WRITE(b,27) P(I)
7      READ(15,*) SUM = 0
   DO 3 J = 1, M
3      READ(5,*) SUM = SUM + AN(I, J)
      WRITE(b,27) SUM

CALCULATION OF THE NEG. MULTINOMIAL PARAMETERS BEGINS HERE

HN = S2/(S2-S1)
   IF (HN.LE.0) GO TO 30

11 WRITE(b,8) HN
   WRITE (6,26)
   DO 7 I = 1, K
5      WRITE(b,27) P(I)
7      SUM = 0
   DO 3 J = 1, K
3      AN(I, J) = (P(I, J) - EE(I, J))
      IF (AN(I, J).LE.0) GO TO 30
13 SUM = SUM + AN(I, J)
   SUM = SUM/G
   WRITE (6,24) SUM
   WRITE (6,26)
   DO 15 I = 1, K
15      WRITE(b,27) P(I)
   D = AN(I, J)
AE=AN(I,J)
DO 31 I=1, KK
DO 31 J=1, KK
IF (AN(I,J)-D) 33,33,32
C IN THE NEGATIVE MULTINOMIAL DISTRIBUTION
32 Q=AN(I,J)
33 IF (AN(I,J)+AE) 34, J=31
34 AE=AN(I,J)
31 CONTINUE
C D IS NOW THE MAX ESTIMATE OF N AND AE IS THE MIN
DU 62 I=1, K
DO 62 J=1, M
62 SP(I,J)=SPI(I,J)+NJ(J,I)
DU 63 I=1, M
DO 63 J=1, K
63 S(I,J)=S(I,J)+NJ(I,J)
MM=M-
DO 64 I=1, MM, 2
II=II+1
IF (S(I,J)-S(I,J)) 75,75,70
75 Q=Q(I,J)
S(I,J)=S(I,J)-II+OD
DU 64 J=II, M
DO 64 I=II, K
S(I,J)=O
S(I,J)=Q
75 CONTINUE
DU 66 J=1, L
DO 67 I=1, M
IF (S(I,J)-Q) 67, 67, 58
68 F(J)=M-I+1DO
GU TO 66
67 CONTINUE
66 CONTINUE
SH=O
DO 78 I=1, M
78 SH=SH+S(I)
39 DO 35 I=1, L
35 INF(I,J)=INF(I,J)+1.0/(D+I-C)**2*F(I)
SPI=0
DO 36 I=1, K
36 PI(I)=E(I)/0
SPI=SPI+PI(I)
DO 37 I=2, K1
37 INF(I,J)=H/(DL+SPI)
DO 38 I=2, K1
28 INF(I,J)=-(D*M+SH)/(1.0+SPI)**2
38 IF (I.EQ.J) INF(I,J)=INF(I,J)+SPI(J-J)/P(J-1)**2
IF (D.EQ.AE) WRITE (O,44)
IF (D.NE.AE) WRITE (O,43)
CALL ARRAY (Z,K1,K1,INF)
CALL SINV (RM,K1,35,IER)
CALL ARRAY(RM,K1, RM, INF)
DO 23 I=1,K
I=1+1
DO 23 J=1,K
23 INF(I,J)=INF(I,J)
DO 4 1=1,K
41 WRITE (6,42) (INF(I,J),1=1,J)
42 FORMAT(5F14.8/4F14.8)
3 FORMAT('CFISHER'S INFORMATION MATRIX USING THE MAXIMUM ESTIMATE OF CN')
IF (O.EQ.AE) GO TO 28
O=E
44 FORMAT('CFISHER'S INFORMATION MATRIX USING THE MINIMUM ESTIMATE OF CN')
C CALCULATIONS OF MULTIVARIATE POISSON PARAMETERS BEGIN HERE
16 PE=PE**2/6.
17 IF (PE.EQ.0.) GO TO 90
SUM=0.0
WRITE (6,18) PE
WRITE (6,21)
DO 20 I=1,K
111=EI(1)-PE
20 I=DE(0,F+1)
DU .7=1.1,KK
EM(I,J)=EM(I,J)-EI(1,J)
21 IF (EM(I,J).LE.0.1) GO TO 28
17 SUM=SUM+EM(I,J)
19 WRITE (6,19) SUM
WRITE (6,21)
DO 22 I=1,K
211=EI(1)+SUM
22 WRITE (6,27) T(I)
27 FORMAT(31X,F14.8)
19 FORMAT('THE MOMENT ESTIMATE OF THE POISSON PARAMETER FOR U OBTAINED BY AVERAGING ALL POSSIBLE MOMENT ESTIMATES IS*3X,F14.8)'
8 FORMAT('THE MOMENT ESTIMATE OF N OBTAINED BY AVERAGING THE RAW MOMENTS FIRST IS*3X,F14.8)
C CALCULATIONS OF MULTIVARIATE POISSON PARAMETERS BEGIN HERE
16 PE=PE**2/6.
17 IF (PE.EQ.0.) GO TO 90
SUM=0.0
WRITE (6,18) PE
WRITE (6,21)
DO 20 I=1,K
111=EI(1)-PE
20 I=DE(0,F+1)
DU .7=1.1,KK
EM(I,J)=EM(I,J)-EI(1,J)
21 IF (EM(I,J).LE.0.1) GO TO 28
17 SUM=SUM+EM(I,J)
19 WRITE (6,19) SUM
WRITE (6,21)
DO 22 I=1,K
211=EI(1)+SUM
22 WRITE (6,27) T(I)
27 FORMAT(31X,F14.8)
19 FORMAT('THE MOMENT ESTIMATE OF THE POISSON PARAMETER FOR U OBTAINED BY AVERAGING ALL POSSIBLE MOMENT ESTIMATES IS*3X,F14.8)'
8 FORMAT('THE MOMENT ESTIMATE OF N OBTAINED BY AVERAGING THE RAW MOMENTS FIRST IS*3X,F14.8)
C CALCULATIONS OF MULTIVARIATE POISSON PARAMETERS BEGIN HERE
16 PE=PE**2/6.
17 IF (PE.EQ.0.) GO TO 90
SUM=0.0
WRITE (6,18) PE
WRITE (6,21)
DO 20 I=1,K
111=EI(1)-PE
20 I=DE(0,F+1)
DU .7=1.1,KK
EM(I,J)=EM(I,J)-EI(1,J)
21 IF (EM(I,J).LE.0.1) GO TO 28
17 SUM=SUM+EM(I,J)
19 WRITE (6,19) SUM
WRITE (6,21)
DO 22 I=1,K
211=EI(1)+SUM
22 WRITE (6,27) T(I)
27 FORMAT(31X,F14.8)
19 FORMAT('THE MOMENT ESTIMATE OF THE POISSON PARAMETER FOR U OBTAINED BY AVERAGING ALL POSSIBLE MOMENT ESTIMATES IS*3X,F14.8)'
8 FORMAT('THE MOMENT ESTIMATE OF N OBTAINED BY AVERAGING THE RAW MOMENTS FIRST IS*3X,F14.8)
C THE CORRESPONDING VARIANCES ARE')
24 FORMAT('THE CORRESPONDING PROBABILITIES ASSOCIATED WITH THE RESPECTIVE VARIABLES ARE')
26 FORMAT('THE CORRESPONDING PROBABILITIES ASSOCIATED WITH THE RESPECTIVE VARIABLES ARE')
25 FORMAT('THE CORRESPONDING PROBABILITIES ASSOCIATED WITH THE RESPECTIVE VARIABLES ARE')
28 FORMAT('THE CORRESPONDING PROBABILITIES ASSOCIATED WITH THE RESPECTIVE VARIABLES ARE')
C END OF THE PROGRAM
SUBROUTINE ARRAT(MODE,N,RM,INF)
IMPLICIT REAL*8 (A-H,O-Z)
REAL*8 RM(9.), INF(13,3)
IF (MODE-1) 1,3,5,125
100 IJ=0
10 DO 10 LG K=1,N
10 DO 10 LG L=1,K
10 IJ=IJ+1
110 INF(IJ,K)=RM(IJ)
100 GO TO 140
120 IJ=G
12 DO 122 LG K=1,N
12 DO 123 LG L=1,K
12 IJ=IJ+1
125 RM(IJ)=INF(L,K)
140 RETURN
END

SUBROUTINE MFSO(A,N,EPS,E)
IMPLICIT REAL*8 (A-H,O-Z)
DIMENSION A(9,)
IF (N-.2) 2,3,4
1 IER=1
KPIV=1
DO 1 LG K=1,N
KPIV=KPIV+K
IND=KPIV
LEND=K-
TUL=DABS(EPS*4(A(KPIV)))
DO 2 LG L=1,K,N
DSUM=L
IF (L-LEND) 2,3,4
2 DO 2 LG L=1,LEND
LNF=KPIV-L
IND=IND-L
3 DSUM=DSUM+A(LNF)-A(IND)
4 DSUM=DSUM-A(IND)-DSUM
IF (I-K) 5,8,9
5 IF (DSUM-TOL) 1,3,4
6 IF (DSUM) 1,7,4
7 IF (IER) 1,2,9
8 IF (I-K) 5,6,9
9 DPIV=A(SURV(USUM))
A(KPIV)=DPIV
DPIV=DPIV/DPIV
GO TO 5
10 IF (IND=INF*EPS+PIV
11 IND=IND+1
RETURN
12 IER=-1
RETURN
END
SUBROUTINE SINV(A,N,eps,Ier)
  IMPLICIT REAL*8 (a-h.o-z)
  DIMENSION A(19)
  CALL MFSC(A,F,LPS,EK)
  IF (IER) 3,9,9
  1    IPIV=N*(N+1)/2
  INO=IPIV
  DT=1:
  D=
  R=:
  A(IPIV)=DIN
  MIN=N
  KEND=I-1
  LANF=N-KEND
  IF (KEND) 3,9,9
  2    J=IND
  DO 4 K=1,KEND
  WORK=0
  MIN=MIN-1
  LPO=LPIV
  LVER=J
  DO 3 L=LANF,MIN
  LV=LVER+1
  LPO=LPO+L
  WORK=WORK+A(LVEC)-2*(LPO)
  A(J)=WORK*cin
  3    J=J-MIN
  4    IPIV=IPIV-MIN
  INO=INO-1
  DO 9 I=1,N
  IPIV=IPIV+1
  J=IPIV
  DO 8 K=1,N
  WORK=0
  LPO=J
  DO 7 L=K,N
  LVER=LPO+K-1
  WORK=WORK+A(LVEK)-2*(LVER)
  7    LPO=LPO+L
  A(J)=WORK
  8    J=J+K
  9    RETURN
END
A PROCEDURE TO PREDICT CLOUD COVER FREQUENCIES IN THE BIVARIATE CASE

Summary

The purpose of this report is to present a procedure for approximating cloud cover probabilities for two different locations or for the same location at different times. In addition a monte carlo procedure is presented for integrating the bivariate normal distribution. This program is used for computing the approximate probabilities.

If one assumes that the density function for the bivariate cloud cover model is approximately bell-shaped, then it is shown that the desired conditional probabilities can be approximated using the bivariate normal distribution. Examples illustrating the feasibility of this procedure are included. However, if the bivariate density for the cloud cover model is highly J or U shaped this procedure provides results which are less than satisfactory. Examples illustrating this situation are also included.

I. INTRODUCTION

The purpose of this report is to present a procedure for estimating joint probabilities for the degree of cloud cover over two regions or one region at subsequent time intervals.

Falls (1974) demonstrated that the beta distribution adequately describes the variation in the amounts of cloud cover. This conclusion was based upon analysing cloud cover data from diverse locations, for different times of the year and for
different times of the day. Thus, we may expect that the multivariate beta distribution, sometimes called the Dirichlet distribution would be a natural extension for describing the bivariate case. However, a theoretical requirement of the Dirichlet distribution is that the variables be negatively correlated, and this constraint seems to intuitively disagree with the actual situations. Consequently, a different approach was required, one allowing for both positive and negative correlations.

Peizer and Pratt (1968) provide a possible approach, that of using the normal distribution for approximating tail probabilities in the beta distribution. Thus, if one assumes that the correlation between the two sites is structurally related to the correlation present in the bivariate normal distribution, one may be able to extend the work of Peizer and Pratt to the multivariate setting, that of approximating joint probabilities using the bivariate normal distribution (BVN). This approximation would appear to work adequately for those cases where the univariate normal approximation gives satisfactory approximations to the beta distribution.

This report consists of three main sections. The first section describes a program for integrating the BVN over rectangular regions. This section is basically self contained, and it provides the user the needed explanation for integrating the BVN. The second section illustrates how this procedure is used in approximating the bivariate cloud cover model. Applications and examples of this procedure are presented in section 3. The program documentation and listings are presented in the Appendix.
II. BVN PROGRAM

A procedure was required for integrating the bivariate normal distribution over a specified region. The BVN program provides an approximation to the above integral. This section consists of three subsections, 1) introduction to the monte carlo theory, 2) application of this theory to the BVN distribution, 3) examples.

2.1 General Monte Carlo Technique

An excellent summary on the general principles of monte carlo theory can be found in Newman and Odell (1971). The following is a discussion of this method as related to double integration.

Let \( x = (x_1, x_2) \) denote an arbitrary two dimensional vector and \( f(x) \) a real valued function of \( x \). Consider the integral

\[
\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} f(x_i) g(x) \text{d}x_1 \text{d}x_2
\]  

(2.1)

where \( g(x) \) denotes a probability density function on the plane. The integral (2.1) is the expected value of \( f(x) \) and can be estimated by

\[
\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} f(x_i)
\]

where \( x_i, i=1,...,N \) are random samples from the pdf \( g(x) \). The variance of \( \hat{\theta} \), is given by
\[
\text{var}(\hat{\theta}) = \frac{1}{N} \text{var}(f(x)) = \frac{1}{N} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (f(x) - \theta)^2 g(x) \, dx_1 \, dx_2
\]

which can be estimated by

\[
s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (f(x_i) - \bar{\theta})^2.
\]

The estimated standard error is given by, \( \hat{\sigma} = s/\sqrt{n} \).

The following describes a procedure for reducing the magnitude of the \( \text{var}(\hat{\theta}) \). Suppose that there exists a function \( h(x) \) on \( \mathbb{R}^2 \) (two dimensional real) which approximates \( f(x) \) on \( \mathbb{R}^2 \) and suppose that

\[
x = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x) g(x) \, dx_1 \, dx_2
\]

is known. Then

\[
h(x) = x + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (f(x) - h(x)) g(x) \, dx_1 \, dx_2.
\]

The variance of \( f(x) - h(x) \), is given by

\[
\text{var}(f(x) - h(x)) = \text{var}(f(x)) + \text{var}(h(x)) - 2 \text{cov}(f(x), h(x)).
\]

If \( \text{var}(h(x)) < 2 \text{cov}(f(x), h(x)) \), we have that

\[
\text{var}(f(x) - h(x)) < \text{var}(f(x)).
\]

Note that if \( f-h \) and \( h \) are positively correlated then \( \text{var}(f-h) \) is less than \( \text{var}(f) \). This is true since

\[
\text{var}(f) = \text{var}[h + f-h]
\]
\[
= \text{var}(h) + \text{var}(f-h) - 2 \text{cov}(h, f-h)
\]

Thus we have

\[
\text{var}(f-h) = \text{var}(f) - \text{var}(h) - 2 \text{cov}(h, f-h).
\]
Assume the correlation of \((f-h)\) and \(h\) is positive. Hence,
\[
\text{var}(f-h) < \text{var}(f) - \text{var}(h)
\]
which implies that
\[
\text{var}(f-h) < \text{var}(f).
\]
Therefore the larger the correlation of \((f-h)\) and \(h\), the greater the reduction of the variance by removal of the regular part \(h(x)\).

2.2 Program Explanation

The object is to integrate
\[
\theta = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x | \mu, \Sigma) dx_1 dx_2.
\]

where \(\mu = (\mu_1, \mu_2); \Sigma = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}\) and
\[
f(x | \mu, \Sigma) = \text{BWN distribution} =
\frac{1}{2\pi \sigma_1 \sigma_2 (1-\rho^2)^{1/2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left( \frac{x_1-\mu_1}{\sigma_1} \right)^2 - \frac{(x_1-\mu_1)(x_2-\mu_2)}{\rho \sigma_1 \sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right)\}
\]

In formula (2.1) we define \(g(x) = \frac{1}{(b_1-a_1)(b_2-a_2)}\), i.e.
g\((x)\) represents a bivariate uniform distribution, and evaluate the integral
\[
\theta = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x | \mu, \Sigma) \frac{dx_1 dx_2}{(b_1-a_1)(b_2-a_2)}.
\]
It follows that
\[ e = \emptyset (b_1-a_1)(b_2-a_2) \]
and the estimate
\[ \hat{e} = \hat{\emptyset} (b_1-a_1)(b_2-a_2) \]
where
\[ \emptyset = \frac{1}{N} \sum_{i=1}^{N} f(x_i | u, l) \]
when \( x_i \) is a random vector from the pdf
\[ g(x) = \begin{cases} \frac{1}{(b_1-a_1)(b_2-a_2)} & ; a_j \leq x_j \leq b_j \text{, } j=1, 2 \\ 0 & ; \text{Otherwise} \end{cases} \]
Since \( g(x) \) is the product of two independent uniform distributions, a random vector is generated using the equations \( x_j = a_j + u_j(b_j - a_j), \)
\( j=1,2 \) where \( u_j \) is distributed uniform over the interval \((0,1)\).
In the BVII program the regular part \( h(x) \) is defined to be all the terms up to the coefficient \( 1/8! \) in the two dimensional Taylor's expansion (Fulks 1969, p. 260). The two dimensional Taylor's expansion about the point \((a_1, a_2)\) is given by

\[
f(x_1, x_2) = f(a_1, a_2) + (x_1-a_1) \frac{\partial f}{\partial x_1} (a_1, a_2)
+ (x_2-a_2) \frac{\partial f}{\partial x_2} (a_1, a_2) + \frac{1}{2!} \left[ (x_1-a_1)^2 \frac{\partial^2 f}{\partial x_1^2} (a_1, a_2)
+ 2(x_1-a_1)(x_2-a_2) \frac{\partial^2 f}{\partial x_1 \partial x_2} (a_1, a_2) + (x_2-a_2)^2 \frac{\partial^2 f}{\partial x_2^2} (a_1, a_2) \right]
+ \frac{1}{3!} \left[ (x_1-a_1)^3 \frac{\partial^3 f}{\partial x_1^3} (a_1, a_2)
+ 3(x_1-a_1)^2 (x_2-a_2) \frac{\partial^3 f}{\partial x_1^2 \partial x_2} (a_1, a_2)
+ 3(x_1-a_1)(x_2-a_2)^2 \frac{\partial^3 f}{\partial x_1 \partial x_2^2} (a_1, a_2)
+ (x_2-a_2)^3 \frac{\partial^3 f}{\partial x_2^3} (a_1, a_2) \right] + \ldots \]

\[(2.4)\]
Hence it was necessary to find all partials (up to 8th order) of the BVN distribution function, $f(x_1,x_2)$.

Let $(a_1,a_2) = (\mu_1, \mu_2)$ the mean vector of the BVN distribution. Then equation (2.4) becomes

$$\frac{\partial^2 f}{\partial x_1^2} (u_1,u_2) = f(x_1,x_2) \left[ \frac{1}{-2(1-p^2)} \left( \frac{2}{\sigma_1^2} (x_1 - \mu_1) \right)^2 \right]$$

$$- \left. \frac{2 \rho}{\sigma_1 \sigma_2} (x_2 - \mu_2) \right| \begin{align*}
  x_1 &= \mu_1 \\
  x_2 &= \mu_2
\end{align*}$$

$$\frac{\partial^2 f}{\partial x_2^2} (u_1,u_2) = \frac{-1}{2 \pi \sigma_1 \sigma_2 (1-p^2)^{3/2}}$$

$$\frac{\partial^4 f}{\partial x_1^4} = \frac{\rho}{2 \pi \sigma_1^2 \sigma_2^2 (1-p^2)^{3/2}}$$

$$\frac{\partial^4 f}{\partial x_2^4} = \frac{-1}{2 \pi \sigma_1 \sigma_2^3 (1-p^2)^{3/2}}$$

$$\frac{\partial^4 f}{\partial x_1^2 \partial x_2^2} = \frac{3}{2 \pi \sigma_1^5 \sigma_2 (1-p^2)^{5/2}}$$

$$\frac{\partial^4 f}{\partial x_2^3 \partial x_1} = \frac{-3 \rho}{2 \pi \sigma_1^4 \sigma_2^2 (1-p^2)^{5/2}}$$
\[
\frac{\partial^4 f}{\partial x_1^2 \partial x_2^2} = \frac{2\rho^2 + 1}{2 \pi \sigma_1^3 \sigma_2^3 (1 - \rho^2)^{5/2}} \\
\frac{\partial^4 f}{\partial x_2^3 \partial x_1} = \frac{-3\rho}{2 \pi \sigma_1^5 \sigma_2^4 (1 - \rho^2)^{5/2}} \\
\frac{\partial^4 f}{\partial x_2^4} = \frac{3}{2 \pi \sigma_1^5 \sigma_2^5 (1 - \rho^2)^{5/2}} \\
\frac{\partial^6 f}{\partial x_1^6} = \frac{-15}{2 \pi \sigma_1^7 \sigma_2^5 (1 - \rho^2)^{7/2}} \\
\frac{\partial^6 f}{\partial x_1^5 \partial x_2} = \frac{15\rho}{2 \pi \sigma_1^7 \sigma_2^6 (1 - \rho^2)^{7/2}} \\
\frac{\partial^6 f}{\partial x_1^4 \partial x_2^2} = \frac{-3 - 12\rho^2}{2 \pi \sigma_1^7 \sigma_2^5 (1 - \rho^2)^{7/2}} \\
\frac{\partial^6 f}{\partial x_1^3 \partial x_2^3} = \frac{9\rho + 6\rho^3}{2 \pi \sigma_1^7 \sigma_2^4 (1 - \rho^2)^{7/2}} \\
\frac{\partial^6 f}{\partial x_1^2 \partial x_2^4} = \frac{-3 - 12\rho^2}{2 \pi \sigma_1^7 \sigma_2^5 (1 - \rho^2)^{7/2}} \\
\frac{\partial^6 f}{\partial x_1 \partial x_2^5} = \frac{15\rho}{2 \pi \sigma_1^7 \sigma_2^6 (1 - \rho^2)^{7/2}}
\]
\[
\begin{align*}
\frac{a_6 f}{a x_2^5} &= \frac{-15}{2 \sigma_1^2 c_2^7 (1-c^2)^{7/2}} \\
\frac{a_8 f}{a x_1^8} &= \frac{105}{2 \sigma_1^9 \sigma_2^8 (1-c^2)^{9/2}} \\
\frac{a_8 f}{a x_1^7 a x_2} &= \frac{-105 \rho}{2 \sigma_1^8 \sigma_2^2 (1-c^2)^{9/2}} \\
\frac{a_8 f}{a x_1^6 a x_2^2} &= \frac{15 \cdot 9 \sigma_2^2}{2 \sigma_1^7 \sigma_2^3 (1-c^2)^{9/2}} \\
\frac{a_8 f}{a x_1^5 a x_2^3} &= \frac{-45 \sigma_2^3 - 60 \sigma_3^3}{2 \sigma_1^6 \sigma_2^4 (1-c^2)^{9/2}} \\
\frac{a_8 f}{a x_1^4 a x_2^4} &= \frac{72 \sigma_2^4 + 24 \sigma_4^4}{2 \sigma_1^5 \sigma_2^5 (1-c^2)^{9/2}} \\
\frac{a_8 f}{a x_1^3 a x_2^5} &= \frac{-45 \sigma_2^3 - 60 \sigma_3^3}{2 \sigma_1^4 \sigma_2^6 (1-c^2)^{9/2}} \\
\frac{a_8 f}{a x_1^2 a x_2^6} &= \frac{15 \cdot 9 \sigma_2^2}{2 \sigma_1^3 \sigma_2^7 (1-c^2)^{9/2}} \\
\frac{a_8 f}{a x_1 a x_2^7} &= \frac{-105 \rho}{2 \sigma_1^2 \sigma_2^8 (1-c^2)^{9/2}}
\end{align*}
\]
\[
\frac{\partial^8 f}{\partial x_2^8} = \frac{105}{2\pi a_1 a_2^9(1-\rho^2)^{9/2}}
\]

However, since all odd ordered partials of the BVN distribution evaluated at the mean are zero, equation (2.4) can be simplified as follows

\[
f(x_1, x_2) = \frac{1}{2\pi \sigma_1 \sigma_2 (1-\rho^2)^{3/2}} - \frac{1}{2} (x_1 - \mu_1)^2 \frac{1}{2\pi \sigma_1^3 \sigma_2 (1-\rho^2)^{3/2}} + \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{2\pi \sigma_1^2 \sigma_2^2 (1-\rho^2)^{3/2}} - \frac{1}{2} (x_2 - \mu_2)^2 \frac{1}{2\pi \sigma_1^3 \sigma_2^3 (1-\rho^2)^{3/2}}
\]

\[+ \frac{1}{2^4} (x_1 - \mu_1)^4 \frac{3}{2\pi \sigma_1^5 \sigma_2^2 (1-\rho^2)^{5/2}}
\]

\[- \frac{1}{6} (x_1 - \mu_1)^3 (x_2 - \mu_2) \frac{3\rho}{2\pi \sigma_1^4 \sigma_2^2 (1-\rho^2)^{5/2}} + \ldots \quad (2.5)
\]

From equation (2.5) we observe that \( \| f(x_1, x_2) - h(x_1, x_2) \| \) becomes large as \((x_1, x_2)\) deviates from \((\mu_1, \mu_2)\), where \( h(x_1, x_2) \) are the first 25 terms in (2.5) and \( \| \cdot \| \) is some distance function. For this reason

\[
\int_A f(x) - h(x) g(x) \, dx
\]

may not be bounded, especially for large region \( A \). However, if the regular part \( h(x) \) is not removed, the convergence would be very slow. To accelerate the convergence and allow for
integration over large regions, the BVN program divides the original integration region into four rectangular regions and integrates each region separately. The program divides the four regions as follows.

Let $L_1 \leq x_1 \leq u_1$ and $L_2 \leq x_2 \leq u_2$ be the integration region. When divided into the four desired regions this becomes

\[
\begin{array}{|c|c|}
\hline
(L_1, L_2) & (u_1, u_2) \\
3 & 4 \\
1 & 2 \\
\hline
\end{array}
\]

Region 1 limits are $L_1 \leq x_1 \leq \frac{L_1 + u_1}{2}$; $L_2 \leq x_2 \leq \frac{L_2 + u_2}{2}$.

Region 2 limits are $\frac{L_1 + u_1}{2} \leq x_1 \leq u_1$; $L_2 \leq x_2 \leq \frac{L_2 + u_2}{2}$.

Region 3 limits are $L_1 \leq x_1 \leq \frac{L_1 + u_1}{2}$; $\frac{L_2 + u_2}{2} \leq x_2 \leq u_2$.

Region 4 limits are $\frac{L_1 + u_1}{2} \leq x_1 \leq u_1$; $\frac{L_2 + u_2}{2} \leq x_2 \leq u_2$.

After obtaining the approximate integral for each region the results are then added together for the final answer. The final standard error is computed as the average of the standard errors corresponding to the four regions.

Since it is difficult to determine if $\text{var}(h) < 2 \cdot \text{cov}(f,h)$, the BVN program is currently set up to integrate both the BVN
function and the BVN function after extraction of the regular part. Convergence is currently checked by computing the estimated standard error of \( \hat{\theta} \) after every 1000 random samples.

There are six input items. These are the means, \((\mu_1, \mu_2)\), the standard deviations, \(\sigma_1, \sigma_2\), the correlation \(\rho\), the maximum standard error, starting value for random number generation (odd integer 15), and the limits of integration. The estimates for each of the four regions are outputed along with their estimated standard error. If the regular part is removed, the correlation between \(f-h\) and \(h\) is output. The output also indicates whether or not the regular part has been removed. Finally, the sum of the values obtained by integrating over each of the four regions is displayed as the final answer.

2.3 Specific Examples

This section presents the output of four examples along with the correct answers Pearson (1931). The four integrals chosen are

1. \[ \int_0^\infty \int_0^\infty f(x_1, x_2, \xi) dx_1 dx_2 \]
   \[ \text{where } \xi = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix} \]

2. \[ \int_{-\infty}^\infty \int_{-\infty}^\infty f(x_1, x_2, \xi) dx_1 dx_2 \]
   \[ \text{where } \xi = \begin{bmatrix} 1 & -.5 \\ -.5 & 1 \end{bmatrix} \]
3. \[ \int_0^1 \int_0^{a(x, t)} f(x, t) \, dx_1 \, dx_2 \]

where \( \mathbf{t} = \begin{bmatrix} 1 & -0.75 \\ -0.75 & 1 \end{bmatrix} \)

4. \[ \int_{\frac{1}{2}}^1 \int_1^{a(x, t)} f(x, t) \, dx_1 \, dx_2 \]

where \( \mathbf{t} = \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix} \)

The results of the BVN program are given in the Tables (1-4).
The respective means are 0.0 0.0
The respective standard deviations are 1.00000000 1.00000000
The correlation is 0.50000000
The maximum error allowed is 0.00300000
The upper bounds are 4.00000000 4.00000000
The lower bounds are 0.0 0.0

An approximation for the 1 region
The value is 0.2930342318 with a standard error of 0.002139714
And a correlation of 0.6115916511
The regular part is positively correlated with the integral and thus extracted.

An approximation for the 2 region
The value is 0.0161355461 with standard error of 0.0006924657
The regular part is not removed.

An approximation for the 3 region
The value is 0.0164091896 with standard error of 0.0007069048
The regular part is not removed.

An approximation for the 4 region
The value is 0.004017887 with standard error of 0.0002146251
The regular part is not removed.

The total probability is 0.32963076 with a standard error of 0.00103199

The correct answer is .33333

Table 1.
THE RESPECTIVE MEANS ARE 0.0 0.0
THE RESPECTIVE STANDARD DEVIATIONS ARE 1.00000000 1.00000000
THE CORRELATION IS -0.50000000
THE MAXIMUM ERROR ALLOWED IS 0.00300000
THE UPPER BOUNDS ARE 4.00000000 4.00000000
THE LOWER BOUNDS ARE 0.50000000 1.00000000

AN APPROXIMATION FOR THE 1 REGION
THE VALUE IS 0.0111994202 WITH STANDARD ERROR OF 0.0006024142
THE REGULAR PART IS NOT REMOVED

AN APPROXIMATION FOR THE 2 REGION
THE VALUE IS 0.0000608119 WITH STANDARD ERROR OF 0.0000054074
THE REGULAR PART IS NOT REMOVED

AN APPROXIMATION FOR THE 3 REGION
THE VALUE IS 0.0000904058 WITH STANDARD ERROR OF 0.0000072492
THE REGULAR PART IS NOT REMOVED

AN APPROXIMATION FOR THE 4 REGION
THE VALUE IS 0.00000000995 WITH STANDARD ERROR OF 0.00000000122
THE REGULAR PART IS NOT REMOVED

THE TOTAL PROBABILITY IS 0.01135074
WITH A STANDARD ERROR OF 0.00015377

The correct answer is .012447

TABLE 2.
<table>
<thead>
<tr>
<th>Region</th>
<th>Value</th>
<th>Standard Error</th>
<th>Regular Part Removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1118712551</td>
<td>0.0028780424</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>0.0001379567</td>
<td>0.0000210493</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>0.0001607447</td>
<td>0.0000219862</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>0.0000000005</td>
<td>0.0000000001</td>
<td>No</td>
</tr>
</tbody>
</table>

The total probability is 0.11216996 with a standard error of 0.00073027.

The correct answer is .115027.

TABLE 3.
THE RESPECTIVE MEANS ARE

0.0  0.0

THE RESPECTIVE STANDARD DEVIATIONS ARE

1.00000000  1.00000000

THE CORRELATION IS

0.75000000

THE MAXIMUM ERROR ALLOWED IS

0.00300000

THE UPPER BOUNDS ARE

4.00000000  4.00000000

THE LOWER BOUNDS ARE

0.50000000  1.00000000

AN APPROXIMATION FOR THE 1 REGION

THE VALUE IS  0.1133274387 WITH STANDARD ERROR OF  0.0027673633
THE REGULAR PART IS NOT REMOVED

AN APPROXIMATION FOR THE 2 REGION

THE VALUE IS  0.0084165793 WITH STANDARD ERROR OF  0.0003595563
THE REGULAR PART IS NOT REMOVED

AN APPROXIMATION FOR THE 3 REGION

THE VALUE IS  0.0033200334 WITH STANDARD ERROR OF  0.0001715015
THE REGULAR PART IS NOT REMOVED

AN APPROXIMATION FOR THE 4 REGION

THE VALUE IS  0.0027903045 WITH STANDARD ERROR OF  0.0001249913
THE REGULAR PART IS NOT REMOVED

THE TOTAL PROBABILITY IS  0.12785436
WITH A STANDARD ERROR OF  0.00085585

The correct answer is .128133

TABLE 4.
III. APPROXIMATION

The introduction briefly presented the reason why the Dirchlet distribution was not applicable in the multivariate case. As the beta distribution seemed firmly established as a proper model in the univariate case, it seemed more reasonable to build a prediction process utilizing the beta distribution than to seek a new model applicable to both univariate and multivariate cases. This led to the BVN distribution.

The reason why the Dirchlet would not work was the theoretical requirement of a negative covariance between the variables—a situation not frequently encountered in most applications. However, the BVN distribution imposes fewer constraints on the value of the covariance. Also, the normal distribution has been shown to yield excellent approximations for "tail" probabilities in the univariate beta case (See Peizer and Pratt, 1968, pg. 1418). Also, the normal approximation exists for the beta probabilities over any interval. If the covariance (or correlation) is thought of as effecting an increase or decrease in probabilities (compared with uncorrelated probabilities) rather than depicting the underlying association between the variables, then one should be able to determine this effect using either the approximations to the beta probabilities or
the beta probabilities themselves. The only reason why a bivariate model is required is because we know cloud cover frequencies at the sites are related. Otherwise an assumption of independence would allow one to compute the joint probabilities via a direct multiplication of the univariate beta probabilities.

Finally, it is important to stress that the BVN, as we use it, is only a mechanism to calculate probabilities. In conversations with MSFC personnel it was noted that some persons in the meteorological profession had proposed the normal distribution as a model to describe cloud cover frequencies. Such a model may or may not be plausible and we did not investigate it. The beta model serves as the basis for our analysis, i.e., we assume the beta model fits the data--all we must do is calculate the parameters. Falls (1973) did encounter months, time intervals and sites where the beta model was not a good fit. It would be proper to preface all our remarks and, indeed, the whole report with the condition that the beta distribution must yield a good fit on the data at hand. However, it is also proper to assert, based on proper evidence, that the beta model is always adequate, at least for the purposes envisioned. The result is the same--situations where the results obtained from applying the model differ substantially from empirical results.
3.1 Normal Approximation to the Beta Distribution

Peizer and Pratt (1968) show that the tail probabilities for a wide range of distributions can be approximated using a normal distribution. Much of the article is not germane to our discussion and will not be discussed. However, it is informative to trace their procedure for approximating the univariate beta distribution.

The density function for the beta distribution is given by

\[ h(x; a, \beta) = \frac{\Gamma(a + \beta)}{\Gamma(a) \Gamma(\beta)} x^{a-1} (1-x)^{\beta-1}; 0 < x < 1, a, \beta > 0. \]  

(3.1)

To approximate the probability that \( 0 < x < x_0 \), i.e.

\[ \Pr \{ x \leq x_0 \} = \int_0^{x_0} h(x; a, \beta) \, dx \]

calculate the quantities

\[ d_1 = (a + \beta - 2/3) x_0 - (a - 1/3) \]

\[ d_2 = d_1 + .02 \left( \frac{x_0}{\beta} - \frac{1-x_0}{a} + \frac{x_0 - \beta}{a + \beta} \right), \]

and

\[ z = \frac{d_2}{|b - 0.5(a + \beta - 1)(1-x_0)|} \left( \frac{12(a + \beta - 1)}{6(a + \beta - 1)} \left( \frac{\gamma - 0.5}{\gamma} \right) \left( \frac{\gamma - 0.5}{\gamma} \right) \right) \]

(3.2)

The approximate probability is given by

\[ P = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \, dy. \]
Of course, should you desire to have a right tail probability.

The approximate value for the right tail probability is

\[ P = \int_{z}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy. \]

The error in these approximations is less than .01 if \( a, \beta > 1 \) and less than .001 if \( a, \beta > 2 \). It also follows that

\[ \Pr \{ x_0 < x < x_1 \} \text{ can be approximated as} \]

\[ \Pr \{ x_0 < x < x_1 \} = \int_{x_0}^{x_1} h(x; a, \beta) dx = 1 - \int_{x_0}^{x} h(x; a, \beta) dx - \int_{x}^{x_1} h(x; a, \beta) dx \]

or

\[ 1 - \int_{-\infty}^{z_0} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy - \int_{z_1}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \int_{z_0}^{z_1} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy. \]

However, the error is potentially doubled for this case.

The approximation is not valid for \( a, \beta \leq 0.5 \) which implies the data must be highly U-shaped for the approximation to fail. This could further restrict the applicability to some locations and for some seasons. However, Falls has shown that this situation is infrequent.

3.2 The Bivariate Case

Assuming that \( x \) and \( y \) are beta distributed,

\[ x_0 < x < x_1, \quad y_0 < y < y_1 \]

can be approximated by

\[ \int_{z_0}^{z_1} \int_{z_{y_0}}^{z_{y_1}} f(z_x, z_y) dz_x dz_y. \]

where \( f(z_x, z_y) \) is the BVN distribution defined in equation (2.2).
IV. THE APPROXIMATION PROGRAM

In order to use the BVN approximation, a computer program was developed to convert raw data and desired beta intervals into the z-values and correlations (BVN program inputs). This program takes raw data and calculates means, variances, correlations and estimated beta parameters for both raw and categorized data. Then for each inputed beta interval value (lower and upper values for each variate) it calculates a corresponding z-value.

Two aspects of the program need explanation. The formulas in Section 3.1 are not defined for the beta values of 0 or 1. Consequently, the program cannot handle such values. For this reason, 0 or 1 must be inputed as $0 + \varepsilon$ or $1 - \varepsilon$ where $\varepsilon > 0$ is some arbitrary real number. Likewise $-4$ is used for $-\infty$, $+4$ for $+\infty$ in the BVN program.

Since the approximation fails if $\alpha, \beta \leq .5$, the program resets the parameters to .51 and prints a notice to the user if the estimated beta parameter value falls below .5. It is then left to the user to decide whether or not he wants to use this acknowledged poor approximation.

The beta parameters are estimated using the method of moments as described by Hahn and Shapiro (1967, pg. 95). The estimated beta parameters for the original data are

$$B = \frac{(1-\bar{X})}{s^2} \left[ \bar{X}(1-\bar{X}) - s^2 \right]$$

$$A = \frac{\bar{X}B}{1-\bar{X}}$$
where $\bar{X}$ and $S^2$ are the sample mean and variance.

A frequency table for both original and category data is given in order to compute the empirical probabilities which are used to check the corresponding approximate BVN probabilities.

V. DATA

The data used in this study was compiled by ESSA, National Weather Records Center, Asheville, North Carolina and was provided to the authors by Organization ES-42, Marshall Space Flight Center, Alabama. The sites selected were Fort Worth and Houston, Texas. Daily records (January 1971 to December 1975) on cloud cover, measured in tenths, were recorded every third hour.

The data was grouped into the categories shown in Table 5 (Fall 1973).

Table 5
Cloud Cover Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Tenths</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1,2,3</td>
</tr>
<tr>
<td>3</td>
<td>4,5</td>
</tr>
<tr>
<td>4</td>
<td>6,7,8,9</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

Since Falls (1971) demonstrated that the beta distribution adequately describes variation in categorical data, our primary investigation was restricted to categorical data. However, the approximation program is not restricted to categorical data.
VI. EXAMPLES

A complete set of probabilities (25 values) have been calculated for the Fort Worth 9 a.m. and Fort Worth 3 p.m. combination. These values are presented in Figure 1. Each of the five portions of figure represents a category level for 9 a.m. and the abseissas represent the categories for 3 p.m. Table 6 presents a portion of the approximation program and Table 7 gives the corresponding BVN computations.

Figure 1 values were determined based on observed and expected frequencies for 5 years (155 values). As can be noted, the agreement is quite satisfactory with a couple of exceptions. Values for Category 1 for 9 a.m. and Category 2 for 3 p.m. shows a wide divergence. Also the five values predicted for 3 p.m. and Category 4 for 9 a.m. show substantial disagreement.

These discrepancies between observed and predicted values can be explained by analyzing how well the beta model describes univariate cloud cover in the various data sets.

From Table 6 the category frequencies for Site 1 (9 a.m.) are 46, 29, 20, 39, 21 respectively and the estimated beta parameters are .862646 and 1.06241. These parameters are for a very U-shaped density which decreases as x → 1. Consequently, the fitted distribution does not reflect the variation in these
Categories - 3 p.m.

**CATEGORY 1**

**CATEGORY 2**

**CATEGORY 3**

**CATEGORY 4**

**CATEGORY 5**

9 a.m. Cloud Cover Mean = .449387
Standard Deviation = .290846
3 p.m. Cloud Cover Mean = .510323
Standard Deviation = .240997
Correlation = .570333

Beta Parameters for 9 a.m. are .862646  1.061241
Beta Parameters for 3 p.m. are 1.685583  1.617392

FIGURE 1.

OBSERVED AND PREDICTED FREQUENCIES FOR FORT WORTH AT 9 A.M.
AND FORT WORTH AT 3 P.M. BASED ON JULY DATA FOR 1971-75.
data for categories 1 and 4, which is reflected in the approximate probability.

Some additional comments are necessary. First it is important to note that we have only 155 data points and more data would, in most such cases, give better fit to the true distribution hence a better approximation. Secondly, this problem is not restricted to this one isolated case. Based upon our analyses, we feel that the substantial disagreement between observed and predicted probabilities were based upon the inadequacy of the beta distribution. It does not seem likely that large errors will occur because of this condition but if the parameter values are low the approximation error could contribute substantially to the disagreement between the values. Thirdly, it must be noted that Figure 1 is based upon integration limits (determined by the transformation from categories to the (0,1) interval) that should give the best results. The category values 1, 2, 3, 4, 5 are transformed to .1, .3, .5, .7, .9 respectively. The corresponding limits of integration are found in Table 6.

<table>
<thead>
<tr>
<th>Category</th>
<th>Integration Limits</th>
<th>Midpoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.01 to .2</td>
<td>.1</td>
</tr>
<tr>
<td>2</td>
<td>.2 to .4</td>
<td>.3</td>
</tr>
<tr>
<td>3</td>
<td>.4 to .6</td>
<td>.5</td>
</tr>
<tr>
<td>4</td>
<td>.6 to .8</td>
<td>.7</td>
</tr>
<tr>
<td>5</td>
<td>.8 to .99</td>
<td>.9</td>
</tr>
</tbody>
</table>
The values in Table 6 are the usual "continuity" corrections for approximating probabilities for discrete variables. It must be noted that the intervals selected will not always reflect the underlying situation and hence could contribute to the differences in values. However, if the above limits are a source of error then its effect will be minor compared with the other errors and its effect will decrease over wider intervals.

As noted, we have elected to use categorical data throughout the analyses. However, one might consider using the original data in that the beta model might actually fit whereas the categorical fit was inadequate. Another reason for using the original data is the greater flexibility in selecting the integration limits which can be made to closely agree with the original situation (cloud cover measured in tenths).

6.2 Application of the Programs

The approximation programs must be run to obtain the approximate integration limits used in integrating the BVN distribution. The input needed for this program consists of two parts. The first part consists of the raw data (read pairwise with the first value corresponding to the first site and the second value corresponding to the second site or the data can represent one site at two different times). The second part consists of the inputed boundary numbers for the
regions to be integrated. Before continuing one should inspect the outputed beta parameters and corresponding frequency tables. If the estimated beta parameters are significantly less than .5, then one must proceed with caution since the calculated integration limits are probably unreliable (for reason explained previously).

The outputed correlations and the integration limits are then used as inputs into the BVN program. Note that since the approximated integration limits pertain only to the standard normal distribution, the mean vector will be (0,0) and the standard deviation will be (1,1). The main output of the BVN program is the total probability. This value represents the approximate probability of a specified category or categories at Site 1 intersected with a specified category or categories at site 2.

For example, Table 7 lists the output of the approximation program for the percent of cloud cover over Fort Worth, Texas, at 9 a.m. and 3 p.m. during the month of July (1971-1975). Since the beta parameters for the original data is significantly less than .5, we decided to work with the category data. The category data z's are the approximate integration limits corresponding to category 1 at 9 a.m. and category 1 at 3 p.m. These values were then used as input for the BVN program along with the correlation of .57. The output of the BVN program is found in Table 8. The total probability of having cloud cover in category 1, (i.e.
essentially no cloud cover) at 9 a.m. and of having cloud cover in category 1 at 3 p.m. during July at Fort Worth is shown to be approximately .063. Whereas the empirical value, found in the category frequency table, is $\frac{10}{155} = .0645$. 
***** RESULTS USING ORIGINAL DATA *****

FREQUENCY TABLE: 155 VALUES

<table>
<thead>
<tr>
<th>10</th>
<th>3</th>
<th>9</th>
<th>11</th>
<th>9</th>
<th>1</th>
<th>0</th>
<th>3</th>
<th>3</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

MEANS = 0.4151 0.5065
ST. DEV = 0.3742 0.3162
CORR = 0.6411

ESTIMATED BETA PARAMETERS

SITE I  0.2849  0.7406
SITE II  0.7603  0.7406

***** RESULTS USING CATEGORICAL DATA *****

FREQUENCY TABLE: 155 VALUES

<table>
<thead>
<tr>
<th>10</th>
<th>23</th>
<th>10</th>
<th>3</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>7</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>4</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>8</td>
</tr>
</tbody>
</table>

MEANS = 0.4434 0.5103
ST. DEV = 0.2308 0.2410
CORR = 0.5759

SITE I  0.4626  1.0612
SITE II  1.6855  1.5174

***** NOTE *****
IF A PARAMETER OR PARAMETERS IS LESS THAN OR EQUAL TO .5, THE Z-VALUE IS UNDEFINED. FOR FURTHER COMPUTATION THE PARAMETER IS RESET TO .51

FIRST SITE

<table>
<thead>
<tr>
<th>INTEGRAL LIMITS</th>
<th>CATEGORY DATA 2:5</th>
<th>ORIGINAL DATA 2:5</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPPER LOWER</td>
<td>0.8000000000</td>
<td>0.2000000000</td>
</tr>
<tr>
<td>CATEGORY DATA 2:5</td>
<td>1.024939124</td>
<td>-0.649395533</td>
</tr>
<tr>
<td>ORIGINAL DATA 2:5</td>
<td>0.59792524</td>
<td>-0.697925290</td>
</tr>
</tbody>
</table>

TABLE 7
**INPUT PARAMETERS**

<table>
<thead>
<tr>
<th>MEANS</th>
<th>ST. DEV</th>
<th>CORRELATION</th>
<th>MAX ERROR</th>
<th>UPPER BOUNDS</th>
<th>LOWER BOUNDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>1.0000</td>
<td>0.5709</td>
<td>0.0300</td>
<td>1.3241</td>
<td>-0.6494</td>
</tr>
</tbody>
</table>

***** APPROX. FOR REGION NO. = 1 *****

+++ THE REGULAR PART HAS BEEN REMOVED
THE VALUE IS 0.13776
ST. ERR = 0.00001
CORR = 0.56849

***** APPROX. FOR REGION NO. = 2 *****

+++ THE REGULAR PART HAS BEEN REMOVED
THE VALUE IS 0.07902
ST. ERR = 0.00063
CORR = 0.55641

***** APPROX. FOR REGION NO. = 3 *****

+++ THE REGULAR PART HAS BEEN REMOVED
THE VALUE IS 0.12317
ST. ERR = 0.00006
CORR = 0.56100

***** APPROX. FOR REGION NO. = 4 *****

+++ THE REGULAR PART HAS BEEN REMOVED
THE VALUE IS 0.13429
ST. ERR = 0.00003
CORR = 0.74074

THE TOTAL PROBABILITY IS 0.47424 WITH A STANDARD ERROR OF 0.00018

---

TABLE 8
REFERENCES


APPENDICES

Appendix A gives a description of the card inputs followed by a listing of the BVN program. Appendix B gives a similar listing for the approximation program. Both programs are written in Fortran and, the approximation program can compute 100 individual integration limits in less than a minute (on IBM 370/155). The BVN program also takes less than a minute to calculate one total probability.
### APPENDIX A

**BVN Program - Card Input**

<table>
<thead>
<tr>
<th>Card</th>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-14</td>
<td>mean of the first variable of the BVN distribution</td>
</tr>
<tr>
<td></td>
<td>15-28</td>
<td>mean of the second variable</td>
</tr>
<tr>
<td>2</td>
<td>1-14</td>
<td>standard deviation of the first variable</td>
</tr>
<tr>
<td></td>
<td>15-28</td>
<td>standard deviation of the second variable</td>
</tr>
<tr>
<td></td>
<td>29-42</td>
<td>correlation between the two variables</td>
</tr>
<tr>
<td>3</td>
<td>1-14</td>
<td>maximum standard error allowed</td>
</tr>
<tr>
<td></td>
<td>21-25</td>
<td>odd, five digit random integer</td>
</tr>
<tr>
<td>4</td>
<td>1-14</td>
<td>upper integration limit for the first variable</td>
</tr>
<tr>
<td></td>
<td>15-28</td>
<td>lower integration limit for the first variable</td>
</tr>
<tr>
<td>5</td>
<td>1-14</td>
<td>upper integration limit for the second variable</td>
</tr>
<tr>
<td></td>
<td>15-28</td>
<td>lower integration limit for the second variable</td>
</tr>
</tbody>
</table>
CARD COLS. VARIABLE

1 1-14 MEAN AT FIRST SITE, REAL NUMBER
1 15-28 MEAN AT SECOND SITE, REAL NUMBER
2 1-14 STD. DEV. AT FIRST SITE, REAL NUMBER
2 15-28 STD. DEV. AT SECOND SITE, REAL NUMBER
3 1-14 ERROR BOUND
3 21-25 INTEGER RANDOM NUMBER
4 1-14 UPPER BOUND AT FIRST SITE
4 15-28 LOWER BOUND AT FIRST SITE
5 1-14 UPPER BOUND AT SECOND SITE
5 15-28 LOWER BOUND AT SECOND SITE

IMPLICIT REAL*8(A-H,O-Z)
COMMON M,M2,KJ, SIG1, SIG2,
REAL*8 C(1/3*0.,/X(4.), NT, M1, M2, M(8,9), Q(4)/4*0./, LUI(4), UP1(4), LJP
c(4), UP2(4), L1, L2
INTEGER RAND
15 READ(5,1,E=100) M1, M2, SIG1, SIG2, RJ, ERR, RAND, U1, L1, U2, L2
1 FORMAT(2F14.8,13F14.8,15/2F14.8/2F14.8)
WRITE(5,200)
700 FORMAT (*) INPUT PARAMETERS,:///)
WRITE(5,2) M1, M2, SIG1, SIG2, RJ
2 FORMAT (*) MEAN=,13X,2F10.4,///
* ST. DEV=,13X,2F10.4,///
* CORRELATION=,14X,F10.4)
WRITE(5,3) ERR
3 FORMAT (*) MAX ERROR=,16X,F10.4)
WRITE(5,4) U1, U2, L1, L2
4 FORMAT (*) UP BND=,13X,2F10.4,///
* LOWER BND=,13X,2F10.4,///
UP1(1)=U1
UP1(2)=U1
UP1(3)=U1
UP1(4)=U1
UP2(1)=U2
UP2(2)=U2
UP2(3)=U2
UP2(4)=U2
L01(1)=L1
L01(2)=L1
L01(3)=L1
L01(4)=L1
L02(1)=L2
L02(2)=L2
L02(3)=L2
L02(4)=L2
B12(1)=1.000
B12(2)=1.000
B12(3)=1.000
B12(4)=300
B4(1)=300
B4(2)=300
B4(3)=300
B4(4)=300
B4(5)=300
B(6,7)=150
B(6,2)=1500*RO
9(1,5)=1500*RO
9(1,3)=300-L200*RJ**2
B(1,1)=300+300*RJ**2
B(1,1)=300-L200*RJ**2
B(1,1)=1050
B(1,2)=(-10500)*KJ
B(1,3)=1500+QDO*RO**2
B(1,4)=4500*RO-6000*RJ**2
B(1,5)=7200*RO**2+2400*RJ**4+900
B(1,6)=9.8
2(1,7)=9.8
B(2,3)=9.8
B(2,4)=9.8
B(2,5)=9.8
B(2,6)=9.8
B(2,7)=9.8
B(3,4)=(-10500)*KJ
C(1)=(1,000)/2DO*3.145926536900*SIG1*SIG2*RUSQ**(.5Q))
DO 4 J=1,4
4 J=2+J
C(1)=C(1)/2JSQ**J
SU=300
SF=300
JG 16 I=1,4
16 CALL TAYLOR(WQ1(I),C,JPI1(I),UP2(I),LO1(I),L02(I),1)
SU 17 I=1,4
18 FORMAT(///**** APPROX. FOR REGION NJ. =*+14,* ****,///)
<EGP=Q(1)
NT=1000
CGS=000
SCS#003
FSU=000
FGS=000
PR0=(UP1(I)-LO1(I))*(UP2(I)-LO2(I))
DO 5 I=1,1000
5 DO 7 J=1,2
7 CALL RAYUJ(RAND,1Y,YFL)
RND-1Y
X(1)=YFL
X(1)=LO1(I)+X(1)*(UP1(I)-LO1(I))
X(2)=LO2(I)+X(2)*(UP2(I)-LO2(I))
HOL=1-000/(2000+RUSQ))*X(1)-M1)**2/SIG1**2-2DO*K0*(X(1)-M1)**2
C(X(1)-M2)/(SIG1*SIG2)*(X(2)-M2)**2/SIG2**2
F=C(1)*DEXP(HOL)
F=F*PR0
FSU=FSQ+(F**2)
FSU=FSUM+F
CALL TAYLOR(G,C,X(1),X(2),LO1(I),LO2(I),2)
G=G*PR0
GSU=GSUM+G
FGSQ=FGS+((F-G)**2)
FGS=CSUM+F*F-(F-G)
6 CONTINUE
FSUM=FSUM-3SUM
FSUM=FSUM/NT
FVAR=(FSQ-I(SUM**2/NT))/(NT-1,000)
VVAR=(FSQ-FSUM**2/NT)/(NT-1,000)


```
53
SF=DSQT(FVAR/NT)
CM=SUM(NT)
SE=SE+SF
SU=SU+SF
GO TO 17
9 FG=FGM+FGP
SU=SU+SF
GO TO 17
114 FORMAT(' ***THE REGULAR PART IS NOT REMOVED***')
 WRITE(6,114) FG,CM
 WRITE(5,111) SE
111 FORMAT(' ***THE REGULAR PART HAS BEEN REMOVED***')
 WRITE(6,111) FG,CM
14 FORMAT(' THE VALUE IS ',F10.5,' ST. ERR = ',F11.5)
15 CONTINUE
 SE=SE+SF
19 FORMAT(' THE TOTAL PROBABILITY IS ',F12.5)
 GO TO 15
100 STOP
END
```

```
## APPENDIX B

**Approximation Program - Card Input**

<table>
<thead>
<tr>
<th>Cols.</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Card 1</td>
<td>1-4</td>
<td>number of data pairs</td>
</tr>
<tr>
<td></td>
<td>5-80</td>
<td>19 pairs of data with each element of each pair right justified in a two column space; no decimal points</td>
</tr>
<tr>
<td>Card 2+</td>
<td>1-76</td>
<td>19 pairs of data with each element of each pair right justified in a two column space; no decimal points. That is the data is read with an 19F2.1 format. There will be as many cards of this type as necessary to punch all data.</td>
</tr>
<tr>
<td>Last Card</td>
<td>1-10</td>
<td>lower integration limit for the first site</td>
</tr>
<tr>
<td></td>
<td>11-20</td>
<td>upper integration limit for the first site</td>
</tr>
<tr>
<td></td>
<td>21-30</td>
<td>lower integration limit for the second site</td>
</tr>
<tr>
<td></td>
<td>31-40</td>
<td>upper integration limit for the second site</td>
</tr>
</tbody>
</table>
CARD COLS. VARIABLE

1 1-40  TITLE
2 5-80  ALTERNATING DATA:  FIRST SITE CLOUD COVER THEN SECOND SITE CLOUD COVER WITH AN INTEGER BETWEEN 1 AND 10 IN CONSECUTIVE TWO COLUMN SPACES.
3+ 1-70  CONTINUE DATA INPUT AS ABOVE
4 1-10  LOWER INTEGRATION LIMIT FOR SITE ONE
4 11-20  LOWER INTEGRATION LIMIT FOR SITE TWO
4 21-30  UPPER INTEGRATION LIMIT FOR SITE ONE
4 31-40  UPPER INTEGRATION LIMIT FOR SITE TWO

ALL DATA ON CARD 4 MUST BE LESS THAN 1 SINCE WE ARE DEALING WITH THE BETA DISTRIBUTION (AND GREATER THAN ZERO)

IMPLICIT REAL*8 (A-H,O-Z)
REAL*8 X(155),Y(155),CX(155),CY(155),MX,AA(10)
INTEGER FI5,51/25*0/,FI0(11,11)/121*0/
REAL(5,76) (AA(I),I=1,10)
WRITE(5,70) (AA(I),I=1,10)
70 FORMAT(1CA8)
59 READ(*,37) END=100 I*(X(I),Y(I),I=1,4)
37 FORMAT(14,(38F2.1))
DO 50 I=1,5
DO 50 J=1,5
50 F(I,J)=0
DO 71 I=1,11
DO 71 J=1,11
71 F0(I,J)=0
DO 72 I=1,N
72 F0(I,J)=F(J)(N1,N2)+1
CALL CAT(X,CX,Y,CY)
DO 51 I=1,N
51 F(V1,N2)=F(V1,N2)+100
WRITE(6,200)
200 FORMAT(//'**** RESULTS USING ORIGINAL DATA '****',//)
WRITE(6,78) 4
78 FORMAT(10O,,22X,15,2X,'VALUES',//)
DO 73 J=1,11
73 WRITE(6,52) (FI0(I,J),J=1,11)
52 FORMAT(10X,1115)
DO 55 I=1,N
55 CX(I)=CX(I)-500/500
CALL STAT(X,Y,MX,MY,SVX,SVY,R01,N)
CALL STAT(CX,CY,MCX,MCY,SVCX,SVCY,RO2,N)
SIG1=DSQRT(SVX)
SIG2=DSQRT(SVY)
SIGC1=DSQRT(SVCX)
SIGC2=DSQRT(SVCY)
WRITE (6,48) MX,MY,SIG1,SIG2,RO1
48 FORMAT(//, ' MEANS=', 10X, 2F10.4, //, ' ST. DEVS=', 8X, 2F10.4, //
     10X, 'XI', /*XI*/)
     B1=((ID0-MX)/SVCX)*4X1(ID0-MX)-SVX)
     A1=(MX-1)/1(ID0-MX)
     A2=(MY-32)/1(IDO-MY)
     WRITE(*,201)
     201 FORMAT(ESTIMATED BETA PARAMETERS, ///)
     WRITE(*,202)
     202 FORMAT(///, ' STORED CATEGORICAL DATA ///)
     54 FORMAT(*, ' FREQUENCY TABLE:', 15X, 'VALUES', ///)
     WRITE(*,54) N
     56 53=1
     59 WRITE(b,52) (f(1,J), J=1,5)
     WRITE(b,51) (MX,MY, SIGC1, SIGC2, RO2
     WRITE(b,50) (MCX,MCY, SVCX, SVCY)
     WRITE(b,50) (MCX,MCY, SVCX, SVCY)
     WRITE(6,50) HA1,HA1,HA2,HA2,HR2
     WRITE(6,50) HA1,HA1,HA2,HA2,HR2
     203 FORMAT(///, ' STORED RESULTS ///)
     75 FORMAT(' IF A PARAMETER OR PARAMETERS IS LESS THAN OR EQUAL TO
     77 .5, THE Z-VALUE IS UNDEFINED. FOR FURTHER COMPUTATION THE PARAMETER
     78 IS RESET TO .51, ///)
     79 WRITE(b,56) Y1, Y3, Y2, Y4
     85 FORMAT(8F13.0)
     CALL CALZ(H1,HA1,Y1,N,ZC1)
     CALL CALZ(H32,HA2,Y2,N,ZC2)
     CALL CALZ(H31,HA1,Y3,N,ZL3)
     CALL CALZ(H32,HA2,Y4,N,ZL4)
     CALL CALZ(H1,A1,Y1,N,Z21)
     CALL CALZ(H32,HA2,Y2,N,Z22)
     CALL CALZ(H31,HA1,Y3,N,Z21)
     CALL CALZ(H32,HA2,Y4,N,Z22)
     WRITE(6,74)
     74 FORMAT(0, 'FIRST SITE', 20X, 'SECOND SITE', 20X, 'UPPER', 10X, 'LOWER
     76 FORMAT(0, 'INTegral LIMITS', 1X, 2F14.8, 1X, FL4.8, 1X, FL4.8, 1X, 'CATEGORICAL DATA', 1X, 2F14.8, 1X, 'ORIGINAL DATA', 1X, C2(F14.8, 1X, 1X, 1X))
     WRITE(6,74) Y1, Y3, Y4, Y2, ZL3, ZC1, ZL4, ZC2, ZL1, ZL2, Z2
     85 STOP
     100 END
SUBROUTINE STAT(X,Y,M1,M2,SVARX,SVARY,RO,N)
IMPLICIT REAL*8 (A-H,M-O,Z)
REAL*8 X(155),Y(155)
SX=0.0
SY=0.0
SXY=0.0
SYX=0.0
DO 47 I=1,N
SX=SX+X(I)
SY=SY+Y(I)
SXY=SXY+(X(I))*Y(I)
SYX=SYM+(Y(I))*X(I)
M1=M1+1
M2=M2+1
SVARX=(SXS-Y*M1**2)/(N-1.0)
SVARY=(SYS-Y*M2**2)/(N-1.0)
30+30/30**RRTI(SVARX*SVARY)
LET JLN
END

SUBROUTINE CALZ(B1,A1,Y,N,Z)
IMPLICIT REAL*8 (A-H,M-O,Z)
IF (A1.LE.5.0) A1=5.10-2
IF (B1.LE.50) B1=5.10-2
SX1=B1-.500
SX2=A1-.500
SNX=A1+B1-100
P=12-Y
C2=(SXY+3.333333333D0)*Y-(A1+.333333333D0)+20-2*(Y/B1-P/A1+Y-.5D0)/
G(A1+1)
DA=ABS(SX1-SXN*P)
DLJ=(SX1/SXN*P))
JLT=(DLJ((SX2/SXN*Y))
Z=2/DATI()**RRTI12CO=SNX/1600*SNX+100)*(SXI*DL+SXZ*DLT))
RETURN
END

SUBROUTINE CAT(X,H,N)
REAL*8 X(155),H(155)
DO 38 I=1,N
IF (X(I)-.1D0) 45,46,46
45 H(I)=100
GO TO 38
38 CONTINUE
RETURN
END
A PROGRAM TO COMPUTE CONDITIONAL BIVARIATE NORMAL PARAMETERS

Summary

This report derives the conditional bivariate normal parameters from an original quadrivariate distribution. The paper presents the theory and appended is a computer program developed to give numerical results. An example is presented in the paper.

I. INTRODUCTION

This report presents a sketch of the theory and a computer program designed to calculate the bivariate normal conditional distribution derived from the quadrivariate normal distribution. The required computer inputs are described and an example is presented. The computer program is appended.

Theory

The general multivariate normal distribution has the density

\[ f(x_1, x_2, \ldots, x_k) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x-\mu)' \Sigma^{-1} (x-\mu) \right\} \]  

where \( \mu = (\mu_1, \mu_2, \ldots, \mu_k) \), the vector of mean values and

\[
\Sigma = \begin{bmatrix}
\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\
\sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk}
\end{bmatrix}
\]
A property of the multivariate normal distribution is that marginal and conditional distributions are also normally distributed. The general expression for these distributions are found often in the literature [see Morrison (1967)]. We shall confine remarks here to the specific case.

Assume we wish to derive \( f(x_1, x_2, x_3, x_4) \). If we define

\[
\mathbf{u} = \begin{bmatrix}
\mu_1 \\
\mu_2 \\
\mu_3 \\
\mu_4
\end{bmatrix}
\quad \text{and} \quad
\Sigma = \begin{bmatrix}
\sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\
\sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\
\sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\
\sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44}
\end{bmatrix},
\]

then letting

\[
\mathbf{x} = \begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{bmatrix}
\quad \text{we have}
\]

\[
f(x_1|x_2) = \frac{1}{(2\pi)^{\frac{1}{2}}} \exp \left( -\frac{1}{2} \left( x_1 - \mu^* \right)' \Sigma^{-1} \left( x_1 - \mu^* \right) \right). \tag{2}
\]

Where

\[
\Sigma^* = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}
\]

and

\[
\mu^* = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2). \tag{4}
\]

Computation of the parameters for this conditional distribution really reduces to computation of the quantities \( \Sigma^* \) and \( \mu^* \). Carefully note that the value of \( \mu^* \) includes values of \( x_2 = [x_3, x_4]' \) that must be specified before numerical values for \( \mu^* \) can be calculated.

Even for this rather easy case the actual expressions for \( \Sigma^* \) and \( \mu^* \) and therefore for the quadratic form in (2) are very complicated algebraically. They are, however, very amenable to numerical computation via computer. The least complicated for the expressions is that for \( \mu^* \) and the actual form is given below (letting \( \sigma_{34} = \sigma_{43} \) for convenience).
\[
\mu^* = \begin{bmatrix}
\mu_1^* & \left(\sigma_{13} \sigma_{33} - \sigma_{14} \sigma_{34}\right) (x_3 - \mu_3) + (\sigma_{14} \sigma_{33} - \sigma_{13} \sigma_{34}) (x_4 - \mu_4) \\
\mu_2^* & \left(\sigma_{23} \sigma_{33} - \sigma_{24} \sigma_{34}\right) (x_3 - \mu_3) + (\sigma_{24} \sigma_{33} - \sigma_{23} \sigma_{34}) (x_4 - \mu_4)
\end{bmatrix}
\]

The matrix triple product \( \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \) makes \( \Sigma^* \) a complicated expression and this, of course, causes \( (\Sigma^*)^{-1} \) and, therefore, the quadratic form in (2) to be almost incomprehensible in an expanded form.

**Computer Program and Required Inputs**

The computer program is written to accept quadrivariate data and return the conditional bivariate parameter. The conditional variance-covariance matrix and the associated standard deviations and correlations are initially calculated and printed. The program is designed to take as many pairs of "conditioning values" of \( x_3 \) and \( x_4 \) as desired and print out both the values of \( x_3 \) and \( x_4 \) plus the associated values of \( \mu^* \).

**Example:** The following data was input to the program

\[
\mu = [21.58, -.04, 43.35, 1.25]^T
\]

\[
\sqrt{\sigma_{11}} = 11.03, \quad \rho_{12} = .0503, \quad \rho_{13} = .7382, \quad \rho_{14} = -.0199
\]

\[
\sqrt{\sigma_{22}} = 11.52, \quad \rho_{23} = 1614, \quad \rho_{24} = .8134
\]

\[
\sqrt{\sigma_{33}} = 15.47, \quad \rho_{34} = .1524
\]

\[
\sqrt{\sigma_{44}} = 14.59
\]

\([x_3, x_4] = [43.35, 0]\)

Attached as Appendix I is the output giving the calculated parameters for the bivariate conditional. Note carefully that the standard deviations and correlations are printed in matrix form for convenience.
not to be confused with the variance-covariance matrix printed above it. Below the standard deviation and correlation matrix the values conditioned on and the resulting conditional means are printed. The original inputs and matrices will be printed only once but the values conditioned on, followed by the conditional means calculated using those values, will be repeated for each set of conditioning values read in.

Input to the program consists of the following cards:

Card 1 The 4 means for the quadrivariate normal in 4F10.4 Format.

Card 2 Standard deviation for variable 1 followed by correlations for variables 1&2, 1&3, and 1&4 in 4F10.4 Format.

Card 3 Standard deviation for variable 2 followed by correlations for variables 2&3 and 2&4 in 3F10.4 Format.

Card 4 Standard deviation for variable 3 followed by correlation between variables 3&4 in 2F10.4 Format.

Card 5 Standard deviation for variable 4 in F10.4 format.

Card 6 Number of sets of $x_3, x_4$ values to be conditioned on in 12 Format.

Card 7 1st set of $x_3, x_4$ values to be conditioned on

Card 8 2nd set of $x_3, x_4$ values to be conditioned on

Card 9 3rd $x_3, x_4$ values to be conditioned on

Card 10 etc.

The source deck listing is given in Appendix II.

References

APPENDIX I

**MEANS VECTOR**

\[
\begin{pmatrix}
21.5800 \\
-0.4070 \\
33.3500 \\
1.2500
\end{pmatrix}
\]

**VARIANCE-COVARIANCE MATRIX**

\[
\begin{bmatrix}
121.6609 & 6.3914 & 125.9621 & -3.2025 \\
6.3914 & 132.7104 & 28.7630 & 136.7336 \\
125.9621 & 28.7630 & 239.3202 & 34.3973 \\
-3.2025 & 136.7336 & 34.3973 & 212.8631
\end{bmatrix}
\]

**COND. VAR. COV. MATRIX**

\[
\begin{bmatrix}
53.17973 & 4.03824 \\
4.03824 & 44.71625
\end{bmatrix}
\]

**SUBCORR. MATRIX**

\[
\begin{bmatrix}
7.29244 & 0.00922 \\
0.00922 & 6.66762
\end{bmatrix}
\]

**VALUES CONDITIONED ON**

\[
\begin{pmatrix}
43.3500 \\
-0.3070
\end{pmatrix}
\]

**CONDITIONAL MEANS**

\[
\begin{pmatrix}
21.7081 \\
-6.8079
\end{pmatrix}
\]
APPENDIX II

```plaintext
DIMENSION SIGMA(2,2),X0(2),VX(2),US(2),COR(2,2),V(2,2),
      V2(2,2),V3(2,2),V4(2,2),U1(2),U2(2),X(2),W24(2,2),VVV(2,2),SD(4),
      RHO(4,4),U(4),S(4,4)
READ(5,1) (U(I),I=1,4)
1 FORMAT(4F10.4)
   DO 2 I=1,4
      L=1
      IF (I.LE.3) L=I+1
2 READ(5,3) SD(I),(RHO(I,J),J=L,4)
3 FORMAT(4F10.4)
   DO 4 I=1,4
      L=I+1
      S(I,I)=SD(I)**2
      IF (L.LE.5) GO TO 4
   DO 4 J=L,4
      S(I,J)=RHO(I,J)*SD(I)*SD(J)
     ^(J,I)=S(I,J)
4 CONTINUE
   WRITE(G,5) (U(I),I=1,4)
5 FORMAT('U(1) = ',4(F10.4,2X))
   WRITE(G,6)
6 FORMAT('VARIANCE - COVARIANCE MATRIX')
   DO 7 J=1,4
7 WRITE(6,8) (S(I,J),J=1,4)
8 FORMAT(5X,4(F10.4,4X))
   DO 9 I=1,2
      DO 9 J=1,2
         V1(I,J)=S(I,J)
         V2(I,J)=S(I,J+2)
         V3(I,J)=S(I+2,J)
9 V4(I,J)=S(I+2,J+2)
   D=V4(1,1)*V4(2,2)-V4(1,2)*V4(2,1)
   C=V4(1,1)
   V4(1,1)=V4(1,1)/D
   V4(2,1)=V4(2,1)/D
   V4(1,2)=C/D
   DO 10 I=1,2
      U1(I)=U(I)
10 U2(I)=U(I)
```

APPENDIX II (CONTINUED)

DO 11 I=1,2
DO 11 J=1,2
VV24(I,J)=0
DO 11 K=1,2
11 VV24(I,J)=VV24(I,J)+V2(I,K)*V4(K,J)
DO 12 I=1,2
DO 12 J=1,2
VVV(I,J)=0
DO 12 K=1,2
12 VVV(I,J)=VVV(I,J)+VV24(I,K)*V3(K,J)
DO 13 I=1,2
DO 13 J=1,2
SRTIA(I,J)=V1(I,J)-VVV(I,J)
COR(1,1)=SORT(SRTIA(1,1))
COR(2,2)=SORT(SRTIA(2,2))
COR(2,1)=SRTIA(1,1)/(COR(1,1)*COR(2,2))
COR(1,2)=COR(2,1)
WRITE(6,14) SRTIA
WRITE(6,15) COR
14 FORMAT('OCOND. VAR. COV. MATRIX'/2(5X,F10.5))
15 FORMAT('OSTD.CORR. MATRIX'/2(5X,F10.5))
READ(5,16) ::
16 FORMAT(12)
DO 23 N=1,2
READ(5,17) (X(I),I=1,2)
17 FORMAT(2F10.5)
DO 18 I=1,2
XU(I)=X(I)-U1(I)
DO 19 I=1,2
VX(I)=0
DO 19 J=1,2
19 VX(I)=VX(I)+VV24(I,J)*XU(J)
DO 20 I=1,2
20 US(I)=U1(I)+VX(I)
WRITE(6,21) (X(I),I=1,2)
WRITE(6,22) (US(I),I=1,2)
21 FORMAT('VALUES CONDITIONED ON',2(5X,F10.4))
22 FORMAT('CONDITIONAL MEANS',2(5X,F10.4))
STOP
END
TRANSFORMATION OF NON-NORMAL MULTIVARIATE DATA
TO NEAR-NORMAL

Summary

A procedure for transforming non-normal multivariate data to near-normal data is presented. The procedure is based upon a multivariate generalization of a technique proposed by Box and Cox (1964). Several examples of the procedure are included along with a documentation of the computer software.

I. INTRODUCTION

Investigators are often confronted with the problem of analysing multivariate data. Upon investigating the existing procedures for analysing this type of data, one soon realizes that a majority of the existing techniques are restricted to the normal distribution. However, real data often violates this normality assumption. Thus the investigator is confronted with two possible approaches: 1) determine a non-normal multivariate distribution which provides a satisfactory model, 2) determine a technique for transforming the non-normal data to near-normal data. If the investigator is mainly interested in modeling the multivariate data, then the first approach is probably most appropriate, however, if the main interests are in making statistical inferences or probabilistic forecasts then the second approach could prove to be adequate. In this paper, we
have presented a procedure which addresses this second approach. The procedure is a multivariate generalization of a procedure proposed by Box and Cox (1964). They proposed the following univariate transformation

\[
y(\lambda) = \begin{cases} 
\frac{y^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0, \\
\log(y) & \text{for } \lambda = 0.
\end{cases}
\]

Andrews et al. (1971) extended this transformation to the bivariate case. In their paper, they were able to find approximate maximum likelihood estimates for \(\lambda\) by examining the contours of the likelihood function. In this paper, the method of Box and Cox is extended to the multivariate case, where the maximum likelihood estimate for \(\lambda\) is determined using a numerical analysis approach. The procedure is presented in a multivariate analysis of variance setting, however, several examples are presented which demonstrate the versatility of the technique.

II. Procedure

Let \(Y_{i1}, \ldots, Y_{in_i}\) denote a random sample of \(n_i\) \(p\)-dimensional observations from a population with finite mean \(\mu_i\) and finite covariance \(\Sigma_i\), for \(i = 1, 2, \ldots, m\). The problem can be stated as; find \(\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_p)^T\) such that \(Y_{ij}^{(\lambda)}\) is distributed normally with
mean $\mu_1$, and common covariance $\Sigma$, where

$$y_{ij}^{(\lambda)} = \begin{pmatrix} (\lambda_1) \\ \vdots \\ (\lambda_p) \end{pmatrix}^T (y_{ij1}, ..., y_{ijp})$$

$$y_{ij} = \begin{cases} \frac{(\lambda_k)}{(y_{ijk} - 1)/\lambda_k} & \text{for } \lambda_k \neq 0 \\ \log(y_{ijk}) & \text{for } \lambda_k = 0 \end{cases}$$

for $i=1,2,\ldots,m$, $j=1,2,\ldots,n_i$, and $k=1,2,\ldots,p$. For $Q$, $Y_{ij}^{(\lambda)}$ can be written as

$$Y_{ij}^{(\lambda)} = D^{-1}(Y_{ij} - J)$$

where

$$D = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$$

$J$ is a px1 vector of 1's

$$Y_{ij}^{\lambda} = (y_{ij1}, y_{ij2}, ..., y_{ijp})^T.$$  

Since $Y_{ij}^{(\lambda)} \sim N(\mu_1, \Sigma)$, its density function can be written as

$$f(z) = \exp\left(-\frac{1}{2}(z-\mu)^T \Sigma^{-1}(z-\mu)\right) (2\pi)^{-p/2} |\Sigma|^{-1/2}$$

where $z = Y_{ij}^{(\lambda)}$. From this, one can determine the density function for the
untransformed data $w = Y_{ij}$ as $g(w) = K_{ij} f(z)$ where,

$$K_{ij} = \prod_{k=1}^{p} \frac{\partial z_k}{\partial w} = \prod_{k=1}^{p} (y_{ijk})^{\lambda_k-1}.$$  \hspace{1cm} (6)

Hence the joint likelihood function becomes

$$L(\lambda) = \left( \frac{m}{\pi} \right)^n \left( \frac{\pi}{\Sigma} \right)^{(2n-p)/2} \left( \frac{n}{\Sigma} \right)^{-n/2}$$

$$\cdot \exp \left\{ -\frac{1}{2} \Sigma \left( \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^T \Sigma^{-1} (Y_{ij} - \mu_i) \right) \right\} \hspace{1cm} (7)$$

where $n = \sum_{i=1}^{m} n_i$. The likelihood function can be written as

$$L(\lambda) = K(2\pi)^{-np/2} |\Sigma|^{-n/2} \exp \left\{ -\frac{1}{2} \Sigma \left( \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^T \Sigma^{-1} (Y_{ij} - \mu_i) \right) \right\} \hspace{1cm} (8)$$

where $K = \prod_{i=1}^{m} \prod_{j=1}^{n_i} K_{ij}$ and $\mu$ and $\Sigma$ are replaced by their maximum likelihood estimates

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_{1i} \hspace{1cm} (9)$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{1i}) (Y_{ij} - \bar{Y}_{1i})^T.$$
Equation (8) follows from equation (7) since

\[
\sum_{i=1}^m \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_i)^T \Sigma^{-1} (Y_{ij} - \hat{\mu}_i)
\]

\[
= \sum_{i=1}^m \sum_{j=1}^{n_i} (\text{tr} \Sigma^{-1} (Y_{ij} - \hat{\mu}_i)^T (Y_{ij} - \hat{\mu}_i) T)
\]

\[
= \text{tr} \Sigma^{-1} \sum_{i=1}^m \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_i)^T (Y_{ij} - \hat{\mu}_i) T
\]

\[
= n \text{tr} \Sigma^{-1} \Sigma = np.
\]

Equation (8) can be further simplified as

\[
L(\lambda) = C \cdot h(\lambda)
\]

(10)

where \( C = (2\pi)^{-np/2} \exp \left\{-np/2\right\} \)

\[
h(\lambda) = |K^{2/n} \Sigma^{-1/2}|^{-n/2}
\]

(11)

Note that maximizing the likelihood function \( L(\lambda) \) is equivalent to minimizing the function \( h(\lambda)^{-1} \). This function can be further simplified by considering

\[
K^{2/n} = \left( \prod_{i=1}^m \prod_{j=1}^{n_i} k_{ij} \right)^{2/n}
\]

\[
= \prod_{k=1}^p \left( \prod_{i=1}^m \prod_{j=1}^{n_i} (y_{ijk}^{\lambda k - 1} 1/n)^2 \right)
\]

\[
= \prod_{k=1}^p (\hat{\gamma}_k^{\lambda k - 1})^2
\]

(12)
where \( \hat{y}_k = \left( \prod_{i=1}^{m} \prod_{j=1}^{n_i} y_{ijk} \right)^{1/n} \) is the geometric mean for the \( k \)th variate, \( k = 1, 2, \ldots, p \). From equation (4) \( \Sigma \) can be written as

\[
\hat{\Sigma} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) (y_{ij} - \bar{y}_i)^T
\]

\[
= \sum_{i=1}^{m} \sum_{j=1}^{n_i} D^{-1}(y_{ij}^\lambda - \bar{y}_i^\lambda) (y_{ij}^\lambda - \bar{y}_i^\lambda)^T D^{-1} \tag{13}
\]

Hence \( |\Sigma| \), becomes

\[
|\hat{\Sigma}| = |D^{-1}| \left| \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij}^\lambda - \bar{y}_i^\lambda) (y_{ij}^\lambda - \bar{y}_i^\lambda)^T \right| D^{-1}
\]

\[
= |D^{-2}| |G| \tag{14}
\]

where \( G = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij}^\lambda - \bar{y}_i^\lambda) (y_{ij}^\lambda - \bar{y}_i^\lambda)^T \). Thus the minimization of \( h(\lambda)^{-1} \) is equivalent to minimizing

\[
\phi(\lambda) = \frac{|G|}{|k^{2/N} \cdot \sigma^2|}
\]

\[
= \frac{|G|}{\prod_{k=1}^{p} \lambda_k \hat{y}_k^{\lambda - 1/2}} \left( \prod_{k=1}^{p} \lambda_k \hat{y}_k \right)^{\lambda - 1/2} \tag{15}
\]

Note that equation (15) reduces to

\[
\frac{\sum_{i=1}^{m} \lambda \hat{y}_i^{\lambda - 1/2}}{(\lambda \hat{y}_i^{\lambda - 1/2})^2} \tag{16}
\]
which was proposed by Box and Cox (1964) for the univariate case. The function \( \varphi(\lambda) \) in equation (15) can now be minimized using a standard numerical technique. In this paper the Flecher-Powell algorithm of deflected steepest descent is used (see Appendix A).

III. Application

The first example illustrates a violation of the equality of covariance matrix assumption in a multivariate analysis of variance problem. The data set is R.A. Fisher's classical iris data (Fisher, 1936) where the response measurements are sepal length, width and petal length, width for three iris species: virginica, versicolor, and setosa. Although this data was originally presented as an application of linear discriminate analysis, Morrison (1967) uses this as an example in multivariate analysis of variance, for which he states, "we shall of course assume... a common covariance matrix". However, in applying Bartlett's likelihood ratio test for equality of covariance, we obtain a test statistic of 141 for 20 degrees of freedom. Hence the hypothesis of equality of covariance can easily be rejected with a high level of significance. In figure 1, the confidence ellipse for the two untransformed variables: sepal length and sepal width, clearly illustrate the difference in covariance matrices. The data is then transformed, and the corresponding confidence ellipses are presented in figure 2. Although the confidence ellipses for the transformed data are more nearly identical, Bartlett's test statistic has been reduced to 63, however, this value is still significant at the .01 level.
Figure 1. Untransformed 95% Confidence Ellipses

Sepal Width

Sepal Length

Setosa

Versicolor

Virginica
Figure 2. Transformed 95% Confidence Ellipses
In the second example, we are interested in obtaining probabilistic forecasts. The data was originally presented in a paper by Haggard et. al. (1973), in which the author was able to model the maximum rainfall from tropical cyclone systems across the Appalachians using the Gamma distribution. Since one of their primary objectives was to obtain estimates for the probability of rainfall exceedence in the Appalachian regions, I felt that comparative results could be obtained by transforming the data then using the well tabulated normal distribution. The results are given in Table 1.

IV. Conclusions

A method transforming non-normal multivariate data to nearly-normal data is presented. The method extends the univariate transformation of Box and Cox (1964) to the multivariate case. A numerical method for approximating the optimal transformation is also included (see Appendix A). The procedure was then applied in two applications. The first was in the area of multivariate analysis of variance where the primary objective was to achieve equality of covariance matrices. It was shown that the transformed data was less heterogeneous than the untransformed data. However, the population covariances were still unequal. The second application illustrated that this type of procedure can be used when the primary objective is the estimation of tail probabilities. This method allows the use of the normal distribution on the transformed data, rather than determining the appropriate non-normal distribution for the untransformed data.
TABLE 1

Expected Probabilities of Exceeding Arbitrary Precipitation Amounts Over the Appalachian Region

<table>
<thead>
<tr>
<th>Precipitation in inches</th>
<th>Data Set*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>I **  I</td>
</tr>
<tr>
<td>1</td>
<td>.978 .966</td>
</tr>
<tr>
<td>2</td>
<td>.913 .903</td>
</tr>
<tr>
<td>3</td>
<td>.821 .819</td>
</tr>
<tr>
<td>4</td>
<td>.717 .723</td>
</tr>
<tr>
<td>5</td>
<td>.613 .624</td>
</tr>
<tr>
<td>6</td>
<td>.515 .528</td>
</tr>
<tr>
<td>7</td>
<td>.427 .439</td>
</tr>
<tr>
<td>8</td>
<td>.349 .359</td>
</tr>
<tr>
<td>10</td>
<td>.227 .232</td>
</tr>
<tr>
<td>15</td>
<td>.070 .066</td>
</tr>
<tr>
<td>20</td>
<td>.019 .016</td>
</tr>
<tr>
<td>25</td>
<td>.005 .003</td>
</tr>
<tr>
<td>30</td>
<td>.001 .001</td>
</tr>
</tbody>
</table>

* A - maximum 24-hour precipitation all storms. B - maximum 24-hour precipitation from no more than one storm per year. C - maximum precipitation totals from all storms. D - maximum precipitation totals from no more than one storm per year.

** I - gamma parameters from Haggard et al. (1973); II transformed normal probabilities.
V. REFERENCES


Appendix A

Flecher-Powell method of deflected steepest descent, requires the gradient vector

\[ \vartheta(\lambda) = \begin{bmatrix} \frac{\partial \vartheta}{\partial \lambda_1} \\ \frac{\partial \vartheta}{\partial \lambda_2} \\ \vdots \\ \frac{\partial \vartheta}{\partial \lambda_p} \end{bmatrix} \]  

(A.1)

where

\[ \vartheta(\lambda) = \frac{|G|}{\prod_{k=1}^{p} \lambda_k^\gamma_k^{\lambda_k-1} \cdot \lambda_k-1} \]  

(A.2)

\[
\frac{\partial \vartheta(\lambda)}{\partial \lambda_h} = \frac{\partial}{\partial \lambda_h} \left( \frac{|G|}{\prod_{k=1}^{p} \lambda_k^\gamma_k^{\lambda_k-1}} \right) - 2 \frac{\partial}{\partial \lambda_h} \left( \frac{\prod_{k=1}^{p} \lambda_k^\gamma_k^{\lambda_k-1}}{\prod_{k=1}^{p} \lambda_k^\gamma_k^{\lambda_k-1}} \right) + \frac{\partial}{\partial \lambda_h} \left( \frac{\prod_{k=1}^{p} \lambda_k^\gamma_k^{\lambda_k-1}}{\prod_{k=1}^{p} \lambda_k^\gamma_k^{\lambda_k-1}} \right) 
\]  

(A.3)

\[ \frac{\partial}{\lambda_h} \left( \prod_{k=1}^{p} \lambda_k^{\gamma_k^{\lambda_k-1}} \right) = -2 \left( \prod_{k=1}^{p} \lambda_k^{\gamma_k^{\lambda_k-1}} \right) \]  

(A.4)

Since \( |G| = \prod_{j=1}^{p} g_{ij} a_{ij} \) where

\[ G = (g_{ij}) \]  

\[ a_{ij} \] is the cofactor of \( g_{ij} \)
Also, since $g_{ij}$ only depends upon $\lambda_i, \lambda_j$ using the chain rule we have

$$\frac{\partial |G|}{\partial \lambda_h} = \sum_{i=1}^{p} \sum_{j=1}^{\nu} \frac{\partial |G|}{\partial g_{ij}} \frac{\partial g_{ij}}{\partial \lambda_h}$$  \hspace{1cm} (A.6)

where

$$\frac{\partial |G|}{\partial g_{ij}} = a_{ij}$$  \hspace{1cm} (A.7)

and

$$\frac{\partial g_{uv}}{\partial \lambda_h} = \left\{ \begin{array}{ll}
0 & \text{if } u, v \neq h \\
b_2 & \text{if } u \text{ or } v = h \\
b_3 & \text{if } u = v = h
\end{array} \right.$$  \hspace{1cm} (A.8)

and

$$b_2 = \frac{\partial}{\partial \lambda_h} \left( \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{\lambda_u}{y_{ij} - y_{1u}} \right) \left( \frac{\lambda_h}{y_{ij} - y_{1h}} \right)$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{\lambda_u}{y_{ij} - y_{1u}} \left( \frac{\lambda_h}{y_{ij} - y_{1h}} \right) \left( \frac{\lambda_h}{y_{ij} - y_{1h}} \right)$$

$$b_3 = 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{\lambda_h}{y_{ij} - y_{1h}} \left( \frac{\lambda_h}{y_{ij} - y_{1h}} \right) \left( \frac{\lambda_h}{y_{ij} - y_{1h}} \right)$$  \hspace{1cm} (A.9)

From this, equation (A.3) becomes

$$\frac{\partial \Theta(\lambda)}{\partial \lambda_h} = \frac{2}{\left( \frac{\lambda_h}{y_{1h}} \right)^2} \left[ \sum_{k=1}^{p} \frac{\alpha_{kh}}{\lambda_h} \frac{\partial g_{kh}}{\partial \lambda_h} + \frac{\alpha_{hh}}{2} \frac{\partial g_{hh}}{\partial \lambda_h} - \frac{|G|}{\lambda_h} \frac{\partial g_{ll}}{\partial \lambda_h} \right]$$  \hspace{1cm} (A.10)
Test of Fit for the Extreme Value Distribution
Based Upon the Generalized Minimum Chi-Square

Summary

A goodness of fit test for the extreme value distribution is developed. The procedure is based upon the generalized minimum chi-square distribution [Gurland and Dahiya (1970)]. Application of the test is given for some extreme value data [Gumbel (1964)].

I. Introduction

There are several difficulties with using the Pearson chi-square test of fit for continuous distributions [c. f. Dahiya and Gurland (1970)]. These difficulties are primarily concerned with the choice of cell width and the number of cells. However, to the applied statistician or non-statistician who must use test of fit procedures on a frequent basis, the primary difficulty of the procedures is in the users set up. That is, the user must have knowledge of the tabular values for the null hypothesis. Dahiya and Gurland (1970, 1972) presented a goodness of fit test for several continuous distributions which eliminate most of the user's set up. Their procedure was based upon the generalized minimum chi-square statistic. In this paper, I have developed a test of fit for the extreme value distribution based upon this generalized minimum chi-square technique.
II. Procedure

Suppose that one would like to test the null hypothesis given by

\[ H_0: X_1, X_2, \ldots, X_n \sim F_X(x; \theta) \]  

where \( X_1, X_2, \ldots, X_n \) denotes a random sample of \( n \) observations from a distribution function \( F_X(x; \theta) \). \( F_X \) is an asymptotic Fisher-Tippett type 1 distribution, that is,

\[ F_X(x; \theta) = \exp\left\{-\exp\left(-\frac{x-a}{\beta}\right)\right\} \]
\[ -\infty < a < \infty \]
\[ \beta > 0. \]

Let \( T \) denote a transformation from the population raw moments to \( \xi \), which can be written as a linear function of the parameters \( \theta \) where

\[ n' = (\eta_1', \eta_2', \ldots, \eta_s')^T \]
\[ \xi = (\xi_1, \xi_2, \ldots, \xi_s)^T \]

and \( \eta_j' \) is the \( j \)th raw population moment for \( F_X \) and \( \xi = W\theta \), \( W \) is a known \( s \times 2 \) matrix, and \( \theta = (a, \beta)^T \). That is,

\[ T: n \rightarrow \xi = W\theta. \]

Let \( m' = (m_1', m_2', \ldots, m_s')^T \) denote the sample raw moments corresponding to \( n \) and let \( h = (h_1, h_2, \ldots, h_s)^T \) denote the sample values corresponding to \( \xi \), that is,

\[ T: m' \rightarrow h. \]
By the central limit theorem, we know

\[ n(m' - n') \sim n(\emptyset, G) \]  \hspace{1cm} (6)

where the \( ij \)th element of the matrix \( G \) is

\[ g_{ij} = n_i' - n_j' \]  \hspace{1cm} (7)

for \( i, j = 1, 2, \ldots, s \). It also follows that

\[ n(h - \xi) \sim n(\emptyset, \Sigma) \]  \hspace{1cm} (8)

where \( \Sigma = TGT^T \). Now using the distributional properties for the quadratic forms, we know that

\[ Q^* = n(h - \xi)^T \Sigma^{-1} (h - \xi) \]  \hspace{1cm} (9)

has an asymptotic chi-square distribution with \( s \) degrees of freedom where \( \hat{\Sigma} \) is a consistent estimator for \( \Sigma \). Since \( \xi = W\theta \), an estimate for \( \theta \) can be found by minimizing \( Q^* \). In which case, the estimate becomes

\[ \theta = (W^T \hat{\Sigma}^{-1} W)^{-1} W^T \hat{\Sigma}^{-1} h. \]  \hspace{1cm} (10)

By letting \( \hat{\xi} = W\hat{\theta} \), \( Q^* \) becomes

\[ \hat{Q} = nh^T \hat{\Lambda} h \]  \hspace{1cm} (11)

where

\[ \hat{\Lambda} = \hat{\Sigma}^{-1} (I - \hat{R}) \]

\[ \hat{R} = W(W^T \hat{\Sigma}^{-1} W)^{-1} W^T \hat{\Sigma}^{-1}. \]  \hspace{1cm} (12)

Again by the distributional properties of the quadratic forms, \( \hat{Q} \) has a non-central chi-square distribution with degrees of freedom \( \text{tr} \hat{\Sigma} \hat{\Lambda} \) and non-centrality parameter \( \lambda = \xi^T \hat{\Lambda} \xi \) if and only if \( \hat{\Sigma} \hat{\Lambda} \) is an idempotent matrix. It is easy to verify that \( (\hat{\Sigma} \hat{\Lambda})^2 = \hat{\Sigma} \hat{\Lambda} \), and \( \lambda = 0 \), so \( \hat{Q} \) has a chi-square
distribution with $s-q$ degrees of freedom. Using this distribution, one can reject the null hypothesis (1) with type I error if $Q > \chi^2_{\alpha}(s-q)$, where

$$\Pr( X > \chi^2_{\alpha}(s-q)) = \alpha. \quad (13)$$

Dahiya and Gurland (1970) developed the non-null distribution for $Q$, using this distribution one can compute the power of the test for a specified non-null distribution. In order to test (1), the transformation $T$ and the matrix $W$ need to be specified. Since we know that the populations cumulants for the extreme value distribution are

$$\kappa_j = (-B)^{j-1} \psi(j-1) \quad \text{for } j = 2, 3, \ldots \quad (14)$$

where

$$\psi(n) = (-1)^{n+1} n! \delta(n+1) \quad (1)$$

$$\delta(n) = \sum_{i=1}^{n} i^{-n}. \quad (15)$$

By letting $\xi = (\kappa_3 \kappa_2^{-1}, \kappa_4 \kappa_3^{-1}, \ldots, \kappa_{s+2} \kappa_{s+1}^{-1})^T$ and $W = (\psi(2)/\psi(1), \ldots, \psi(s+1)/\psi(s))^T$, and $\theta = B$ it is possible to map $n^* \rightarrow \xi$ where $s = 4$ and $q = 1$. By letting $h_i = k_j+2/k_j+1$, for $j = 1, 2, \ldots, 4$, and $k_j$ is the $j^{th}$ sample cumulant. We are now able to compute $Q$, where

$$\Sigma = JGJT \bigg|_{\theta = B} \quad (16)$$

$$J = (j_{mn}); \quad j_{mn} = \frac{\partial \mathcal{L}}{\partial \kappa_n^m} \text{ for } m, n = 1, 2, \ldots, s$$

and $\hat{B}$ is the maximum likelihood estimate for $B$. 
The values in equation (15) can be found in Abrahomovich, hence J becomes

\[ J = \frac{1}{8} \begin{bmatrix} 1 & 0 & 0 & 0 \\ -0.885 & 0.6079 & 0 & 0 \\ 0 & -1.131 & 0.4174 & 0 \\ 0 & 0 & -0.5901 & 0.154 \end{bmatrix} \]  

(17)

From these values, we are able to compute \( \hat{Q} \) in (11) for the sample values \( X_1, X_2, \ldots, X_n \). Hypothesis (1) can be rejected if \( \hat{Q} > \chi^2(3) \) since \( s = 4, q = 1 \).

**Application**

In this section, an extreme value data set given in Gumbel and Goldstein (1964) is analysed using this test of fit procedure. The data set consists of the oldest ages at death for men and women in Sweden from the period 1905-1958. The data for male and female are fitted separately. Gumbel and Goldstein (1964) estimated the extreme value distribution parameters using a modified method of moments. Tables 1 & 2 contain a comparison of the two different procedures in term of estimated parameters and cumulative tail probabilities. It must be noted, that the null hypothesis of the extreme value distribution being the null distribution could not be rejected at a significance level of greater than 70%.

In the second example, extreme monthly temperatures and winds for three United States locations were analysed. The data set taken from the daily meteorological records, 1970-1971, for New Orleans, LA., Orlando, FL., and Daytona Beach, FL. The results are summarized in Tables 3 and 4.
Table 1: Comparison of Procedures using Swedish Men

<table>
<thead>
<tr>
<th>Method of Moments</th>
<th>Generalized minimum ( x^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\alpha} )</td>
<td>( \hat{\beta} )</td>
</tr>
<tr>
<td>102.49</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* the values \( X \) represent the 5, 10, 20, 30, 40, 50, 54th smallest sample value. \( F_X \) and \( G_X \) are the corresponding c.d.f.

Table 2: Comparison of Procedures using Swedish Women

<table>
<thead>
<tr>
<th>Method of Moments</th>
<th>Generalized minimum ( x^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\alpha} )</td>
<td>( \hat{\beta} )</td>
</tr>
<tr>
<td>103.83</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* same as in Table 1
### Table 3

Extreme Monthly Temperatures

<table>
<thead>
<tr>
<th>Site</th>
<th>Extreme Value Distribution</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}$</th>
<th>$Q^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Orleans</td>
<td></td>
<td>83.8</td>
<td>.98</td>
<td>.001</td>
</tr>
<tr>
<td>Orlando</td>
<td></td>
<td>84.8</td>
<td>.88</td>
<td>.003</td>
</tr>
<tr>
<td>Daytona Beach</td>
<td></td>
<td>81.7</td>
<td>.67</td>
<td>.002</td>
</tr>
</tbody>
</table>

* null distribution of extreme valued distribution can not be rejected.

### Table 4

Extreme Monthly Winds

<table>
<thead>
<tr>
<th>Site</th>
<th>Extreme Value Distribution</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}$</th>
<th>$Q^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Orleans</td>
<td></td>
<td>15.4</td>
<td>2.9</td>
<td>.9</td>
</tr>
<tr>
<td>Orlando</td>
<td></td>
<td>13.6</td>
<td>2.5</td>
<td>.6</td>
</tr>
<tr>
<td>Daytona Beach</td>
<td></td>
<td>13.0</td>
<td>2.2</td>
<td>.4</td>
</tr>
</tbody>
</table>

* same as in Table 3
IV. Conclusions

A procedure for testing the goodness of fit for the extreme value distribution, based upon a generalized minimum chi-square is presented. The procedure is applied to several data sets where the extreme value distribution is a potential fit, although it must be mentioned that the meteorological data set was included in a manner which lends itself to program utility rather than for meteorological interpretation.

V. References


Test of Fit for Continuous Distributions Based Upon the Generalized Minimum Chi-Square

Summary

A procedure for test of fit for several continuous probability distributions based upon the generalized minimum chi-square method is presented. The procedure was first presented in a series of papers by Dahiya and Gurland ((1970a), (1970b), (1972)). Examples of the procedure are included, along with the corresponding computer listing.

I Introduction

Dahiya and Gurland (1970a) discuss the difficulties with using the Pearson chi-square test of fit for continuous distributions. These difficulties are primarily concerned with the choice of cell numbers and widths. However, to the applied statistician who must use test of fit procedures on a frequent basis the main disadvantage is in the users setup. That is, the user must have knowledge of the parameters and the tabular values for the specified null distribution. These demands severely hamper the investigator who must determine an appropriate distribution from potentially many distribution functions. The purpose of this paper is to present a test of fit for continuous distributions which minimizes the users interface in the estimation of parameters for the specified null distribution or in specifying the tabular values of the null distribution. In fact, several different families of distributions can be tested for fit using a single
setup. The procedure is based upon the generalized minimum chi-square (GMCS) statistical method. Section 3 contains the GMCS procedure for the univariate normal and gamma distributions.

Procedure

Suppose that we want to test the null hypothesis

$H_0: x_1, x_2, \ldots, x_n \sim F(x; \theta) \in \mathcal{F}(x; \theta)$  \hspace{1cm} (1)

where $x_1, x_2, \ldots, x_n$ is a random sample of $n$-observations from an unknown distribution function $F(x; \theta)$; $\theta$ is a $q \times 1$ vector of parameters and $\mathcal{F}(x; \theta)$ is a specified family of distributions with admissible parameters $\theta$.

The (GMCS) procedure can be used for testing any family of distribution $\mathcal{F}(x; \theta)$, provided there exists a transformation $T$, where

$T: \mu \rightarrow \xi$  \hspace{1cm} (2)

where $\mu' = (\mu_1, \mu_2, \ldots, \mu_s)^T$, $\mu_j$ is the $j^{th}$ raw population moment $1, 2, \ldots, s$ and $\xi = (\xi_1, \xi_2, \ldots, \xi_s)^T$ can be expressed as $\xi = W \theta$  \hspace{1cm} (3)

for a known $s \times q$ matrix $W$ and $s > q$. Let $m' = (m_1, m_2, \ldots, m_s)^T$ denote a $s \times 1$ vector of raw sample moments and define $h = (h_1, h_2, \ldots, h_s)^T$ to be the image of the transformation $T$, that is $T: m' \rightarrow h$. Using the central limit theorem, we have

$n(m' - \mu') \rightarrow n(\theta, G)$  \hspace{1cm} (4)

where $G = (g_{ij}), g_{ij} = \mu_j - \mu_j \mu_i, i, j = 1, 2, \ldots, s.$
From this, it can be shown that
\[ n(h - \xi) \to N(0, \Sigma) \]  
(5)

where \( \Sigma = J G J^T \), \( J \) the Jacobian matrix for the transformation \( T \). Now using the properties of quadratic forms, we know that
\[ Q = n(h - \xi)^T \Sigma^{-1}(h - \xi) \]  
(6)

has a chi-square asymptotic null distribution with \( s \) degrees of freedom. Furthermore, this distribution does not change when we estimate \( \Sigma \) in (6) by \( \hat{\Sigma} \), where \( \hat{\Sigma} \) is a consistent estimator for \( \Sigma \). Since \( \xi = W \theta \), we can estimate \( \theta \) by finding \( \hat{\theta} \) which minimizes \( Q \). This estimate is given by
\[ \hat{\theta} = (W^T \Sigma^{-1} W)^{-1} W^T \Sigma^{-1} h \]  
(7)

By letting \( \xi = W \theta \), the minimal \( Q \) is
\[ \hat{Q} = n(h - \hat{\xi})^T \Sigma^{-1}(h - \xi) = nh^T A h \]  
(8)

where
\[ \hat{A} = \hat{\Sigma}^{-1}(I - \hat{R}) \]  
(9)
\[ \hat{R} = W(W^T \Sigma^{-1} W)^{-1} W^T \]

Again, using the properties of the quadratic forms, we know that \( \hat{Q} \) has a non-central chi-square distribution with degrees of freedom = \( \text{tr}(\hat{\Sigma} \hat{A}) \) and asymptotic non-centrality parameter \( \lambda = \xi^T \hat{\Sigma} \hat{A} \xi \), if and only if \( \hat{\Sigma} \hat{A} \) is idempotent. Under the null hypothesis, \( \text{tr}(\hat{\Sigma} \hat{A}) = s - q \) and \( \lambda = 0 \). Hence the asymptotic distribution of \( \hat{Q} \) is \( \chi^2(s - q) \). Using this distribution, we can reject the null hypothesis with a type I error if \( \hat{Q} > \chi^2_{\alpha}(s - q) \).

Gurland and Vahiya (1970) developed the non-null distribution for \( \hat{Q} \). Using this result, they were able to compute the power of the test for selective alternative distributions.
In the next section, the general procedure is adapted for two specific distributions, the normal and gamma.

**Normal Distribution**

Suppose one would like to test the following hypothesis

$$H_0: X_1, X_2, \ldots, X_n \sim F_X(x; \theta) \sim N(\mu, \sigma^2)$$

where $\theta = (\theta_1, \theta_2, \sigma^2)^T$, $\mu$ and $\sigma^2$ are unknown parameters. If we let

$$\xi = (\xi_1, \mu, \xi_2 = \frac{\log \theta_2}{\sigma^2}, \xi_3 = \mu, \xi_4 = \log(\frac{3}{\sigma^2}))^T$$

we have

$$\xi = W_0^\theta$$

where

$$\theta^\theta = (\theta_1, \theta_2), \quad \theta_2 = \log \theta_2$$

$$W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 2 \end{bmatrix}$$

The transformation $T$ from $\mu'$ to $\xi$ can be achieved in two steps; $T_1: \mu' \rightarrow \mu$

$T_2: \mu \rightarrow \xi$. Hence, $E$ in equation (5) becomes

$$E = J_2 J_1 G J^T J^T$$

(13)
where

\[ J_1 = (J_{mn}); \quad j = \frac{\partial \mu}{\partial \mu_m} \quad m, n = 1, 2, \ldots, s \]

\[ J_2 = (J_{uv}); \quad j = \frac{\partial \mu}{\partial \xi_v} \quad u, v = 1, 2, \ldots, s. \]

By assuming that \( \mu_1 = 0 \), \( J_1 \) and \( J_2 \) become

\[
J_1 = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
-3\theta_2 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]

(15)

\[
J_2 = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1/\theta_2 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & \frac{1}{3\theta_2^2}
\end{bmatrix}
\]

(16)

and equation (14), becomes

\[
\Sigma = \begin{bmatrix}
\theta_2 & 0 & 0 & 0 \\
0 & 2 & 0 & 4 \\
0 & 0 & 6\theta_2^2 & 0 \\
0 & 4 & 0 & 32/3
\end{bmatrix}
\]

(17)
Since $\theta_2$ is unknown, let $\hat{\theta}_2$ denote the usual maximum likelihood estimate. Then $E = \Sigma |_{\hat{\theta}_2 = \hat{\theta}_2}$. Now by computing $\hat{\theta}$ and $\hat{Q}$ in equation (7) and (8), one can test the hypothesis (10).

**Gamma Distribution**

Test the hypothesis

$$H_0: X_1, X_2, \ldots, X_n \sim F_X(x; \theta) \sim \Gamma(\theta_1, \theta_2)$$

where the density function for the gamma distribution $\Gamma(\theta_1, \theta_2)$ is

$$f_X(x; \theta_1, \theta_2) = \frac{e^{-\frac{x}{\theta_1}}}{\theta_2 \Gamma(\theta_1)} ; y = x/\theta_2$$

$$\theta_1, \theta_2 > 0.$$  

Since $\gamma = (j - 1)! \theta_1^{\theta_1} \theta_2^{\theta_2}$, the $j$th cumulant, we can express $\xi = W\theta^*$, where

$$\xi = (\xi_1 = \kappa_1, \xi_2 = \kappa_2, \xi_3 = \kappa_3^{-1}, \xi_4 = \kappa_4 \kappa_3^{-1})^T$$

$$W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 2 \\ 0 & 3 \end{bmatrix} \quad \theta^* = (\theta_1^* = \theta_1, \theta_2^* = \theta_2)$$

The transformation $T$ from $n'$ to $\xi$ can be obtained in two steps

$$T_1: n' \rightarrow K$$

$$T_2: K \rightarrow \xi$$

(22)
where \( \kappa = (\kappa_1, \kappa_2, \kappa_3, \kappa_4)^T \). In which case \( \Sigma \) becomes

\[
\Sigma = J_2 J_1 G J_1^T J_2^T
\]

where

\[
J_1 = \begin{bmatrix}
1 & 0 & 0 & 0 \\
J_{12} & 1 & 0 & 0 \\
J_{13} & J_{23} & 1 & 0 \\
J_{14} & J_{24} & J_{34} & 1
\end{bmatrix}
\]

\( J_{12} = -2 n_1' \)

\( J_{13} = -3n_1' \)

\( J_{23} = -3n_2 + 6n_1' \)

\( J_{24} = -6 \frac{1}{2} + 12(n_1')^2 \)

\( J_{34} = -4n_1' \)

\( n_j' = \frac{r(\theta_1 + j)}{r(\theta_1)} \theta_2' ; \ j = 1, 2, 3, 4, ... \)

\[
J_2 = \begin{bmatrix}
1 & 0 & 0 & 0 \\
-\kappa_2 & 0 & 0 & 0 \\
-\kappa_3 & 0 & -\kappa_2 & 0 \\
0 & -\kappa_4 & 0 & -\kappa_3
\end{bmatrix}
\]
Since $\theta_1, \theta_2$ are unknown, they can be estimated by $\hat{\theta}_1, \hat{\theta}_2$ where

\[
\hat{\theta}_2 = \frac{X}{\hat{\theta}_1}
\]

\[
\hat{\theta}_1 = y^{-1/4} \left(1 + (1 + \frac{4y}{3})^{1/2}\right)
\]

\[
y = \log \left(\frac{\bar{X}}{\text{GM}}\right)
\]

\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

\[
\text{GM} = \left(\prod_{i=1}^{n} X_i\right)^{1/n}
\]

By replacing $\hat{\theta}_1, \hat{\theta}_2$ in $\Sigma$, we can test the hypothesis (18) using $\hat{Q}$.

Results

In order to demonstrate the GMCS procedure, the procedure was used in three different experiments. The first was to simulate data from several different distributions and determine the test of fit. In the second example the procedure was analysed using meteorological data consisting of several different atmospheric variables. The third experiment consisted of analyzing a meteorological data set from a specified distribution function.
Experiment 1

In this experiment, random observations were simulated from many different distribution functions in order to demonstrate how robust the procedure is to varying sample sizes, shape parameters, etc. This part of the experiment was not meant to provide conclusive evidence that the (GMCS) procedure is better or worse than any other procedure, but was intended to point out any apparent deficiencies. The results have been summarized in Table 1. In this table, I have only included the results for fitting the true distribution, however, the procedure may have indicated that another distribution could have provided satisfactory fit. However, this is explainable since the Gamma and Extreme Value distribution can resemble many other distributions depending upon their shape parameters.
<table>
<thead>
<tr>
<th>True Distribution</th>
<th>Parameters</th>
<th>Sample Size</th>
<th>Estimated Parameters</th>
<th>$\hat{Q}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Gamma(y, \beta)$</td>
<td>$\gamma$</td>
<td>$\beta$</td>
<td>$\hat{\gamma}$</td>
<td>$\hat{\beta}$</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>10</td>
<td>1.2</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>1.0</td>
<td>.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>1.1</td>
<td>.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>1.1</td>
<td>.90</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>10</td>
<td>.97</td>
<td>.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>85</td>
<td>.88</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>1.18</td>
<td>.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>.83</td>
<td>.72</td>
</tr>
<tr>
<td>.5</td>
<td>1</td>
<td>10</td>
<td>1.99</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>.80</td>
<td>.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>1.02</td>
<td>.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>1.17</td>
<td>.91</td>
</tr>
<tr>
<td>$N(\mu, \sigma^2)$</td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
<td>$\hat{\mu}$</td>
<td>$\hat{\sigma^2}$</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>10</td>
<td>12.1</td>
<td>11.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>9.5</td>
<td>31.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>8.9</td>
<td>20.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>10.2</td>
<td>23.9</td>
</tr>
<tr>
<td>Extreme value, $\alpha, \beta$</td>
<td>$\alpha$</td>
<td>$\beta$</td>
<td>$\hat{\alpha}$</td>
<td>$\hat{\beta}$</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>10</td>
<td>5.01</td>
<td>1.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>5.04</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>5.04</td>
<td>.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>4.82</td>
<td>.85</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>10</td>
<td>2.90</td>
<td>.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>2.69</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>1.74</td>
<td>2.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>2.09</td>
<td>1.95</td>
</tr>
<tr>
<td>Exponential $\lambda$</td>
<td>$\lambda$</td>
<td>$\hat{\lambda}$</td>
<td>$\hat{Q}$</td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>10</td>
<td>.69</td>
<td>9.04*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>.56</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>.53</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>.42</td>
<td>4.15*</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>10</td>
<td>1.04</td>
<td>.33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>.83</td>
<td>.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1.28</td>
<td>.29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.1</td>
<td>2.90</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>10</td>
<td>2.60</td>
<td>4.9*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>1.89</td>
<td>1.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1.97</td>
<td>1.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.92</td>
<td>.35</td>
<td></td>
</tr>
</tbody>
</table>

* null hypothesis can be rejected at $\alpha = 0.5$ level
Experiment 2

In this experiment meteorological data sets from three southern United States locations were analysed. The first set consisted of monthly precipitation totals and monthly mean temperature for the years 1936-1975 for sites New Orleans, LA, Orlando, FL, and Daytona Beach, FL. The results for these data sets have been summarized in Tables 2 & 3, where the data sets are partitioned into five year intervals, each containing 60 observations. The second data set consists of daily (high temperature, maximum wind speed) for the three U.S. sites. The observations are partitioned into monthly intervals for the 1970-1971 data. The results are summarized in Tables 4 & 5. Tables 6 & 7 contain the results for test of fit for extreme monthly temperature and wind for the three U.S. locations.

It should be mentioned that the above data set was partitioned for the author's convenience rather than for meteorological interpretation.
**TABLE 2**

Monthly Total Precipitation

<table>
<thead>
<tr>
<th>Site**</th>
<th>Year</th>
<th>Normal</th>
<th>Exp</th>
<th>Gamma</th>
<th>Extreme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>μ</td>
<td>σ²</td>
<td>Q</td>
<td>λ</td>
<td>γ</td>
</tr>
<tr>
<td>I</td>
<td>1936-40</td>
<td>4.8</td>
<td>24</td>
<td>3.3</td>
<td>7.6*</td>
</tr>
<tr>
<td></td>
<td>41-45</td>
<td>4.5</td>
<td>12</td>
<td>1.3</td>
<td>8.2*</td>
</tr>
<tr>
<td></td>
<td>46-50</td>
<td>5.4</td>
<td>20</td>
<td>.8</td>
<td>5.3*</td>
</tr>
<tr>
<td></td>
<td>51-55</td>
<td>4.6</td>
<td>9</td>
<td>1.4</td>
<td>7.8*</td>
</tr>
<tr>
<td></td>
<td>56-60</td>
<td>4.7</td>
<td>8</td>
<td>.3</td>
<td>10.6*</td>
</tr>
<tr>
<td></td>
<td>61-65</td>
<td>4.6</td>
<td>7</td>
<td>.0</td>
<td>9.3*</td>
</tr>
<tr>
<td></td>
<td>66-70</td>
<td>4.4</td>
<td>8</td>
<td>.3</td>
<td>8.6*</td>
</tr>
<tr>
<td></td>
<td>71-75</td>
<td>5.8</td>
<td>10</td>
<td>.3</td>
<td>13.1*</td>
</tr>
</tbody>
</table>

| II     | 1936-40 | 4.2    | 13  | .7    | 3.7*    | 1.6   | .04   | .0    | 7.5   | 2.5   | 2.2    |
|        | 41-45   | 4.0    | 14  | 1.4   | 3.6*    | 1.6   | .04   | "     | 8.8   | 2.4   | 2.3    |
|        | 46-50   | 4.5    | 19  | .5    | .9      | 1.1   | .02   | "     | 7.7   | 3.0   | 3.9*   |
|        | 51-55   | 4.4    | 16  | .9    | .9      | 1.5   | .04   | "     | 8.0   | 2.7   | 2.7    |
|        | 56-60   | 3.4    | 9   | .1    | 2.0     | "     | "     | "     | 5.3   | 2.2   | 1.9    |
|        | 61-65   | 4.3    | 19  | 1.2   | 1.4     | 1.3   | .03   | "     | 9.0   | 2.9   | 3.5*   |
|        | 66-70   | 4.0    | 9   | 1.4   | 4.6*    | 1.7   | .04   | "     | 6.1   | 2.2   | 1.5    |
|        | 71-75   | 3.9    | 15  | 1.2   | 1.3     | 1.2   | .03   | "     | 7.8   | 2.6   | 2.8    |

| III    | 1936-40 | 3.8    | 7   | 1.5   | 6.1*    | 1.8   | .05   | .0    | 5.6   | 2.0   | 1.2    |
|        | 41-45   | 4.5    | 16  | .3    | 2.0     | 1.3   | .03   | "     | 7.4   | 2.8   | 3.0    |
|        | 46-50   | 4.4    | 13  | .3    | 3.5*    | 1.5   | .03   | "     | 7.1   | 2.6   | 2.3    |
|        | 51-55   | 4.1    | 20  | 1.9   | 2.2     | 1.3   | .03   | "     | 9.3   | 2.8   | 3.5*   |
|        | 56-60   | 3.9    | 10  | .3    | 3.3*    | 1.5   | .03   | "     | 6.2   | 2.3   | 1.9    |
|        | 61-65   | 3.9    | 9   | .3    | 9.9*    | 1.7   | .04   | "     | 6.1   | 2.2   | 1.5    |
|        | 66-70   | 3.9    | 14  | .8    | 1.3     | 1.2   | .03   | "     | 7.3   | 2.6   | 2.7    |
|        | 71-75   | 3.9    | 9   | .3    | 5.3*    | 1.7   | .05   | "     | 6.0   | 2.2   | 1.5    |

* null hypothesis can be rejected at α = .05 level

** I - New Orleans; II - Orlando; III - Daytona Beach
<table>
<thead>
<tr>
<th>Site**</th>
<th>Year</th>
<th>Normal</th>
<th>Exp</th>
<th>Gamma</th>
<th>Extreme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
<td>$Q$</td>
<td>$\lambda$</td>
</tr>
<tr>
<td></td>
<td>1936-40</td>
<td>69.7</td>
<td>116</td>
<td>0.01</td>
<td>24.1*</td>
</tr>
<tr>
<td></td>
<td>1936-40</td>
<td>71.0</td>
<td>69</td>
<td>0.01</td>
<td>24.6*</td>
</tr>
<tr>
<td></td>
<td>1936-40</td>
<td>69.7</td>
<td>116</td>
<td>0.01</td>
<td>24.7*</td>
</tr>
</tbody>
</table>

* null hypothesis can be rejected at $\alpha = .05$ level

** I - New Orleans; II - Orlando; III - Daytona Beach
### TABLE 4

**Daily Maximum Temperature**

<table>
<thead>
<tr>
<th>Site**</th>
<th>Date***</th>
<th>Normal</th>
<th>Exp</th>
<th>Gamma</th>
<th>Extreme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
<td>$\lambda$</td>
<td>$\hat{\gamma}$</td>
</tr>
<tr>
<td>I</td>
<td>1/70</td>
<td>47.5</td>
<td>141</td>
<td>.0</td>
<td>02</td>
</tr>
<tr>
<td></td>
<td>3/70</td>
<td>60.0</td>
<td>44</td>
<td>&quot;</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td>6/70</td>
<td>79.7</td>
<td>20</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>10/70</td>
<td>69.0</td>
<td>26</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>1/71</td>
<td>55.3</td>
<td>112</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>3/71</td>
<td>59.4</td>
<td>75</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>6/71</td>
<td>80.2</td>
<td>9</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>10/71</td>
<td>71.8</td>
<td>23</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td>II</td>
<td>1/70</td>
<td>55.9</td>
<td>94</td>
<td>.0</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td>3/70</td>
<td>76.2</td>
<td>2</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>6/70</td>
<td>83.9</td>
<td>1</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>10/70</td>
<td>63.5</td>
<td>66</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>1/71</td>
<td>64.3</td>
<td>87</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>3/71</td>
<td>72.0</td>
<td>53</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>6/71</td>
<td>83.4</td>
<td>3</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>10/71</td>
<td>71.6</td>
<td>17</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td>III</td>
<td>1/70</td>
<td>54.7</td>
<td>94</td>
<td>.0</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td>3/70</td>
<td>65.6</td>
<td>53</td>
<td>&quot;</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td>6/70</td>
<td>80.9</td>
<td>8</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>10/70</td>
<td>82.7</td>
<td>12</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>1/71</td>
<td>58.8</td>
<td>92</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>3/71</td>
<td>60.1</td>
<td>65</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>6/71</td>
<td>71.0</td>
<td>8</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>10/71</td>
<td>76.0</td>
<td>9</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
</tbody>
</table>

* null hypothesis can be rejected at $\alpha = .05$ level

** I - New Orleans; II - Orlando; III - Daytona Beach

*** data set consists of daily observation for a monthly interval, only these selected months are presented.
### TABLE 5

**Daily Maximum Wind**

<table>
<thead>
<tr>
<th>Site</th>
<th>Date***</th>
<th>Normal</th>
<th>Exp</th>
<th>Gamma</th>
<th>Extremely</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>μ</td>
<td>σ</td>
<td>Q</td>
<td>λ</td>
</tr>
<tr>
<td>I</td>
<td>1/70</td>
<td>9.6</td>
<td>7</td>
<td>.0</td>
<td>.10</td>
</tr>
<tr>
<td></td>
<td>3/70</td>
<td>9.8</td>
<td>6</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>6/70</td>
<td>6.8</td>
<td>6</td>
<td>&quot;</td>
<td>.14</td>
</tr>
<tr>
<td></td>
<td>10/70</td>
<td>7.6</td>
<td>9</td>
<td>1.3</td>
<td>.13</td>
</tr>
<tr>
<td></td>
<td>1/71</td>
<td>8.4</td>
<td>12</td>
<td>.0</td>
<td>.11</td>
</tr>
<tr>
<td></td>
<td>3/71</td>
<td>9.7</td>
<td>7</td>
<td>&quot;</td>
<td>.10</td>
</tr>
<tr>
<td></td>
<td>6/71</td>
<td>5.3</td>
<td>2</td>
<td>&quot;</td>
<td>.18</td>
</tr>
<tr>
<td></td>
<td>10/71</td>
<td>4.7</td>
<td>5</td>
<td>.3</td>
<td>.20</td>
</tr>
<tr>
<td>II</td>
<td>1/70</td>
<td>9.6</td>
<td>10</td>
<td>.0</td>
<td>.10</td>
</tr>
<tr>
<td></td>
<td>3/70</td>
<td>10.3</td>
<td>10</td>
<td>&quot;</td>
<td>.04</td>
</tr>
<tr>
<td></td>
<td>6/70</td>
<td>8.4</td>
<td>4</td>
<td>&quot;</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>10/70</td>
<td>8.8</td>
<td>6</td>
<td>&quot;</td>
<td>.11</td>
</tr>
<tr>
<td></td>
<td>1/71</td>
<td>8.8</td>
<td>7</td>
<td>&quot;</td>
<td>.11</td>
</tr>
<tr>
<td></td>
<td>3/71</td>
<td>10.1</td>
<td>11</td>
<td>&quot;</td>
<td>.11</td>
</tr>
<tr>
<td></td>
<td>6/71</td>
<td>7.4</td>
<td>3</td>
<td>&quot;</td>
<td>.13</td>
</tr>
<tr>
<td></td>
<td>10/71</td>
<td>6.8</td>
<td>5</td>
<td>&quot;</td>
<td>.14</td>
</tr>
<tr>
<td>III</td>
<td>1/70</td>
<td>9.2</td>
<td>5</td>
<td>.0</td>
<td>.10</td>
</tr>
<tr>
<td></td>
<td>3/70</td>
<td>8.8</td>
<td>6</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>6/70</td>
<td>9.0</td>
<td>7</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>10/70</td>
<td>10.3</td>
<td>13</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>1/71</td>
<td>8.0</td>
<td>7</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>3/71</td>
<td>9.5</td>
<td>11</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>6/71</td>
<td>7.3</td>
<td>3</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
<tr>
<td></td>
<td>10/71</td>
<td>7.5</td>
<td>6</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
</tbody>
</table>

* null hypothesis can be rejected at α = .05 level

** I - New Orleans; II - Orlando; III - Daytona Beach

*** data set consists of daily observation for a monthly interval, only these selected months are presented.
### TABLE 6
Extreme Monthly Temperatures

<table>
<thead>
<tr>
<th>Site</th>
<th>Normal</th>
<th>Exponential</th>
<th>Gamma</th>
<th>Extreme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\mu}$</td>
<td>$\hat{\sigma}^2$</td>
<td>$\hat{\lambda}$</td>
<td>$\hat{\alpha}$</td>
</tr>
<tr>
<td>I</td>
<td>82.5</td>
<td>1.5</td>
<td>0.012</td>
<td>16.9*</td>
</tr>
<tr>
<td>II</td>
<td>82.9</td>
<td>1.3</td>
<td>0.012</td>
<td>16.9*</td>
</tr>
<tr>
<td>III</td>
<td>81.1</td>
<td>0.8</td>
<td>0.012</td>
<td>16.9*</td>
</tr>
</tbody>
</table>

* null hypothesis can be rejected at $\alpha = .05$ level

** I - New Orleans; II - Orlando; III - Daytona Beach

### TABLE 7
Extreme Monthly Winds

<table>
<thead>
<tr>
<th>Site</th>
<th>Normal</th>
<th>Exponential</th>
<th>Gamma</th>
<th>Extreme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\mu}$</td>
<td>$\hat{\sigma}^2$</td>
<td>$\hat{\lambda}$</td>
<td>$\hat{\alpha}$</td>
</tr>
<tr>
<td>I</td>
<td>11.1</td>
<td>17</td>
<td>.26</td>
<td>.09</td>
</tr>
<tr>
<td>II</td>
<td>11.2</td>
<td>11</td>
<td>.0</td>
<td>.08</td>
</tr>
<tr>
<td>III</td>
<td>10.3</td>
<td>9</td>
<td>.0</td>
<td>.09</td>
</tr>
</tbody>
</table>

* null hypothesis can be rejected at $\alpha = .05$ level

** I - New Orleans; II - Orlando; III - Daytona Beach
Experiment 3

In this section the procedure was applied to a data set found in Haggard et. al. (1973). In their paper, they analysed a meteorological data set consisting of maximum rainfall amounts in the Appalachian region resulting from tropical disturbances. In their paper they satisfactorily modeled the data set with a Gamma distribution. In this section, I wanted to determine if the GMCS procedure would indicate that the Gamma distribution would provide a satisfactory fit. Also, since the original authors were interested in making probabilistic forecasts, I have included the similar forecasts based upon the GMCS fitted distribution. The results for the test of fit are summarized in Table 7. Table 8 contains a comparison of the GMCS fitted Gamma distribution with the results found in Haggard et. al. (1964).
**TABLE 7**

GMCS Procedure for Maximum Rainfall within the Appalachians

<table>
<thead>
<tr>
<th>Data Set**</th>
<th>Normal</th>
<th>Exp</th>
<th>Gamma</th>
<th>Extreme</th>
<th>Haggard et. al. Result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\mu} )</td>
<td>( \hat{\sigma}^2 )</td>
<td>( \hat{\lambda} )</td>
<td>( \hat{\gamma} )</td>
<td>( \hat{\beta}^{-1} )</td>
</tr>
<tr>
<td>A</td>
<td>7.29</td>
<td>0.3</td>
<td>1.75</td>
<td>0.14</td>
<td>5.14*</td>
</tr>
<tr>
<td>B</td>
<td>8.08</td>
<td>53.5</td>
<td>1.24</td>
<td>0.12</td>
<td>4.70*</td>
</tr>
<tr>
<td>C</td>
<td>9.37</td>
<td>55.6</td>
<td>0.42</td>
<td>0.10</td>
<td>3.40*</td>
</tr>
<tr>
<td>D</td>
<td>10.18</td>
<td>55.3</td>
<td>0.32</td>
<td>0.09</td>
<td>3.90*</td>
</tr>
</tbody>
</table>

| A'         | 7.18   | 39.7| 1.23  | 0.13    | 5.05*                  | 2.1    | 3.4  | 0.06 | 14.2 | 4.0 | .02 | 2.2 | 3.1  |
| B'         | 7.94   | 41.8| 0.86  | 0.12    | 4.78*                  | 2.4    | 3.4  | 0.04 | 14.7 | 4.2 | .02 | 2.9 | 2.6  |
| C'         | 9.2    | 47.9| 0.26  | 0.10    | 3.73*                  | 1.9    | 4.8  | 0.02 | 14.5 | 4.9 | .04 | 2.0 | 4.5  |
| D'         | 10.0   | 46.5| 0.18  | 0.09    | 4.24*                  | 2.3    | 4.3  | 0.00 | 15.2 | 4.9 | .02 | 2.7 | 3.6  |

* null hypothesis can be rejected at \( \alpha = .05 \) level

** A - maximum 24-hour precipitation all storms.  B - maximum 24-hour precipitation from no more than one storm per year.  C - maximum precipitation totals from all storms.  D - maximum precipitation totals from no more than one storm per year.  A' - D' - same as A - D except using 27 inches for Camille rather than 31 inches.
TABLE 8

Expected Probabilities of Exceeding Arbitrary Precipitation Amounts Over the Appalachian Region

<table>
<thead>
<tr>
<th>Precipitation in inches</th>
<th>A I*</th>
<th>B II</th>
<th>C I</th>
<th>C II</th>
<th>D I</th>
<th>D II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.976</td>
<td>.966</td>
<td>.992</td>
<td>.980</td>
<td>.980</td>
<td>.980</td>
</tr>
<tr>
<td>2</td>
<td>.909</td>
<td>.890</td>
<td>.954</td>
<td>.926</td>
<td>.926</td>
<td>.930</td>
</tr>
<tr>
<td>3</td>
<td>.817</td>
<td>.797</td>
<td>.887</td>
<td>.850</td>
<td>.850</td>
<td>.863</td>
</tr>
<tr>
<td>4</td>
<td>.716</td>
<td>.698</td>
<td>.801</td>
<td>.764</td>
<td>.764</td>
<td>.788</td>
</tr>
<tr>
<td>5</td>
<td>.615</td>
<td>.602</td>
<td>.705</td>
<td>.674</td>
<td>.674</td>
<td>.705</td>
</tr>
<tr>
<td>6</td>
<td>.519</td>
<td>.513</td>
<td>.607</td>
<td>.587</td>
<td>.587</td>
<td>.632</td>
</tr>
<tr>
<td>7</td>
<td>.433</td>
<td>.432</td>
<td>.513</td>
<td>.505</td>
<td>.505</td>
<td>.559</td>
</tr>
<tr>
<td>8</td>
<td>.357</td>
<td>.362</td>
<td>.427</td>
<td>.431</td>
<td>.431</td>
<td>.490</td>
</tr>
<tr>
<td>9</td>
<td>.292</td>
<td>.301</td>
<td>.351</td>
<td>.364</td>
<td>.364</td>
<td>.427</td>
</tr>
<tr>
<td>10</td>
<td>.237</td>
<td>.248</td>
<td>.286</td>
<td>.306</td>
<td>.306</td>
<td>.371</td>
</tr>
<tr>
<td>15</td>
<td>.077</td>
<td>.090</td>
<td>.091</td>
<td>.118</td>
<td>.118</td>
<td>.172</td>
</tr>
<tr>
<td>20</td>
<td>.023</td>
<td>.030</td>
<td>.024</td>
<td>.042</td>
<td>.042</td>
<td>.075</td>
</tr>
<tr>
<td>25</td>
<td>.006</td>
<td>.009</td>
<td>.006</td>
<td>.014</td>
<td>.014</td>
<td>.031</td>
</tr>
<tr>
<td>30</td>
<td>.002</td>
<td>.003</td>
<td>.001</td>
<td>.005</td>
<td>.005</td>
<td>.010</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A' I</th>
<th>B' II</th>
<th>C' I</th>
<th>C' II</th>
<th>D' I</th>
<th>D' II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.978</td>
<td>.972</td>
<td>.993</td>
<td>.985</td>
<td>.981</td>
</tr>
<tr>
<td>2</td>
<td>.913</td>
<td>.900</td>
<td>.959</td>
<td>.934</td>
<td>.932</td>
</tr>
<tr>
<td>3</td>
<td>.821</td>
<td>.806</td>
<td>.893</td>
<td>.868</td>
<td>.865</td>
</tr>
<tr>
<td>4</td>
<td>.717</td>
<td>.704</td>
<td>.806</td>
<td>.768</td>
<td>.789</td>
</tr>
<tr>
<td>5</td>
<td>.613</td>
<td>.603</td>
<td>.706</td>
<td>.673</td>
<td>.710</td>
</tr>
<tr>
<td>6</td>
<td>.515</td>
<td>.510</td>
<td>.605</td>
<td>.580</td>
<td>.631</td>
</tr>
<tr>
<td>7</td>
<td>.427</td>
<td>.425</td>
<td>.507</td>
<td>.492</td>
<td>.556</td>
</tr>
<tr>
<td>8</td>
<td>.349</td>
<td>.352</td>
<td>.418</td>
<td>.413</td>
<td>.486</td>
</tr>
<tr>
<td>9</td>
<td>.283</td>
<td>.288</td>
<td>.340</td>
<td>.344</td>
<td>.423</td>
</tr>
<tr>
<td>10</td>
<td>.227</td>
<td>.235</td>
<td>.273</td>
<td>.284</td>
<td>.365</td>
</tr>
<tr>
<td>15</td>
<td>.070</td>
<td>.078</td>
<td>.079</td>
<td>.098</td>
<td>.165</td>
</tr>
<tr>
<td>20</td>
<td>.019</td>
<td>.023</td>
<td>.020</td>
<td>.031</td>
<td>.070</td>
</tr>
<tr>
<td>25</td>
<td>.005</td>
<td>.006</td>
<td>.002</td>
<td>.004</td>
<td>.028</td>
</tr>
<tr>
<td>30</td>
<td>.001</td>
<td>.002</td>
<td>.001</td>
<td>.003</td>
<td>.011</td>
</tr>
</tbody>
</table>

* 1- Haggard et al. Gamma distribution; II- GMCS Gamma distribution.

** Same as Table 7
Conclusions

A goodness of fit procedure based upon the theoretical work of Dahiya and Gurland [(1970a), (1970b), (1972)] is presented. The procedure has been documented in the computer software package (Appendix A). Several examples using meteorological data sets are analysed using this procedure. The principle advantages of this procedure over existing goodness-of-fit tests lies in the ability to test for several distributions using a single user setup. This advantage stems from the freedom of testing a distribution without having to specify all the unknown parameters of the tabular values of the null distribution.

References


Appendix A

User setup for Gurland's (GMCS) procedure

### JOB CONTROL PARAMETERS

<table>
<thead>
<tr>
<th>CARD</th>
<th>COL</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-5</td>
<td>IUNIT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>INPUT DEVICE for DATA.</td>
</tr>
<tr>
<td>6-10</td>
<td>NOB</td>
<td>Number of observations to be fitted.</td>
</tr>
<tr>
<td>15</td>
<td>ICOR</td>
<td>ICOR = 0.</td>
</tr>
<tr>
<td>20</td>
<td>IDIST</td>
<td>1 NORMAL distribution fitted.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 NORMAL distribution not fitted.</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>1 Exponential fitted.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 Exponential not fitted.</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>1 Gamma distribution fitted.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 Gamma distribution not fitted.</td>
</tr>
<tr>
<td>35</td>
<td></td>
<td>1 Extreme value distribution fitted.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 Extreme value distribution not fitted.</td>
</tr>
<tr>
<td>2</td>
<td>1-80</td>
<td>NFORMT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Format for input raw data.</td>
</tr>
<tr>
<td>3+</td>
<td></td>
<td>Input raw data</td>
</tr>
</tbody>
</table>
Program Description

MAIN - main program; input job parameters
GCALC - calculates the coefficients for matrix G.
RHAT1 - calculates the matrix \( \hat{R} \) for exponential dist.
RHAT2 - calculates the matrix \( \hat{R} \) for other dist.
TRIPLE - calculates matrix product \( x \times y \times z \).
AHAT - calculates matrix \( \hat{A} \).
QHAT - calculates matrix \( \hat{Q} \).
GREXTR - performs goodness of fit for extreme value distribution.
GRNORM - performs goodness of fit for normal dist.
GEXPO - performs goodness of fit for exponential dist.
GRGAMM - performs goodness of fit for gamma dist.
DGMPRD - IBM matrix multiplication
DMIN - IBM matrix inversion
Subroutines Needed By A Given Routine

MAIN - GRNORM, GREXPO, GREXTR, GRGAMM
GCALC -
RHAT1 - DGMPRD
RHAT2 - DGMPRD, DMINV
TRIPLE - DGMPRD
AHAT - DGMPRD
QHAT - DGMPRD
GREXTR - GCALC, TRIPLE, DMINV, RHAT 1, AHAT, QHAT, DGMPRD
GRNORM - GCALC, TRIPLE, DMINV, RHAT 2, AHAT, QHAT, DGMPRD
GREXPO - Same as GREXTR
GRGAMM - Same as GRNORM
FURTRAN IV G LEVEL

DATE = 78192

14

0001 IMPLICIT REAL*8 (A-H, O-Z)
0002 DIMENSION A(8), CMUL(8), CNRL(5), G(4,4), X(1000), IDIST(10),
0003 & HFURMT(20), XJ1(4,4), EXJB(1000)

0005 9000 READ(5,1,EN=9999) IUNIT, NOB, (IDIST(I), I=1,5)
0007 1 FORMAT (915)

0008 READ(I,N2) (HFURMT(I), I=1,20)
0009 2 FORMAT (20A4)

0010 IF (IUNIT - Equ. 5) GO TO 9004
0011 READ(IUNIT, HFURMT) (X(J), J = 1, NOB)
0012 DO 152 I=1, NOB, 12
0013 XMAX = X(1)

0014 XMIV = X(11)
0015 K = I+1
0016 L = I+11

0017 DO 151 J = K, L
0018 IF (X(J) .GT. XMAX) XMAX = X(J)

0019 IF (X(J) .LT. XMIN) XMIN = X(J)

0020 151 CONTINUE
0021 152 WRITE(6,153) XMAX, XMIN
0022 153 FORMAT (25, 8F5.1)

0023 ICHECK = 0

0024 DO 111 J = 1, NOB
0025 111 IF (X(J) .LE. 0.) ICHECK = 1

0026 WRITE(6,125)
0027 WRITE(6,123) (X(J), J=1, NOB)

0028 XDIV = DFL3AT(NOB)
0029 XMEAN = 0.0

0030 XVAR = 0.0
0031 SVAR = 0.0

0032 SUM = 0.0
0033 XM3 = 0.0
0034 XM4 = 0.0
0035 SM2 = 0.0
0036 SM3 = 0.0
0037 SM4 = 0.0

0038 PI = 3.1415926
0039 SDIV = XDIV - 1.

0040 DO 9001 I = 1, NOB
0041 XMEAN = XMEAN + X(I) / XDIV

0042 SUM = SJM + OABSIX(I))
0043 IF (SUM .LE. 0.) SUM=0.1

0044 SD = DLUG(SUM) / XDIV
0045 XGEJM = DEXP(SD)

0046 9001 CONTINUE
0047 DO 9002 I = 1, NOB

0048 XVAR = XVAR + (X(I) - XMEAN)**2 / XDIV
0049 XM3 = XM3 + (X(I) - XMEAN)**3 / XDIV
0050 XM4 = XM4 + (X(I) - XMEAN)**4 / XDIV

0051 SM2 = SM2 + X(I)**2 / XDIV
0052 SM3 = SM3 + X(I)**3 / XDIV
0053 SM4 = SM4 + X(I)**4 / XDIV

0054 STD = DSRT(XVAR)
0055 9002 CONTINUE
LOOP TILL ALL DISTRIBUTION REQUESTS HAVE BEEN SATISFIED

C

0096  DO 9003 1 = 4
0097  IF ((IDIST(I) .LE. 0) .OR. (IDIST(I) .GT. 4)) GO TO 9003
0098  IDUM = IDIST(I)
0099  GO TO (11,12,13,14), IDUM

C

0100 4 URMAL  IDIST = 1

0101 11 CALL GNVCRM(XM3)
0102  GO TO 9003

C

0103  EXPONENTIAL  IDIST = 2

0104 12 CALL GNEXPO(SM2,SM3,SM4,X)
0105  GO TO 9003

C

0106  GAMMA  IDIST = 3

0107 13 IF (ICHECK .EQ. 0) CALL GRIA44(X,SM2,SM3,SM4)
0108  WRITE (5,121) ICHECK
0109 121 FORMAT (10X,25(12,1X))
0110  GO TO 9003

C

0111  EXTREME VALUE  IDIST = 4

0112 14 CALL GREXTR(X)
0113  GO TO 9003

C

0114  BIVARIATE NORMAL  IDIST(1) = 5

0115 9004 READ(IUNIT,NFORMTM) ((X((J-1)*2+1), X((J-1)*2+2)), J=1,NOB)
0116  CALL BIVAR(X,NUB,IUNIT)
0117  FORMAT(T25,'F12.5)
0118  FORMAT(1411//1HO,T51,'THE OBSERVATIONS',///)
0119  WRITE(6,25)
0120 25 FORMAT('I*)
0121  REWIND 9
0122  READ(9,15,END=20) LINE
0123 15 FORMAT(33A4)
0124  WRITE(6,15) LINE
0125  GO TO 10
0126 20 END
SUBROUTINE SCALC(I,ICOR)

CALCULATE G FOR FIRST FOUR DISTRIBUTIONS

IMPLICIT REAL*8 (A-H,O-Z)
DIMENSION RAW(8),G(4,4),CUML(8),CENRL(8),A(1000),B(1000)
COMMON /MOMENT/ RAW,CUML,CENRL,GNU

XN = DFLUAT(NUB)
DO 100 I = 1,N
DO 100 J = 1,N
GI(J) = RAW(I)*RAW(J) - RAW(I)*RAW(J)
100 CONTINUE
RETURN
END

SUBROUTINE RHAT1(R,SIGI,R)

CALCULATE VECTOR HAT FOR EXPONENTIAL DISTRIBUTION

IMPLICIT REAL*8 (A-H,O-Z)
DIMENSION W(4),SIGI(4,4),DUM(4),X(1),FJUR(4,4)
CALL DGMPRD(W,SIGI,DUM(1,1,4))
CALL DGMPRD(DUM,W,X,1,1,4)
X(1) = 1.0 / X(1)
CALL DGMPRD(X,W,DUM,1,1,4)
CALL DGMPRD(W,DUM,FJUR,1,1,4)
CALL DGMPRD(FOUR,SIGI,R,4,4,4)
RETURN
END

SUBROUTINE RHAT2(R,SIGI,R)

CALCULATE R HAT MATRIX(4X2) FOR GAMMA, NEG BIN, NORMAL

IMPLICIT REAL*8 (A-H,O-Z)
DIMENSION h(4,2),SIGI(4,4),R(4,4),MT(2,2),DUM(2,4),X(2,2),
& FUUK(4,4),M(2,1,2)
DO 9000 I = 1,2
DO 9000 J = 1,4
MT(I,J) = W(J,I)
9000
CALL DGMPRD(WT,SIGI,DUM,2,4,4)
CALL DGMPRD(DUM,W,MT,2,4,2)
CALL DINV(X,2,DETP,L,M)
CALL DGMPRD(W,DUM,FUUK,4,2,4)
CALL DGMPRD(FUUK,SIGI,R,4,4,4)
RETURN
END
SUBROUTINE AHAAT(SIGI,R,A)
C    CALCULATE A HAT
C
IMPLICIT REAL*8 (A-H, O-Z)
DIMENSION SIGI(4,4),R(4,4),A(4,4),RI(4,4)
DO 1 I = 1,4
DO 1 J = 1,4
RI(I,J) = -R(I,J)
IF ( I .NE. J ) GO TO 1
RI(I,J) = RI(I,J) + 1.0
1 CONTINUE
CALL DGMPRD(SIGI,RI,A,4,4,4)
RETURN
END

SUBROUTINE TRIPLE(X,Y,Z)
C    CALCULATE X * Y * X TRANSPOSED AND RETURN VALUE IN Z
C
IMPLICIT REAL*8 (A-H, O-Z)
DIMENSION X(4,4),Y(4,4),Z(4,4),DUM(4,4),XT(4,4)
DO 1 J = 1,4
CALL DGMPRD(X,Y,DUM,4,4,4)
CALL DGMPRD(DUM,XT,Z,4,4,4)
RETURN
END

SUBROUTINE QHAT(XN,H,A,Q)
C    CALCULATE CHI-SQUARE Q HAT
C
IMPLICIT REAL*8 (A-H, O-Z)
DIMENSION H(4),A(4,4),DUM(4,4),XX(1)
CALL DGMPRD(H,A,DUM,1,4,4)
CALL DGMPRD(DUM,H,XX,1,4,4)
Q = XX(1) * XN
RETURN
END
SUBROUTINE GREXTR(X)

C GURLAND ROUTINE FOR EXTREME VALUE DISTRIBUTION

C

IMPLICIT REAL*8 (A-H,O-Z)

DIMENSION XJL(1:4),RAW(8),CUML(8),XMEAN(8),XVAR(8),XGEO(8),XUST(8)

C XJL(1:4), SIGI(1:4),L(1:4),M(1:4),THETA(1:4),R(1:4),A(1:4)

C

CUMM(0) = BATH(0), TOXUM(0)

C

X001 = DFLOAT('UB)

C

C3CALCULATE EXTREME CUMULANT MOMENTS

C

C PUT CUML(1-1) IN PLACE OF CUML(11) IN ORDER TO MAKE THE SAME

C SUBSCRIPTS OF H-VECTOR AS THAT OF THE JACOBIAN MATRIX

C

ESUM = 0.0

C

BETA = CSORT(SIX) * STD / PI.

B = BETA

DO 1 I = 1,NOB

1 ESMU = ESUM + DEXP(X(I)*B)

ALPHA = B * DLOG(ESUM) - B * DLOG(XDIV)

EMEAN = ALPHA - 0.5772168

EMODE = ALPHA

EVAR = 0.1 ** 2 * B ** 2 / 6.

CUML(1) = 1.045*B**2.

CUML(2) = 2.396*B**3.

CUML(3) = 6.494*B**4.

CUML(4) = 24.886*B**5.

CUML(5) = 122.078*B**6.

CUML(6) = 726.019*B**7.

CUML(7) = 5360.545*B**8.

C

C

C1 = CUML(1)

C2 = CUML(2)

C3 = CUML(3)

C4 = CUML(4)

C5 = CUML(5)

C6 = CUML(6)

C7 = CUML(7)

C

RAW(1) = XMEAN

C

RAW(2) = C1 + XMEAN**2

C

RAW(3) = C2 + 3.*C1*XMEAN + XMEAN**3

C

RAW(4) = C3 + 4.*C2*XMEAN + 3.*C1**2 + 5.*C1*XMEAN**2

C

RAW(5) = C4 + 5.*C3*XMEAN + 10.*C2*C1 + 10.*C2*XMEAN**2

C

RAW(6) = C5 + 6.*C4*XMEAN + 15.*C3*C1 + 15.*C3*XMEAN**2

C

RAW(7) = C6 + 7.*C5*XMEAN + 21.*C4*C1 + 21.*C4*XMEAN**2

C

RAW(8) = C7 + 8.*C6*XMEAN + 28.*C5*C1 + 28.*C5*XMEAN**2

C

C
CALL GCALC (ICOR1)

CALL INITIALIZE W

CALL INITIALIZE H

CALL INITIALIZE J1

DO 120 I=1,4
120 IF((.1.E2.J).OR.((1-1).EQ.J)) GO TO 120
XJ1(I) = xJ1(I-1)
12) CONTINUE

CALL TRIPLE(XJ1,G,SIG1)
CALL DMINV(SIG1,DET, DET, Chile)
CALL WHAT(LS2,SIG1)
CALL WHAT(LS1+1)
CALL DGPRD(K,theta,theta,4,4,1)

WRITE(6,126) X(J), J=1,NBR
```fortran
&TRAN IV G LEVEL 21

GEXTR

DATE = 7/8192

14/47/

070 WRITE (6, 123) EMEAN
071 WRITE (6, 124) ESTU
072 WRITE (6, 126) EMODE
073 WRITE (6, 127) EVAR
074 WRITE (6, 128) XVAR
075 WRITE (6, 129) XMEAN
076 WRITE (9, 130) ALPHA, B, Q
077 WRITE (5, 121) ALPHA, B, Q

078 121 FORMAT (/S, 125, ' PARAMETERS: ALPHA = ', E15.5, ', BETA = ', E15.5,
&            /S, 126, ' *** (CHI-SQUARE VALUE) *** ', E15.5)
079 122 FORMAT (T35, 5F10.5)
080 123 FORMAT (//, T37, ' THE MEAN OF THE EXTREME VALUES IS ', F15.7, '')
081 127 FORMAT (//, T37, ' VARIANCE OF THE EXTREME VALUES IS ', F15.7, '')
082 124 FORMAT (//, T37, ' THE STANDARD DEVIATION IS ', F8.3, ', F15.7, '')
083 123 FORMAT (1H17//1H6, T51, ' THE OBSERVATIONS ', /H, 01H41, ' GURLANDS PROCEDURE FOR EXTREME VALUES ', //)
084 126 FORMAT (//, T37, ' THE MODE OF THE EXTREME VALUES IS ', F15.7, '')
085 129 FORMAT (//, T37, ' THE SAMPLE VARIANCE IS ', F10.3, '', F15.7, '')
086 129 FORMAT (//, T37, ' THE SAMPLE MEAN IS ', F8.3, ', F15.7, '')
088 RETURN
089 END
```
SUBROUTINE GREXPUI(SM1, SM2, SM3, SM4, XI)

C GURT AND ROUTINE FOR EXPONENTIAL DISTRIBUTION

C IMPLICIT REAL*8 (A-H, O-Z)
C
C DIMENSION XI(4+4), RAV(8), CUMUL(8), CENKL(8), G(4+4), W(4+4),
C (14+4), U1M(4+4), SIGI(4+4), L(4+4), 4(4+4), THETA(4),
C (14+4: A14+4, X(1000))
C
C COMMON /NUMER/ XDIV, XMAX, XYAX, XGEOM, UNIT, ICOR, PI, STD
C COMMON /MUMENT/ RAW, CUMUL, CENKL, CM, NOB
C WRITE(6,10), UNIT
C
C FORMAT(///*, EXPONENTIAL DISTRIBUTION WITH DATA FROM UNIT *
C (13///))
C
C XN = DFLGAT(INOB)
C ZERJ = 0.0
C JNE = 1.0
C
C CALCULATE EXPONENTIAL MUMENTS
C
C RAW(1) = XMEAN
C RAW(2) = 2 * XMEAN**2
C RAW(3) = 6 * XMEAN**3
C RAW(4) = 24 * XMEAN**4
C RAW(5) = 120 * XMEAN**5
C RAW(6) = 720 * XMEAN**6
C RAW(7) = 5040 * XMEAN**7
C RAW(8) = 40320 * XMEAN**8
C CALL GCALC(ICOR)
C
C INITIALIZE W
C
C DO 9001 J = 1, 4
C 9001 W(I1) = 1
C
C INITIALIZE H
C
C H(I1) = RAW(1)
C H(I2) = SM2 / RAW(1)
C H(I3) = SM3 / SM2
C H(I4) = SM4 / SM3
C
C INITIALIZE J1
C
C DO 9001 J = 1, 4
C 9001 CONTINUE
C
C CALCULATE CHI-SQUARE TEST AND EXPONENTIAL PARAMETER
C
C
SUBROUTINE GKGAMM(X, SM2, SM3, SM4)

C C GURLAND ROUTINE FUR GAMMA DISTRIBUTION
C
IMPLICIT REAL*8 (A-H, O-Z)
DIMENSION XJ(4,4), XJ2(4,4), RAD(8), CUML(8), GENRL(8)
& S(4,4), W(4,2), H(4), DJM(4,4), SIGI(4,4), L(4), M(4),
& THETA(4), R(4,4), A(4,4), X(1000), SCUML(1000)

CUMON / MMTENT / RAW = CUM*, GENRL, G, NJ3
CUMON / NUMBER / XDIV, XMEAN, XVAR, XGEYM, IUNIT, ICOR, PI, STD
WRITE(6, 100) IUNIT
1000 FORMAT( ///, ', GAMMA DISTRIBUTION WITH DATA FROM UNIT *13,///)
XN = DFLAT(NUB)
ZERJ = 0.0
UNE = 1.0
IF (XMEAN .LE. 0.0) WRITE(6,21)
21 FORMAT( //, ', **NEGATIVE VALUES WRONG DISTRIBUTION*** *')

C C CALCULATE GAMMA MOMENTS
C
AX = DABS(XMEAN)
VY = DLOG1O(AX / XGEYM)
YYY = DABS(YY)
T1 = .5 * (1.0 / YYY) * (1.0 + DSQRT(1.0 * 1.333333333300 * YYY))
T2 = AX / T1
CUML(1) = T1 * T2
CUML(2) = T1 * T2**2
CUML(3) = T1 * T2**3 * 2.0
CUML(4) = T1 * T2**4 * 6.0
DO 351 I = 1, 8
351 CONTINUE
TLIMIT = 2.0 ** (-251)
XX = DFLAT(1)
IF ((T1 .GE. 57.57) .OR. (T1 .LE. TLIMIT)) GO TO 98
RAW(1) = DGMMA(T1 + XX) / DGMMA(T1) * T2**XX
GO TO 351
98 RAW(1) = (T1 + XX - 1.0) * T2 ** XX
351 CONTINUE

C C CALCULATE SAMPLE CUMULANTS FOR GAMMA
C
SCUML(1) = RAW(1)
SCUML(2) = SM2 - RAW(1)**2
SCUML(3) = SM3 - 3.0 * SM2 * RAW(1) + 2.0 * RAW(1)**3
SCUML(4) = SM4 - 4.0 * SM3 * RAW(1) - 3.0 * SM2**2 + 12.0 * SM2 * RAW(1)**2
& - 6.0 * RAW(1)**4

CALL GLCALC(ICOR)

W(1+1) = ONE
W(2+1) = ZERO
W(3+1) = ZERO
W(4+1) = ZERO
W(1+2) = ZERO
W(2+2) = ONE
W(3+2) = 2.0
W(4+2) = 3.0
C INITIALIZE M

0043 M(1) = SCUML(1)
0044 M(2) = SCUML(2) / SCUML(1)
0045 M(3) = SCUML(3) / SCUML(2)
0046 M(4) = SCUML(4) / SCUML(3)

C INITIALIZE J1

0047 DO 100 I = 1,4
0048 DO 100 J = 1,4
0049 IF (I .GT. J) GO TO 100
0050 IF (I .EQ. J) XJ1(I,J) = ONE
0051 IF (I .NE. J) XJ1(I,J) = ZERO
0052 100 CONTINUE
0053 XJ1(1,1) = -2 * RAW(1)
0054 XJ1(1,2) = -3 * RAW(2) + 6 * RAW(1)
0055 XJ1(1,3) = -4 * RAW(3) + 12 * RAW(3) * RAW(1) - 24 * RAW(1)**3
0056 XJ1(1,4) = -3 * RAW(1)
0057 XJ1(2,1) = -2 * RAW(1)
0058 XJ1(2,2) = -3 * RAW(2)
0059 XJ1(2,3) = 1.0 / CUML(1)
0060 XJ1(2,4) = 1.0 / CUML(2)
0061 XJ1(3,1) = -CUML(2) / CUML(1)**2
0062 XJ1(3,2) = -CUML(3) / CUML(2)**2
0063 XJ1(3,3) = -CUML(4) / CUML(3)**2
0064 XJ1(4,1) = ONE
0065 XJ1(4,2) = ONE / CUML(1)
0066 XJ1(4,3) = ONE / CUML(2)
0067 XJ1(4,4) = ONE / CUML(3)

C INITIALIZE J2

0068 XJ2(I,1) = -CUML(2) / CUML(1)**2
0069 XJ2(I,2) = -CUML(3) / CUML(2)**2
0070 XJ2(I,3) = -CUML(4) / CUML(3)**2
0071 101 CONTINUE

C CALCULATE CHI-SQUARE TEST AND GAMMA PARAMETERS

0072 CALL TRIPLE(XJ1,G,DOM)
0073 CALL TRIPLE(XJ2,DUM,SGI)
0074 CALL DMINV(SIGI,4,DET,L,M)
0075 CALL RHATZ(SIGI,R)
0076 CALL ANAT(SIGI,R,A)
0077 CALL QHAT(XX,H,A,Q)
0078 CALL QMPRD(R,H,THETA,4,4,1)
0079 XR = THE(TA(1) / THETA(2))
0080 XL = 1.0 / THETA(2)
0081 WRITE(6,123) XR, XL, Q
0082 WRITE(6,124) XR, XL, Q
0083 123 FORMAT(/,T259* PARAMETERS: K = ,E15.5,10X,* LAMDA= ,E15.5)
0084 124 FORMAT(/,T259* GAMMA PARAMETERS: R = ,F6.2,5X,* LAMDA= ,F6.2)
0085 RETURN
0086 END
SUBROUTINE GRNORM (X, M)

C GURLAND NORMAL DISTRIBUTION ROUTINE

IMPLICIT REAL*8 (A-H, O-Z)
DIMENSION X(10001), M(10001)

COMMON /NUMBER/ XDIV,XMEAN,XVAR,XGEOM,UNIT,ICOR,PI,STD
COMMON /MOMENT/ RAW,CUML,CENRL,S,NOS
WRITE(6,10001)UNIT
1000 FORMAT(/,10H NORMAL DISTRIBUTION WITH DATA FROM UNIT *13//)
XN = O.FLOAT(NUB)
ZERJ = 3.0

C CALCULATE NORMAL MOMENTS

CENRL(1) = ZERO
CENRL(2) = XVAR
CENRL(3) = ZERO
CENRL(4) = 3 * XVAR**2
RAW(1) = XMEAN
RAW(2) = XVAR + XMEAN**2
RAW(3) = 3 * XMEAN * XVAR + XMEAN**3
RAW(4) = 3 * XVAR**2 + 6 * XMEAN**2 + XVAR + XMEAN**4
RAW(5) = 15 * XVAR**2 * XMEAN + 10 * XVAR * XMEAN**3 + XMEAN**5
&
RAW(6) = 15 * XVAR**3 + 45 * XVAR * XMEAN**2 + 15 * XVAR * XMEAN**4 +
&
RAW(7) = 105 * XMEAN * XVAR**3 + 84 * XVAR**2 * XMEAN**3 +
&
RAW(8) = 420 * XVAR**3 * XMEAN**2 +
&
RAW(9) = 210 * XVAR**2 * XMEAN**4 + 84 * XVAR * XMEAN**5 + XMEAN**7

CALL GCALC(ICOR)

C INITIALIZE W

W(1,1) = ONE
W(2,1) = ZERO
W(3,1) = ZERO
W(4,1) = ZERO
W(5,1) = ZERO
W(6,1) = ONE
W(7,1) = ZERO
W(8,1) = ZERO
W(9,1) = ZERO
W(10,1) = ZERO
W(11,1) = ZERO

C INITIALIZE H

H(1) = XMEAN
H(2) = DLOG (CENRL(2))
H(3) = XMEAN
H(4) = DLOG (CENRL(4) / 3.0)

C INITIALIZE J1

DO 100 J = 1,4
100 IF (1 .GE. J) GO TO 100
IF (I .EQ. J) XJ1(I,J) = ONE
IF (I .NE. J) XJ1(I,J) = ZERO

CONTINUE

XJ1(2,1) = -2 * RAW(1)
XJ1(3,1) = -3*RAW(2) + 6*RAW(1)**2
XJ1(3,2) = -3 * RAW(1)
XJ1(4,1) = -4*RAW(3) + 12*RAW(2)*RAW(1) - 12*RAW(1)**3
XJ1(4,2) = 5*RAW(1)**2
XJ1(4,3) = -4 * RAW(1)

C

C INITIALIZE J2

DO 10 CI J = 1,4
DO 10 C J = 1,4

IF (I .EQ. J) GO TO 101
XJ2(I,J) = ZERO

CONTINUE

XJ2(1,1) = ONE
IF (XN .EQ. 1.) GO TO 8

7 XJ2(2,2) = 1. / (CENRL(2) * XN / (XN - 1.01))

XJ2(3,3) = ONE
XJ2(4,4) = XJ2(2,2)**2 / 3.0

GO TO 9

CALCULATE CHI-SQUARE TEST AND PARAMETERS

XN = XN + 1.
GO TO 7

CALL TRIPLE(XJ1,6,DM)
CALL TRIPLE(XJ2,6,DM,SIGI)
CALL DINV(SIGI,4,DETL,M)
CALL RHTZ(4*SIGI,R)
CALL AMAT(SIGI,R,A)
CALL QHAT(XN,H,A,Q)
CALL QMPROD(XH,THETA,4*4,1)
TVAR = EXP(THETA(2))
WRITE(6,123) THETA(1),TVAR,Q
WRITE(9,124) THETA(1), TVAR, Q

123 FORMAT //,T25,* PARAMETERS : MJ = *E15.5,10X,* SIGMA=*,E15.5
& //,T39,* ***(CHI-SQUARE VALUE)*** *E15.5)

124 FORMAT //,T25,* NORMAL PARAMETERS: MJ = *F6.2,5X,* SIGMA=*,F6.2
& //,T39,* *** (CHI-SQUARE VALUE) *** *F10.3)

RETJR4

END
CALL TRIPLE(XJ1,G,SIGI)
CALL DMINV(SIGI,R,DET,L,M)
CALL RHA1(SIGI,R)
CALL AHA1(SIGI,R,A)
CALL WHAT1(X4,H,A,T)
CALL DIMPDR1(R,H,THETA,T4,1)
XLAMDA = 1. / THETA11
WRITE(6,123) XLAMDA,Q
WRITE(9,124) XLAMDA,J
123 FORMAT(/,T25*: PARAMETERS : LAMDA = *,E15.5,/, & T39,*** (CHI-SQUARE VALUE )*** ,E15.5)
124 FORMAT(/,T25*: EXPONENTIAL PARAMETERS : LAMDA= *,F0.2,/, & T39,*** (CHI-SQUARE VALUE )*** ,F10.3)
RETURN
END
Effect of Correlated Observations on Confidence Sets Based Upon Chi-Square Statistics

Summary

This paper investigates how the presence of correlation in a multivariate sample effects the confidence coefficients of confidence sets based upon chi-square statistics.

I. Introduction

Basu et al. (1976) investigated the effect that simple equicorrelation within a multivariate normal sample has upon confidence sets based upon chi-square statistics. They suggested that their results could provide a useful application in the area of pattern recognition using remotely sensed LANDSAT data. However, several recent investigations have demonstrated that the equicorrelated correlation structure is not an appropriate model in the Landsat application. In fact, Tubbs and Coberly (1978) demonstrated that the correlation structure in the LANDSAT data is similar to observations obtained from a stationary autoregressive process. In this paper, I have investigated the effect that autocorrelated data have on confidence sets based upon chi-square statistics.

II. Basic Concepts

Let $X_1, \ldots, X_n$ denote a sample of $n$ p-dimensional normal observations with mean $\mu$ and common positive definite covariance matrix $\Sigma$. Suppose
that \( X = [X_1, X_2, \ldots, X_n]^T \) and that

\[
E[(X - E(X))(X - E(X)^T)] = \Gamma_n \otimes \Sigma
\] (1)

where \( \Gamma_n \) is a positive definite \( nxn \) matrix, \( A \otimes B \) denotes the Kronecker product of matrices \( A \) and \( B \), and \( E(\cdot) \) denotes the expectation operator. Note, if the sample \( X_1 \ldots X_n \) is random then \( \Gamma_n = I \), where \( I \) is an identity matrix.

Now suppose that the sample \( X_1 \ldots X_n \) is a realization from a discrete stationary time series \( \{X_t\} \) with continuous density function \( f_X(\cdot) \). If \( \Gamma_n \) denotes the autocorrelation matrix for \( n \) lags. That is,

\[
\Gamma_n = (\rho_{ij}) \text{ for } i,j = 1,2,\ldots,n
\] (2)

\[
\rho_{ij} = \text{corr}(X_i, X_j).
\]

It is well known [Fuller (1972)] that there exists an orthogonal matrix \( U \) such that

\[
U^*\Gamma_n U \sim 2\pi D_X
\] (3)

where

\[
D_X = \text{diag} (d_1, d_2, \ldots, d_n)
\]

\[
d_1 = f_X(0)
\]

\[
d_n = f_X(1)
\]

\[
d_{2k} = d_{2k+1} = f_X \left( \frac{2k}{n} \right); \text{ for } k = 1,2,\ldots,(n-1)/2.
\]
\[
\begin{bmatrix}
2^{-1/2} & 2^{-1/2} & \ldots & 2^{-1/2} \\
1 & \cos(2\pi/n) & \ldots & \cos(2\pi \frac{n-1}{n}) \\
0 & \sin(2\pi/n) & \ldots & \sin(2\pi \frac{n-1}{n}) \\
\vdots & \vdots & \ddots & \vdots \\
1 & \cos(\frac{n-1}{n} 2\pi/n) & \ldots & \cos(\frac{n-1}{n} 2\pi \frac{n-1}{n}) \\
0 & \sin(\frac{n-1}{n} 2\pi/n) & \ldots & \sin(\frac{n-1}{n} 2\pi \frac{n-1}{n})
\end{bmatrix}
\]

By letting
\[ Z = U\mathbf{x} \]

it follows that
\[
\mathbb{E}( (Z - \mathbb{E}(Z)) (Z - \mathbb{E}(Z))^T ) = \mathbf{D}_x \otimes \Sigma.
\]

Furthermore, it follows that
\[ Z_1 = n^{1/2} \overline{X}; \quad \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \]

where \[ Z = [Z_1 \ldots Z_n]^T. \] The distribution for \( Z_j \) is
\[
Z_1 \sim N(n^{1/2} \mu, \sigma^2 \Sigma) \\
Z_j \sim N(0, \sigma^2 \Sigma); \quad j = 2, 3, \ldots, n
\]
where the symbol ~ means "is distributed as". The expectation of $Z_j$ is zero since

$$E(Z_j) = E\left( \sum_{k=0}^{n-1} \cos\left(\frac{j-1}{n} 2\pi k/n\right) X_k \right) \quad (9)$$

or

$$= E\left( \sum_{k=0}^{n-1} \sin\left(\frac{j-1}{n} 2\pi k/n\right) X_k \right)$$

$$= \mu \left( \sum_{k=0}^{n-1} \cos\left(\frac{j-1}{n} 2\pi k/n\right) \right) \text{ or } \mu \left( \sum_{k=0}^{n-1} \sin\left(\frac{j-1}{n} 2\pi k/n\right) \right)$$

$$= 0.$$

Now let

$$Q_1(\mu) = n(\overline{X} - \mu)^T \Sigma^{-1} (\overline{X} - \mu)$$

$$Q_2 = \sum_{j=1}^{n} (X_j - \overline{X})^T \Sigma^{-1} (X_j - \overline{X}). \quad (10)$$

If $\Gamma_n = I$, it is well known that

$$Q_1(\mu) \sim \chi^2(p)$$

$$Q_2 \sim \chi^2(n-1)p \quad (11)$$

where $\chi^2(v)$ denotes a chi-square distribution with $v$ degrees of freedom.
However, if $\Gamma_n$ is given by (1) we have

$$Q_1(u) = n(\bar{x} - u)^T \Sigma^{-1} (\bar{x} - u)$$

$$= (n \bar{x} - n \mu)^T \Sigma^{-1} (n \bar{x} - n \mu)$$

$$= (z_1 - \mu)^T \Sigma^{-1} (z_1 - \mu)$$

$$= d_1 (z_1 - E(z_1))^T (d_1 \Sigma)^{-1} (z_1 - E(z_1)). \quad (12)$$

Hence

$$Q_1(u)/d_1 \sim \chi^2(p). \quad (13)$$

Now consider

$$Q_2 = \sum_{j=1}^{n} (x_j - \bar{x})^T \Sigma^{-1} (x_j - \bar{x})$$

$$= \text{tr} \Sigma^{-1} \left[ \sum_{j=1}^{n} (x_j - \bar{x}) (x_j - \bar{x})^T \right]$$

$$= \text{tr} \Sigma^{-1} \left[ \sum_{j=1}^{n} x_j x_j^T - n \bar{x} \bar{x}^T \right] \quad (14)$$

However, since $U$ is orthogonal (14) becomes

$$Q_2 = \text{tr} \Sigma^{-1} \left[ \sum_{j=1}^{n} z_j z_j^T - n \bar{x} \bar{x}^T \right]$$

$$= \text{tr} \Sigma^{-1} \left[ \sum_{j=2}^{n} z_j z_j^T \right]$$

$$= \sum_{j=2}^{n} z_j z_j^T \Sigma^{-1}$$

$$= \sum_{j=2}^{n} d_j w_j \quad (15)$$
for \( W_j = z_j^T (d_j \Sigma^{-1}) z_j \). We know that \( W_j \) has a chi-square distribution with 
p degrees of freedom and that \( W_i, W_j \) are independent for each
\( i \neq j = 2, 3, \ldots, n \).

III. Confidence Set for Mean

Let \( H_0 \) denote the null hypothesis that \( X_1, \ldots, X_n \) is a random sample from
a \( p \)-dimensional normal population with \( E(X) = \mu \), \( \text{cov}(X) = \Sigma \). The statistic \( Q_1 \),
as given in equation (10) is used to define a confidence set for the unknown
population mean \( \mu \). That is, let

\[
I_\varepsilon = \{ \mu : Q_1(\mu) \leq \chi^2_{\varepsilon}(p) \} \tag{16}
\]

where \( \chi^2_{\varepsilon}(p) \) is the 100 \( \varepsilon \) percentage point of \( \chi^2(p) \). Thus since
\( Q_1 \sim \chi^2(p) \) whenever \( H_0 \) is true, we know that

\[
P[\mu \in I_\varepsilon \mid H_0 \text{ true}] = \varepsilon. \tag{17}
\]

Let \( H_1 \) denote the alternative hypothesis that the sample satisfies equation
(1). If \( H_1 \) is true, then find the value \( \alpha \) such that

\[
P[\mu \in I_\varepsilon \mid H_1 \text{ true}] = \alpha. \tag{18}
\]
From equation (13), we know that \( a \) must satisfy the following relationship

\[ \chi_a^2(p) = \chi_c^2(p)/d_1 \]  \hspace{1cm} (19)

**IV. Confidence Interval for the Dispersion Scalar**

Let \( X_1 \ldots X_n \) denote a sample from a normal distribution with mean \( \mu \) and covariance matrix \( \sigma^2 \Sigma \), where \( \Sigma \) is a known positive definite matrix. Let \( \text{Ho} \) denote the hypothesis that the sample is random and \( \text{H}_1 \) denote the hypothesis that the sample satisfies equation (1). If \( \text{Ho} \) is true, then

\[ Q_2/\sigma^2 \sim \chi^2_{p(n-1)} \]  \hspace{1cm} (20)

where \( Q_2 \) is given by equation (10). Hence the interval

\[ 0 \leq \sigma^2 \leq Q_2/\chi^2_{p(n-1)} \]  \hspace{1cm} (21)

is a 100 \( \varepsilon \) confidence interval for \( \sigma^2 \). However, to find the confidence interval for \( \sigma^2 \) when \( \text{H}_1 \) is true, it is necessary to determine the distribution of \( Q_2 \). From equation (15) we obtain

\[ Q_2/\sigma^2 = \sum_{j=2}^{n} d_j W_j \]  \hspace{1cm} (22)

where \( W_j \), for \( j = 2,3,\ldots,n \) are distributed as independent chi-squares with \( p \) degrees of freedom. The distribution for (22) can be expressed in the
following series representation [c.f. Kotz, Johnson, and Boyd (1967)].

\[ P\left[ \frac{Q_2}{\sigma^2} \leq y \right] = \sum_{k=0}^{\infty} c_k G(v + 2k; y/\beta) \]  \hfill (23)

where \( G(v+2k; y/\beta) \) denotes the cumulative probability density function for a central chi-square with degrees of freedom \( v+2k \), and \( c_k, \beta \) are known functions of the \( d_j \)'s, for \( j=2,3,...,n \). Hence, whenever \( H_1 \) is true, the confidence interval for \( \sigma^2 \) in equation (21) is given by \( \alpha \) where \( \alpha \) is the value which satisfied the following relationship

\[ \alpha = \sum_{k=0}^{\infty} c_k G(n(n-1)+ 2k; \gamma^2/\beta) \]  \hfill (24)

where

\[ \gamma^2 = \chi_{\epsilon, p(n-1)}^2 \]

V. Examples

Suppose that \( X_1,...,X_n \) are a realization from a stationary autoregressive process of order one with parameter \( \phi \). Then the spectral density function is

\[ f_X(\omega) = \frac{1}{2 \pi (1+\phi^2 - 2\phi \cos \omega)} \]  \hfill (25)

Hence

\[ d_{2k} = (1+\phi^2 - 2\phi \cos(2k\pi/n))^{-1} \quad k=1,1/2,\ldots,n-1/2 \]  \hfill (26)

\[ d_1 = (1-\phi)^{-2} \]
The \( \alpha \) -values which satisfy equation (19) are given in Table 1 for \\
\( \varepsilon = .99, 95. \)

<table>
<thead>
<tr>
<th>P / ( \phi )</th>
<th>.0</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.9900</td>
<td>.9795</td>
<td>.9606</td>
<td>.9285</td>
<td>.8776</td>
<td>.8021</td>
</tr>
<tr>
<td></td>
<td>.9500</td>
<td>.9222</td>
<td>.8830</td>
<td>.8298</td>
<td>.7603</td>
<td>.6728</td>
</tr>
<tr>
<td>2</td>
<td>.9900</td>
<td>.9760</td>
<td>.9475</td>
<td>.8953</td>
<td>.8094</td>
<td>.6838</td>
</tr>
<tr>
<td></td>
<td>.9500</td>
<td>.9116</td>
<td>.8529</td>
<td>.7695</td>
<td>.6598</td>
<td>.5270</td>
</tr>
<tr>
<td>5</td>
<td>.9900</td>
<td>.9681</td>
<td>.9145</td>
<td>.8071</td>
<td>.6346</td>
<td>.4174</td>
</tr>
<tr>
<td></td>
<td>.9500</td>
<td>.8896</td>
<td>.2856</td>
<td>.6337</td>
<td>.4485</td>
<td>.2642</td>
</tr>
<tr>
<td>10</td>
<td>.9900</td>
<td>.9570</td>
<td>.8623</td>
<td>.6704</td>
<td>.4055</td>
<td>.1682</td>
</tr>
<tr>
<td></td>
<td>.9500</td>
<td>.8614</td>
<td>.6952</td>
<td>.4648</td>
<td>.2363</td>
<td>.0823</td>
</tr>
</tbody>
</table>
From Table 1, we observe that a 95% confidence ellipse is a 65.98% confidence ellipse if the sample \( X_1 \ldots X_n \) is a bivariate sample from an autoregressive process of order 1 with parameter \( \phi = .4 \).

<table>
<thead>
<tr>
<th>( N )</th>
<th>( p )</th>
<th>( \phi )</th>
<th>.0</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>1</td>
<td>.9500</td>
<td>.9326</td>
<td>.8759</td>
<td>.7901</td>
<td>.6913</td>
<td>.5938</td>
<td>.396</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.9143*</td>
<td>.8617</td>
<td>.7768</td>
<td>.6317</td>
<td>.4822</td>
<td>.3539</td>
<td>.1518</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.9144*</td>
<td>.8211</td>
<td>.5822</td>
<td>.3365</td>
<td>.1666</td>
<td>.0754</td>
<td>.0089</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>.9143*</td>
<td>.8742</td>
<td>.7577</td>
<td>.5996</td>
<td>.4386</td>
<td>.3020</td>
<td>.0902</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.0000*</td>
<td>.8935</td>
<td>.6452</td>
<td>.3869</td>
<td>.1998</td>
<td>.0927</td>
<td>.0081</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.0000*</td>
<td>.7547</td>
<td>.3344</td>
<td>.0934</td>
<td>.0178</td>
<td>.0026</td>
<td>.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>1</td>
<td>1.0000*</td>
<td>.8859</td>
<td>.6223</td>
<td>.3550</td>
<td>.1702</td>
<td>.0712</td>
<td>.0036</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.0000*</td>
<td>.7850</td>
<td>.3872</td>
<td>.1260</td>
<td>.0286</td>
<td>.0050</td>
<td>.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.0000*</td>
<td>.5460</td>
<td>.0933</td>
<td>.0056</td>
<td>.0001</td>
<td>.0000</td>
<td>.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>1</td>
<td>1.0000*</td>
<td>.7822</td>
<td>.3811</td>
<td>.1209</td>
<td>.0266</td>
<td>.0043</td>
<td>.0000</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.0000*</td>
<td>.6123</td>
<td>.1453</td>
<td>.0146</td>
<td>.0007</td>
<td>.0000</td>
<td>.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.0000*</td>
<td>.2932</td>
<td>.0080</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The specified level \( \epsilon = .9500 \)

From Table 2, a 99% confidence interval for \( \sigma^2 \) is a 19.98% confidence based upon a bivariate sample of 25 observations from an AR(1) process with \( \phi = .4 \).
VI CONCLUSIONS

It is well known in applications using atmospheric observations that the data are non-random and in fact are highly correlated. Very little research has been done in the area of determining the effect that correlated samples have upon statistical inference. In this paper, I have investigated the effect that samples taken from a stationary autoregressive process have upon the confidence regions for the parameters of a normal distribution. Tables are included for the effect that sampling from an AR(1) process have upon these confidence regions.

VII REFERENCES


GENERATION OF RANDOM VARIATES FROM SPECIFIED DISTRIBUTIONS

Summary

Due to the complexity of many of the existing statistical problems associated with atmospheric variables, computer simulations have proved to be a very informative technique. However, due to the various types of atmospheric data, thus the different type of statistical distributions one can no longer perform simulations based solely upon normal data. So in anticipating this problem, this paper presents the computer software for generating both random and correlated data for several specified distributions. A brief explanation of the procedure is given along with the program documentation.

I. INTRODUCTION

In order to obtain insight into some of the statistical problems with atmospheric data, it is necessary to be able to simulate some of the environmental situations. However, since most of the data are non-normal it is necessary to generate data from various specified distributions (e.g. Gamma, Beta, Negative Binomial, etc.). The purpose of this paper is to document the procedures used in generating both correlated and uncorrelated observations. The uncorrelated procedures have been documented in Newmann and Odell (1971). The correlated procedures have been compiled from numerous sources, however, Johnson and Kotz (1972) provide the primary reference. In this paper, I have included only a brief description of the statistical distributions. For a more detailed discussion see Falls (1971).
II. UNCORRELATED VARIATES

All of the procedures listed here are transformations of independent random variates from a uniform $U(0,1)$ distribution. The pseudo-random number generator used is a congruential generator (IBM SSP RANDU) whose choice was based solely upon convenience. However, some additional testing will be necessary to determine if the pseudo-random variates procedures are satisfactory for our purposes.

Continuous Distributions

2.1 Univariate Normal Distribution $N(\mu, \sigma^2)$

The Box-Muller transformation [1] has been used. It can be summarized in the following result.

Result: 2.1 If $u$ and $v$ are independently distributed $U(0,1)$ then,

\[
\begin{align*}
x &= (-2 \ln u)^{1/2} \cos 2\pi v \\
y &= (-2 \ln v)^{1/2} \sin 2\pi v
\end{align*}
\]

are independent random variates with the standardized normal distribution $N(0,1)$.

Thus if $u_1, \ldots, u_N$ is a sequence of independent $U(0,1)$ one can generate a sequence $x_1, \ldots, x_N$ of independent $N(0,1)$ using the above procedure. Also if $\sigma, \sigma$ is a fixed known constant, then $y_i = \sigma x_i + \mu$, $i=1,2,\ldots,n$ is a sequence of independent normal with mean $= \mu$, variance $= \sigma^2$. 
2.2 Multivariate Normal $N_p(\mu, \Sigma)$

Let $x_1, \ldots, x_p$ be a sequence of $p$ independent normals with mean $0$ and variance $1$, then $x = (x_1, \ldots, x_p)^T$ is said to be multivariate normal with mean $\mu$ and covariance matrix $I_p$ (pxp identity matrix). However, if $x \sim N_p(\mu, I_p)$ then $y = Bx + \mu$ has a multivariate normal distribution with mean $\mu$ and covariance matrix $\Sigma$, where $\Sigma = BB^T$. From $x$ we can find $y$ for any specified real positive definite symmetric matrix $\Sigma$. This follows from the following result.

Result: 2.2 Let $\Sigma$ be a real p.d. symmetric matrix. Then there exists a lower triangular matrix $B$ with positive elements on the main diagonal such that $\Sigma = BB^T$. This is often referred to as the Crout factorization of $\Sigma$.

2.3 Gamma Distribution $\Gamma(\lambda, k)$

Let $u_1, \ldots, u_k$ be a sequence of $k$ independent random variables each having a $U(0,1)$ distribution. Then

$$x = -\frac{1}{\lambda} \ln \prod_{i=1}^{k} u_i$$

is a gamma with parameters $\lambda$ and $k$. Note the chi-square distribution with $n$ degrees of freedom can be obtained by letting $k=n/2$ and $\lambda = 1$. Also, if $n$ is odd then $y = x + w^2$ is chi-square with d.f. = $n$ if $x \sim \Gamma(k = n-\frac{1}{2}, \lambda = \frac{1}{2})$ with $w \sim N(0,1)$. The exponential distribution with parameter $\lambda$ can also be obtained by letting $k=1$ in $(2)$. 
2.4 Beta Distribution $\beta(p, q)$

If $x_1 \sim \Gamma(1, p)$ and $x_2 \sim \Gamma(1, q)$ are independent then $y = x_1 / (x_1 + x_2)$ has a Beta distribution with parameters $p$ and $q$.

Discrete Distributions

If the distribution function $F_x$ is known then we can generate pseudo-random numbers by using the inverse function $F_x^{-1}$. However, this procedure can be simplified by letting $x$ be the random variate from $F_x$ which satisfied the relation $F_x(x-1) < u < F_x(x)$ where $u$ is a random variate having a $U(0, 1)$ distribution. This procedure could be used to generate Binomials, since the distribution function for the Binomial is easily obtained. Included is a discussion of some other discrete distributions which can be generated without knowledge of $F_x$.

2.5 Poisson Distribution $\mathcal{P}(\lambda)$

If $x_1, \ldots, x_N$ is a sequence of $N$ independent exponentials with parameter $\lambda$, then a non-negative integer $k$ such that $S_k \leq 1$ and $S_{k+1} > 1$ is distributed Poisson with parameter $\lambda$, where

$$S_k = \sum_{i=1}^{k} x_i.$$  

2.6 Negative Binomial Distribution $\text{NB}(p, N)$

The negative binomial distribution can be generated
from a mixture of a Poisson and a Gamma distribution. That is, let $X$ be distributed as a Poisson with parameter $\theta$, where $\theta$ is a random variable from a Gamma distribution with parameters $\lambda, R$. Then $X$ is distributed as a negative binomial with parameters $p = \lambda/(1+\lambda)$ and $N=R$.

III. CORRELATED VARIATES

Continuous Distributions

3.1 Correlated Multivariate Normal Distribution CNORM ($u, \Sigma, A$)

Let $Z_0, Z_1, \ldots, Z_N$ be a sequence of $N+1$ $p$-dimensional independent multivariate normals with common null mean vector $\mathbf{0}$ and $pxp$ covariance matrix $\Sigma$. Then

$$X_i = a_i^2 Z_0 + (1-a_i^2) Z_i + u \quad \text{for } i=1,2,\ldots,N$$

are correlated multivariate normals with mean vector $u$ and dispersion matrix $A \otimes \Sigma$, where $\otimes$ denotes the Kronecker product of $A$ and $\Sigma$, that is

$$A \otimes \Sigma = \begin{bmatrix}
    a_{11} \Sigma & a_{12} \Sigma & \cdots & a_{1n} \Sigma \\
    a_{21} \Sigma & a_{22} \Sigma & \cdots & a_{2n} \Sigma \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{n1} \Sigma & \cdots & \cdots & a_{nn} \Sigma
\end{bmatrix}$$

$(np \times np)$

and $A$ is an $N \times N$ matrix where the $i,j^{th}$ element of $A$ is

$$a_{ij} = \begin{cases} 
    i=j & \text{if } i \neq j, i, j=1,2,\ldots,n \\
    1 & \text{if } i=j
\end{cases}$$

From the dispersion matrix $A \otimes \Sigma$ we have that
\[ \text{COV} (x_i, x_j) = \varepsilon_i \varepsilon_j \mathbb{I} \quad \text{if } i \neq j \]
\[ = \mathbb{I} \quad \text{if } i = j \]

Hence the correlation matrix between vector \( X_i, X_j \) is

\[ \text{CORR}(X_i, X_j) = \varepsilon_i \varepsilon_j \quad \text{if } i \neq j \]
\[ = \mathbb{I}_p \quad \text{if } i = j \]

where \( \mathbb{I}_p \) is a \( p \times p \) identity matrix. When \( p = 1 \) we have the univariate case.

\[ \text{3.2 Correlated Univariate Gamma Distribution } \Gamma(\lambda, R, A) \]

Let \( Z_0, Z_1, \ldots, Z_n \) denote a sequence of independent variables having the following Gamma distributions

\[ Z_0 \sim \Gamma(\lambda, R_0) \]
\[ Z_i \sim \Gamma(\lambda, R_i - R_0) \]

Let \( X_i = Z_0 + Z_i, i = 1, 2, \ldots, n \), then \( X_1, \ldots, X_n \) is a sequence of correlated Gamma variables where \( X_i \sim \Gamma(\lambda, R_i) \) and the correlation between \( X_i \) and \( X_j \) is

\[ \text{CORR}(X_i, X_j) = a_{ij} \]

where \( a_{ij} \) is the \( ij \)th element of the \( n \times n \) matrix \( A \) and

\[ a_{ij} = \begin{cases} 
1 & \text{if } i = j \\
\frac{R_i^2}{R_0^2} & \text{if } i \neq j
\end{cases} \]
3.3 Correlated Beta Distribution $B(p,q,A)$

Let $Z_0, Z_1, ..., Z_n$ be a sequence of independent chi-squares with degrees of freedom $df = v_i$ (Gamma with $\lambda = 1$, $R_i = v_i / 2$) for $i=0,1,2,...,n$. Let

$$X_i = Z_i / \left( \sum_{j=0}^{n} Z_j \right)$$

then the $X_i$'s are correlated Beta with parameter $(p_i, q_i)$ where $p_i = v_i / 2$ and $q_i = p - p_i$ where $p = \sum_{j=0}^{n} p_j$.

Then the correlation between $X_i$ and $X_j$ is given by

$$\text{CORR} (X_i, X_j) = a_{ij}$$

and

$$a_{ij} = \begin{cases} 1 & i=j \\ \frac{\sqrt{E_i P_j}}{(p-p_i)(p-p_j)} & i \neq j \end{cases}$$

Discrete Distributions

3.4 Correlated Poisson $P(\lambda, A)$

Let $Z_0, Z_1, ..., Z_n$ be a sequence of independent Poisson with parameters $\lambda_i, i=0,1,2,...,n$, then

$$X_i = Z_0 + Z_i$$

is a sequence of correlated Poissons with $X_i \sim P(\lambda_i)$ where $\lambda_i = \lambda_0 + \lambda_i$, $i=1,2,...,n$ and the correlation between $X_i$ and $X_j$ is given by

$$\text{Corr} (X_i, X_j) = a_{ij}$$
and

\[ a_{i,j} = \begin{cases} 
1 & i=j \\
\frac{C^2}{\lambda_i \lambda_j} & i \neq j 
\end{cases} \]

IV. CONCLUSIONS

The purpose of this paper is to document the procedure used in programming uncorrelated or correlated number generators for various specified distributions. The results are fairly well known and should prove to be satisfactory for most simulation needs. As mentioned in the introduction, the procedures are dependent upon the choice of the pseudo-random number generator selected, and hence the objective of the situation to be simulated may dictate changes in the random number generator. A simple package is presented which would hopefully satisfy the needs of those researchers interested in generating numbers from the statistical distributions given.

REFERENCES


## APPENDIX A

### JOB CONTROL PARAMETERS

<table>
<thead>
<tr>
<th>CARD</th>
<th>COL</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-5</td>
<td>NREPS - Number of sets of numbers to be generated (I5)</td>
</tr>
<tr>
<td>(215)</td>
<td>6-10</td>
<td>IX - Seed for random number generator. (I5) IX=0, then program will initiate using CPU clock</td>
</tr>
</tbody>
</table>

** Note the following set of cards are repeated NREPS times

<table>
<thead>
<tr>
<th>CARD</th>
<th>COL</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1-5</td>
<td>NOB - Number of observations to be generated (I5)</td>
</tr>
<tr>
<td>6-10</td>
<td>ITYPE - Type distribution to be generated (I5)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Normal</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Gamma</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>Beta</td>
<td>6</td>
</tr>
</tbody>
</table>

11 | ICOR = 1 | correlated data (II) |
| = 0 | uncorrelated data (II) |

12 | ISTAT = 1 | P, Q, t statistics (II) |
| = 0 | N, r |

12-13 | IUNIT = 0 | Generated data output on external device # IUNIT |
| # 0 | Generated data output on external device # IUNIT |

*** Note the following cards depend upon the distribution selected on Card # 2.

- NORMAL -

<table>
<thead>
<tr>
<th>CARD</th>
<th>COL</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1-5</td>
<td>NV = Number of variates (NV=2=bivariate normal (I5)</td>
</tr>
<tr>
<td>6-10</td>
<td>KEY = 0</td>
<td>Standardized normal mean = 0 variance = 1</td>
</tr>
<tr>
<td>KEY = 1</td>
<td>Read Mean, Variance (I5)</td>
<td></td>
</tr>
</tbody>
</table>
IF KEY = 1 Read following cards

<table>
<thead>
<tr>
<th>CARD</th>
<th>COL</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>(16F5.0)</td>
<td>Y(I),I=1,.NV Mean vector</td>
</tr>
<tr>
<td>5</td>
<td>(16F5.0)</td>
<td>S(I),I=1,.NV**2 Covariance matrix</td>
</tr>
</tbody>
</table>

** OF ICOR = 1 on card 2 read following for correlated case

6  Correlation factor (see page iii)
7  Means (same grouping as correlation factors) only need when NV=1

- GAMMA -

3  1-5  R1  Shape parameter (F5.0)
6-10 XLAMDA  Scale parameter (F5.0)

** IF ICOR = 1 Read following

4+  Correlation factor (page iii)

- BETA -

3  1-5  R1  Beta parameter (F5.0)
6-10 R2  Beta parameter (F5.0)

** IF ICOR = 1 Read following

4  1-5  VND  Parameter for Z_o (see page 9)(F5.0)
5+  V(I), same format as correlation factors (page iii)

- POISSON -

3  1-5  XLAMDA  Poisson Parameter (F5.0)

** IF ICOR = 1 Read following

4+  Correlation factors (page iii)
### NEGATIVE BINOMIAL

<table>
<thead>
<tr>
<th>CARD</th>
<th>COL</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1-5</td>
<td>P parameter (F5.0)</td>
</tr>
<tr>
<td>6-10</td>
<td></td>
<td>N parameter (I5)</td>
</tr>
</tbody>
</table>

**IF ICOR = 1 Read following

**Correlation factors (page iii)

### BINOMIAL

No additional inputs needed.

*** The following cards are used to define the A-matrix used in defining correlated observation

### CORRELATION FACTORS

1. (I3) | NG= Number of groups \[1 \leq NG \leq NOB\]
2. (16F5.0) | VALUE (I), I=1,NG, A value for each group

**Example 1**

\[\text{NOB}=25, \quad \text{NG}=1, \quad \text{NO}(1)=25\]

\[\text{VALUE}(1)=.8\]

\[\text{the \ CORR}(X_i, X_j) = (.8) \times (.8) = .64\]

**CARD**

1. \#1
2. \#10\#25
3. \#15\#.8

**Example 2**

\[\text{NOB}=25, \quad \text{NG}=2, \quad \text{NO}(1)=10, \quad \text{NO}(2)=15\]

\[\text{VALUE}(1)=.5, \quad \text{VALUE}(2)=.8\]

\[\text{then \ CORR}(X_i, X_j) = \begin{cases} .25 & i, j \leq 10 \\ .40 & i \leq 10, j > 10 \\ .40 & j \leq 10, i > 10 \\ .64 & i, j > 10 \end{cases}\]

**CARD**

1. \#2
2. \#10\#15
3. \#15\#.8

(Note: \# denote blank column)

iii
PROGRAM DESCRIPTION

MAIN - main program to read in job parameters

SUPER - supervisor routine to direct the generation of data, computation of statistics and printed output.

BETA - generates independent Beta variates.

GAMMA - Gamma variates.

BINOM - Binomial variates.

NORMAL - Normal variates.

POISSN - Poisson variates.

NEGBIN - Negative Binomial variates.

CBETA - correlated Beta variates.

CGAMMA - Gamma variates.

CNORML - Normal variates.

CPOISN - Poisson variates.

CNEGBN - Negative Binomial variates.

PRINT - prints generated values and output on specified unit.

STATS - calculates statistic for generated values.

RANDU - generates random uniform variates.
### Subroutines Needed by a Given Routine

<table>
<thead>
<tr>
<th>Subroutine</th>
<th>Required By</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNORML</td>
<td>RANDA</td>
</tr>
<tr>
<td></td>
<td>NORMAL</td>
</tr>
<tr>
<td>CNEGBN</td>
<td>GAMMA</td>
</tr>
<tr>
<td></td>
<td>POISSN</td>
</tr>
<tr>
<td>CBETA</td>
<td>GAMMA</td>
</tr>
<tr>
<td>CGAMMA</td>
<td>GAMMA</td>
</tr>
<tr>
<td>CPOISN</td>
<td>POISSN</td>
</tr>
<tr>
<td>NORMAL</td>
<td>RANDU</td>
</tr>
<tr>
<td>BETA</td>
<td>GAMMA</td>
</tr>
<tr>
<td>GAMMA</td>
<td>RANDU</td>
</tr>
<tr>
<td></td>
<td>NORMAL</td>
</tr>
<tr>
<td>BINOM</td>
<td>RANDU</td>
</tr>
<tr>
<td>GMETRC</td>
<td>RANDU</td>
</tr>
<tr>
<td>POISSN</td>
<td>GAMMA</td>
</tr>
<tr>
<td>NEGBIN</td>
<td>GMETRC</td>
</tr>
<tr>
<td>RANDA</td>
<td>RANDU</td>
</tr>
<tr>
<td>RANDU</td>
<td></td>
</tr>
<tr>
<td>STATS</td>
<td></td>
</tr>
<tr>
<td>PRINT</td>
<td></td>
</tr>
<tr>
<td>MAIN</td>
<td>SUPER</td>
</tr>
<tr>
<td>SUPER</td>
<td>CNORML</td>
</tr>
<tr>
<td></td>
<td>CBETA</td>
</tr>
<tr>
<td></td>
<td>CGAMMA</td>
</tr>
<tr>
<td></td>
<td>CNEGBN</td>
</tr>
<tr>
<td></td>
<td>CPOISN</td>
</tr>
<tr>
<td></td>
<td>NORMAL</td>
</tr>
<tr>
<td></td>
<td>GAMMA</td>
</tr>
<tr>
<td></td>
<td>BETA</td>
</tr>
<tr>
<td></td>
<td>NEGBIN</td>
</tr>
<tr>
<td></td>
<td>BINOM</td>
</tr>
<tr>
<td></td>
<td>POISSN</td>
</tr>
<tr>
<td></td>
<td>STATS</td>
</tr>
<tr>
<td></td>
<td>PRINT</td>
</tr>
</tbody>
</table>
SUBROUTINE DETA(X,N0)

DIMENSION X(100), Y(100), S(100), Z(100)
COMMON/R, Y, S, Z
COMMON/A, [I = N\*R]*X, LAMDA, R2, P, N
X=1.0

RETURN
END

SUBROUTINE GAMMA(X,N0)

DIMENSION X(N0), Y(100), S(100), Z(100)
COMMON/R, Y, S, Z
COMMON/A, [I = N\*R]*X, LAMDA, R2, P, N
X=1.0

RETURN
END

SUBROUTINE PIFORM(X,N0)

DIMENSION X(N0)
COMMON/R, [I = N\*R]*X, LAMDA, R2, P, N
DO 10 I=1,N0
IN=I-1

DO 2 K=1,IN
2 CALL RANDU(IX*IX*YFL)
1 X(I)=YFL
RETURN
END

ORIGINAL PAGE IS OF POOR QUALITY
SUBROUTINE NORMAL (X,NO,NV,Y,KEY)
INTEGER X,X(100),Y(100)
COMMON/X,NV,Y/ KEY
COMMON/KEY(X,NV+1)*X

* X=0

NV=NV+1
P=P+1

C GENERATE N UNINDEPENDENT UNIFORM (0,1)
DO 2 I=1,NV
TM=I*(1Y/10)@10+1
DO 2 K=1,TM
3 CALL FAMITE X(I,Y,I)
2 X(I)=YF

C TRANSFORM TO N UN NORMAL (0,1)
J=NV/2.
(10) I=1,J
AZ:=K T(-XLOG(X(2* I-1)));)
X(2* I-1)=4COS(P*OX(10))
4 X(2* I)=XESIN(12*OX(10))
IF (X,Y,F0.0) RETURN
WRITE (2,50) K
500 FORMAT (CHAR AA AND CORRELATION )
WRITE (2,51) Y(I)
510 FORMAT (10X+1IF0.3)
511 IF(NV.GT.I) GO TO 6
DO 5 T=1,NV
5 X(I)=S(I)*X(I)+Y(I)
RETURN
6 NV=NV
DO 11 J=1,NV
(10) I=1,J
11 Z(J-1)*I+I=0.
Z(I)=SORT(S(I))
(10) I=1,N
13 Z(I)=Z(I)/Z(1)
DO 14 I=2,N
T=I+1
SUM=0.
IF (I.-J) 14,15,17
15 M=J-1
(10) K=M-1
16 SUM=SUM+Z((K-1)*N+I)*I*
Z((J-1)*N+I)=SORT(S((J-1)*N+I)-SUM)
GO TO 19
17 M=J-1
(10) K=M-1
19 SUM=SUM+Z((K-1)*N+I)*Z((K-1)*N+J))
Z((J-1)*N+I)=(S((J-1)*N+I)-SUM)/Z((J-1)*N+J)

C CONTINUE IF
DO 7 I=1,NO
DO 7 K=1,NV
S(I)*N+K=Y(K)
(10) K=1
7 S(I)*N+K=S(I)*N+K+Z((J-1)*N+J)
* X(NV*J-(NV-J))
(10) I=1,NO
DO 8 J=1,NV
8 X((J-1)*N+I)=S((I-1)*N+J)
RETURN
100 FORMAT(1KF5.0)
END
```
SUBROUTINEamon (XX:*N)
COMMON JA/A [X:*N] P:XLAMDA:R2:F:N
XX:0.
* = 5
ONF= 1.
D=ONE=N
IF (1 == N) CALL PAMMU([I:*I]'I)
SIM=0.
J=0
G2=P
J=J+1
SUM=SUM+N
IF (SUM == 1) GO TO 3
1 XX:XX+J
RETURN
END

SUBROUTINE poisson (X:*N)
DIMENSION X (NO)
COMMON JA/A [X:*N] P:XLAMDA:R2:F:N
P:SUM=0.
J=1.
1 SUM=0.
J=0
2 CALL GAMMA (XX:1.
J=J+1
SUM=SUM*XY
IF (SUM = IF:1.) GO TO 2
X:SUM=1-
NSUM=SUM+1.
IF (NSUM = IF:NO) GO TO 1
RETURN
END

SUBROUTINE typf (X:*NO:*ITYPF)
DIMENSION X (NO)
COMMON JA/A [X:*N] P:XLAMDA:R2:F:N
GO TO (1:2:3:4:5:6:7:8:9:10)
1 IF X:*NO:KEY
NNO:NO X:KEY
CALL NORMAL (X:*NO:NN0:*KEY)
RETURN
2 IF X:*NO:1.1
CALL GAMMA (X:*NO)
RETURN
3 IF X:*NO:PI.*1.
CALL RITA (X:*NO)
RETURN
4 IF X:*NO:PI.*1.
CALL PAMMU ([I:*I]'I)
RETURN
5 CALL RIFIM (X:*NO)
RETURN
6 IF X:*NO:PI.*1.
CALL MGEMN (X:*NO)
RETURN
100 FORMAT (2*5)
101 FORMAT (2*5:0)
102 FORMAT (5*0:15)
END
```
SUBROUTINE A4GRIN (X, N0)
DIMENSION X(N0)
COMMON /A/ [X]*N0*P*2+P+N
RETURN
END

SUBROUTINE RANGH(IY*YFL)
IY=IY*A4SKA39
IF (IY)*A4A47)
YFL=YFL*A4SKA36E-9
RETURN
END

SUBROUTINE PRTN (X*NO*II*IT*II)
DIMENSION X(N0)
COMMON /X*/X*N0*1+N*N0*1+N
NOV=NVE+1
GO TO (1,2,3,4,5,6,7,8)
1 WRITE (6,100) II
WRITE (4,200) (X(I), I=1, NOV)
IF (II)*F0.2) WRITE (9,300) (X(I), I=1, NOV)
RETURN
2 WRITE (6,101) II*1+R1*XLAMDA
WRITE (4,200) X
IF (II)*F0.2) WRITE (9,300) X
RETURN
3 WRITE (6,102) II*1+R2
WRITE (4,200) X
IF (II)*F0.2) WRITE (9,300) X
RETURN
4 WRITE (6,103) II*1*XLAMDA
WRITE (4,200) X
IF (II)*F0.2) WRITE (9,300) X
RETURN
5 WRITE (6,104) II
WRITE (4,200) X
IF (II)*F0.2) WRITE (9,300) X
RETURN
6 WRITE (6,105) II*P*N
WRITE (4,200) X
IF (II)*F0.2) WRITE (9,300) X
RETURN
100 FORMAT (1, NORMA] DATA FOR REP.*16, N.*NO. OF VARIATES=16.
* //)
101 FORMAT (C, GAMMA DATA FOR REP.*16, N=1.*F10.3, //)
102 FORMAT (C, BETA DATA FOR REP.*16, ALPHAS=1.*F10.2, RETA=1. //)
103 FORMAT (C, POISSN DATA FOR REP.*16, LAMBDA=1.*F10.2, //)
104 FORMAT (C,UniFORM (0,1) DATA FOR INTEGRAL TRANSFORM REP.*16, //)
105 FORMAT (C,UNIFORM DATA FOR REP.*16, //)
200 FORMAT (10X, 10F10.3)
900 FORMAT (5DF10.3)
END