UNIVERSITY OF ARKANSAS

STATISTICAL ANALYSIS OF MULTIVARIATE ATMOSPHERIC VARIABLES

(NASA-CR-161472) STATISTICAL ANALYSIS OF MULTIVARIATE ATMOSPHERIC VARIABLES Final Report (Arkansas Univ.) 158 p HC A08/MF A01 CSCL 04B

N80-25991

Unclas

G3/47 23463

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

Department of Mathematics
University of Arkansas
Fayetteville, Arkansas

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

ESTIMATION IN DISCRETE MULTIVARIATE DISTRIBUTIONS

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 computor 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 occuring 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

$$P(n_1, n_2, \ldots, n_k) = N! \prod_{j=1}^{k} (P_j^{n_j} / n_j!) \quad (0 \le n_j;\ \sum_{j=1}^{k} n_j = N). \tag{1}$$

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

$$(Q - \sum_{i=1}^{k} P_i t_i)^{-N} \tag{2}$$

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

From formula (2) we have the following distribution function

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

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

$$P(x_1,\ldots,x_k) = \exp(-\xi-\theta_1-\ldots-\theta_k) \sum_{j=0}^{m} \left[\frac{\xi^j}{j!}\ \frac{\theta_1^{x_1-j}}{(x_1-j)!}\ \frac{\theta_2^{x_2-j}}{(x_2-j)!} \cdots \frac{\theta_k^{x_k-j}}{(x_i-j)!}\right] \qquad (4)$$

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

$$\hat{p}_j = n_j/N \qquad (j=1,\ldots,k) \tag{5}$$

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

$$E(n_j) = \left.\frac{\partial m(t_1,\ldots,t_k)}{\partial t_j}\right|_{\underline{t}=\underline{0}} = NP_j \text{ for } j=1,\ldots,k$$

$$E(n_i n_j) = \left. \frac{\partial^2 m(t_1, \ldots, t_k)}{\partial t_j\, t_i} \right|_{\underline{t}=\underline{0}}$$

$$= N(N+1)P_i P_j$$

$$= N^2 P_i P_j + N P_i P_j$$

$$= E(n_i)\, E(n_j) + \frac{E(n_i)E(n_j)}{N}$$

giving

$$N = \frac{E(n_i)E(n_j)}{E(n_i n_j) - E(n_i)E(n_j)} \tag{6}$$

and

$$P_j = E(n_j)/N. \tag{7}$$

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

$$\hat{N} = \frac{\bar{n}_i \bar{n}_j}{\overline{n_i n_j} - \bar{n}_i \bar{n}_j}$$ and $\hat{P}_j = \bar{n}_j/\hat{N}$ for i, $j=1,\ldots,k$ and $i \neq j$ where

$$\bar{n}_\ell = \frac{\sum_{i=1}^{n} n_{\ell_i}}{n} \; ; \; \overline{n_\ell n_k} = \frac{\sum_{i=1}^{n} n_{\ell_i} n_{h_i}}{n}$$ given n 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 $\overline{n_i n_j}$. The program first averages the $k(k-1)/2$ values of $\bar{n}_i \bar{n}_j$ and $\overline{n_i n_j}$ 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 estimated 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

$$m(t_1,\ldots,t_k) = \exp\left[-\xi\left(1-\exp\left(\sum_{i=1}^{k} t_i\right)\right) - \sum_{i=1}^{k} \theta_i\left(1-e^{t_i}\right)\right] \quad (8)$$

It follows that

$$\left.\frac{\partial m(t_1,\ldots,t_k)}{\partial t_i}\right|_{\underline{t}=\underline{0}} = E(x_i) = \xi + \theta_i \quad (9)$$

$$\left.\frac{\partial^2 m(t_1,\ldots,t_k)}{\partial t_i \partial t_j}\right|_{\underline{t}=\underline{0}} = E(x_i x_j)$$

$$= (\xi+\theta_i)(\xi+\theta_j) + \xi$$

$$= E(x_i)\, E(x_j) + \xi$$

Therefore

$$\xi = E\left[x_i x_j\right] - E\left[x_i\right] E\left[x_j\right]. \quad (10)$$

Substituting raw estimates for expected values we have

$$\hat{\xi} = \overline{x_i x_j} - \bar{x}_i \bar{x}_j \quad \text{where } \bar{x}_\ell = \frac{\sum_{i=1}^{n} x_{\ell i}}{n} \; ; \; \overline{x_1 x_k} = \frac{\sum_{i=1}^{n} x_{\ell i} x_{k i}}{n} \,. \quad (11)$$

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

3.2 Variances of Parameter Estimates

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

$$\begin{aligned} \operatorname{var}(\hat{P}_j) &= \operatorname{var}(n_j/N). \\ &= E(n_j/N)^2 - \{E(n_j/N)\}^2 \\ &= \frac{1}{N^2}(N^2P_j^2 + NP_jq_j) - P_j^2 \\ &= \frac{P_jq_j}{N} = \frac{P_j(1-P_j)}{N} , \end{aligned} \tag{12}$$

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

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

$$V(a_1,a_2,\ldots,a_k) = \left(E\left[-\frac{\partial^2 \log L}{\partial a_i \partial a_j}\right]\right)^{-1} \tag{13}$$

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

$$L = \frac{\prod_{j=1}^{n} \Gamma(N+\sum_{i=1}^{k} n_{ij})}{\prod_{j=1}^{n}\prod_{i=1}^{k} n_{ij}!)(\Gamma(N))^n} Q^{-Nn} \prod_{i=1}^{k} (P_i/Q)^{\sum_{j=1}^{n} n_{ij}} \tag{14}$$

$$\ln L = \sum_{j=1}^{n} \ln \Gamma(N+S_j) - \ln (\prod_{j=1}^{n} \prod_{i=1}^{k} n_{ij}!) - n \ln\Gamma(N) \tag{15}$$

$$-Nn \ln Q + \sum_{i=1}^{k} (\sum_{S_j^1}^{n} n_{ij})(\ln P_i - \ln Q)$$

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

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

$$\frac{\partial^2 \ln L}{\partial^2 N} = \sum_{j=1}^{n} \sum_{k=0}^{E(S_j)-1} \frac{-1}{(N+k)^2} = \sum_{k=1}^{\infty} - (\hat{N}+k-1)^{-2} E(F_j) \tag{17}$$

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

$$\frac{\partial^2 \ln L}{\partial N \partial P_i} = \frac{-n}{1+\sum_{i=1}^{k} \hat{P}_i} \quad \text{for } i=1,\dots,k \tag{18}$$

$$\frac{\partial \ln L}{\partial P_i} = \frac{-\hat{N}n - \sum_{\ell=1}^{k} F(S_\ell^1)}{1 + \sum_{j=1}^{k} \hat{P}_j} + \frac{E(S_i^1)}{P_i} \tag{19}$$

$$\frac{\partial^2 \ln L}{\partial^2 P_j} = \frac{Nn + \sum_{\ell=1}^{k} E(S_\ell^1)}{(1 + \sum_{\ell=1}^{k} \hat{P}_\ell)^2} - \frac{E(S_j^1)}{P_j^2} \tag{20}$$

$$\frac{\partial^2 \ln L}{\partial P_j \partial P_i} = \frac{\hat{N}n + \sum_{\ell=1}^{k} E(S_\ell^{1})}{(1 + \sum_{\ell=1}^{k} \hat{P}_\ell)^2} \quad \text{for } i \neq j \tag{21}$$

Using the two sets of estimates for N, $P_1,\ldots,P_k$ and numerical values for $E(S_j)$, $E(S_j^{1})$, and $E(F_j)$, we can obtain approximate bounds for $V(\hat{N},\hat{P}_1,\ldots,\hat{P}_k)$.

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 internal about the variance of the parameter estimates.

TABLE 1.

THE MOMENT ESTIMATE OF N OBTAINED BY AVERAGING THW RAW MOMENTS FIRST IS	3.350
THE CORRESPONDING PROBABILITIES ASSOCIATED WITH THE RESPECTIVE VARIABLES ARE	0.295 0.385
THE MOMENT ESTIMATE OF N OBTAINED BY AVERAGING ALL POSSIBLE MOMENT ESTIMATES	3.350
THE CORRESPONDING PROBABILITIES ASSOCIATED WITH THE RESPECTIVE VARIABLES ARE	0.295 0.385

FISHER'S INFORMATION MATRIX USING THE MINIMUM ESTIMATE OF N

1.207		
-0.108	0.010	
-0.140	0.012	0.017

REFERENCES

Arbous, A.G. and Kerrich, J.E. (1951). Accident Statistics and the Concept of Accident Proneness, Biometrics 7, pp. 340-432.

Bates, Grace E. and Neyman, J. (1952). Contributions to the Theory of Accident Proneness, University of California, Publications in Statistics, 1, pp. 215-253.

IBM Application Program (1968). System/360 Scientific Subroutine Package, Fourth edition.

Johnson, N. and Kotz, S. (1969). Distributions in Statistics: Discrete Distributions. Boston: Houghton-Mifflin.

Kendall, Maurice G. and Stuart, Alan (1961). The Advanced Theory of Statistics: Inference and Relationship, 2, p. 28.

A P P E N D I X

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 multinomial 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*8 (A-H,O-Z)
      REAL*8 E(12),EE(11,11),EP(11,11),NJ(350,12),P(12),AN(11,11),EM(11,
     C11),T(12),F(350),INF(13,13),S(350),SP(12),RM(91)
   28 READ(5,1,END=100) IH
      IF (IH.EQ.2) GO TO 30
      READ(5,1) K,M,((NJ(I,J),J=1,K),I=1,M)
    1 FORMAT(2I3,(7F10.0))
      K1=K+1
      KK=K-1
      DO 9 I=1,K
      E(I)=0D0
    9 SP(I)=0D0
      DO 10 I=1,KK
      DO 10 J=1,KK
   10 EP(I,J)=0D0
      DO 12 I=1,M
   12 S(I)=0D0
      DO 14 I=1,K1
      DO 14 J=1,K1
   14 INF(I,J)=0D0
      DO 2 I=1,K
      DO 3 J=1,M
    3 E(I)=E(I)+NJ(J,I)
    2 E(I)=E(I)/M
      DO 4 J=1,KK
      JJ=J+1
      DO 4 L=JJ,K
      LL=L-1
      DO 5 I=1,M
    5 EP(J,LL)=EP(J,LL)+NJ(I,J)*NJ(I,L)
      EE(J,LL)=E(J)*E(L)
    4 EP(J,LL)=EP(J,LL)/M
      S1=0D0
      S2=0D0
      DO 6 I=1,KK
      DO 6 J=I,KK
      S1=S1+EE(I,J)
    6 S2=S2+EP(I,J)
      G=K*(K-1)/2D0
      IF (IH.EQ.1) GO TO 16
C     CALCULATION OF THE NEG. MULTINOMIAL PARAMETERS BEGINS HERE
      HN=S1/(S2-S1)
      IF (HN.LE.0D0) GO TO 60
   11 WRITE(6,8) HN
      WRITE (6,26)
      DO 7 I=1,K
      P(I)=E(I)/HN
    7 WRITE(6,27) P(I)
      SUM=0D0
      DO 13 I=1,KK
      DO 13 J=I,KK
      AN(I,J)=EE(I,J)/(EP(I,J)-EE(I,J))
      IF (AN(I,J).LE.0D0) GO TO 60
   13 SUM=SUM+AN(I,J)
      SUM=SUM/G
      WRITE (6,24) SUM
      WRITE (6,26)
      DO 15 I=1,K
      P(I)=E(I)/SUM
   15 WRITE(6,27) P(I)
      D=AN(1,1)
```

```
      AE=AN(1,1)
      DO 31 I=1,KK
      DO 31 J=I,KK
      IF (AN(I,J)-D) 33,33,32
C   N>0 IN THE NEGATIVE MULTINOMIAL DISTRIBUTION
   32 D=AN(I,J)
   33 IF (AN(I,J)-AE) 34,31,31
   34 AE=AN(I,J)
   31 CONTINUE
C   D IS NOW THE MAX ESTIMATE OF N AND AE IS THE MIN
      DO 62 I=1,K
      DO 62 J=1,M
   62 SP(I)=SP(I)+NJ(J,I)
      DO 63 I=1,M
      DO 63 J=1,K
   63 S(I)=S(I)+NJ(I,J)
      MM=M-2
      DO 64 I=1,MM,2
      II=I+2
      I1=I+1
      IF (S(I+1)-S(I)) 75,75,76
   75 DD=S(I)
      S(I)=S(I+1)
      S(I+1)=DD
   76 DO 64 J=II,M
      IF (S(J)-S(I)) 65,65,64
   65 DD=S(I)
      CE=S(I1)
      S(I)=S(J)
      S(I1)=DD
      S(J)=CE
   64 CONTINUE
      L=S(M)
      DO 66 J=1,L
      DO 67 I=1,M
      IF (S(I)-J) 67,68,68
   68 F(J)=M-I+1D0
      GO TO 66
   67 CONTINUE
   66 CONTINUE
      SH=0D0
      DO 78 I=1,M
   78 SH=SH+S(I)
   39 DO 35 I=1,L
   35 INF(1,1)=INF(1,1)+(1D0/(D+I-1D0)**2)*F(I)
      SPI=0D0
      DO 36 I=1,K
      P(I)=E(I)/D
   36 SPI=SPI+P(I)
      DO 37 I=2,K1
      H=M
   37 INF(1,I)=H/(1D0+SPI)
      DO 38 J=2,K1
      DO 38 I=J,K1
      INF(J,I)=-(D*M+SH)/(1D0+SPI)**2
   38 IF (I.EQ.J) INF(I,J)=INF(I,J)+SP(J-1)/P(J-1)**2
      IF (D.EQ.AE) WRITE (6,44)
      IF (D.NE.AE) WRITE (6,43)
      CALL ARRAY (2,K1,RM,INF)
      CALL SINV (RM,K1,.05,IER)
      CALL ARRAY(1,K1,RM,INF)
```

```
      DO 23 I=1,K
      I1=I+1
      DO 23 J=I1,K1
   23 INF(J,I)=INF(I,J)
      DO 41 J=1,K1
   41 WRITE (6,42) (INF(I,J),I=1,J)
   42 FORMAT(9F14.8/4F14.8)
   43 FORMAT('CFISHERS INFORMATION MATRIX USING THE MAXIMUM ESTIMATE OF
     CN')
      IF (D.EQ.AE) GO TO 28
      D=AE
      GO TO 39
   44 FORMAT('CFISHER''S INFORMATION MATRIX USING THE MINIMUM ESTIMATE
     COF N')
C     CALCULATIONS OF MULTIVARIATE POISSON PARAMETERS BEGIN HERE
   16 PE=(S2-S1)/G
      IF (PE.LE.ODO) GO TO 80
      SUM=ODO
      WRITE (6,18) PE
      WRITE (6,21)
      DO 20 I=1,K
      T(I)=E(I)-PE
   20 WRITE(6,27) T(I)
      DO 17 I=1,KK
      DO 17 J=I,KK
      EH(I,J)=EP(I,J)-EE(I,J)
      IF (EH(I,J).LE.ODO) GO TO 80
   17 SUM=SUM+EH(I,J)
      SUM=SUM/G
      WRITE (6,19) SUM
      WRITE (6,21)
      DO 22 I=1,K
      T(I)=E(I)-SUM
   22 WRITE (6,27) T(I)
   27 FORMAT(31X,F14.8)
   19 FORMAT('0','THE MOMENT ESTIMATE OF THE POISSON PARAMETER FOR U OBT
     CAINED BY AVERAGING ALL'/' POSSIBLE MOMENT ESTIMATES IS',3X,F14.8)
    8 FORMAT('0','THE MOMENT ESTIMATE OF N OBTAINED BY AVERAGING THE RAW
     C MOMENTS FIRST IS'/31X,F14.8)
   26 FORMAT(' THE CORRESPONDING PROBABILITIES ASSOCIATED WITH THE RESPE
     CCTIVE VARIABLES ARE')
   24 FORMAT('0','THE MOMENT ESTIMATE OF N OBTAINED BY AVERAGING ALL POS
     CSIBLE MOMENT ESTIMATES'/' OF N IS',24X,F14.8)
   18 FORMAT('0','THE MOMENT ESTIMATE OF THE POISSON PARAMETER FOR U OBT
     CAINED BY AVERAGING THE RAW MOMENTS FIRST IS',F14.8)
   21 FORMAT (' THE CORRESPONDING ESTIMATE OF THE POISSON PARAMETER FOR
     C THE RESPECTIVE V VARIABLES ARE')
      GO TO 28
C     CALCULATION OF MULTINOMIAL PARAMETERS BEGIN HERE
   30 READ (5,49) K,(T(I),I=1,K)
   49 FORMAT(I3,(7F10.0))
      WRITE (6,52)
      SUM=ODO
      DO 50 I=1,K
   50 SUM=SUM+T(I)
      DO 51 I=1,K
      P(I)=T(I)/SUM
      SP(I)=P(I)*(1D0-P(I))/SUM
   51 WRITE(6,48) P(I),SP(I)
   52 FORMAT('CPROBABILITIES',10X,'APPROXIMATE VARIANCES')
   48 FORMAT (F14.8,14X,F14.8)
      GO TO 28
   80 WRITE (6,81)
   81 FORMAT(' A NON-NEGATIVE PARAMETER HAS BEEN ESTIMATED AS NEGATIVE')
  100 STOP
      END
```

```
      SUBROUTINE ARRAY(MODE,N,RM,INF)
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL*8 RM(91),INF(13,13)
      IF (MODE-1) 100,100,120
  100 IJ=0
      DO 110 K=1,N
      DO 110 L=1,K
      IJ=IJ+1
  110 INF(L,K)=RM(IJ)
      GO TO 140
  120 IJ=0
      DO 125 K=1,N
      DO 125 L=1,K
      IJ=IJ+1
  125 RM(IJ)=INF(L,K)
  140 RETURN
      END


      SUBROUTINE MFSD(A,N,EPS,IER)
      IMPLICIT REAL*8 (A-H,O-Z)
      DIMENSION A(91)
      IF (N-1) 12,1,1
    1 IER=0
      KPIV=0
      DO 11 K=1,N
      KPIV=KPIV+K
      IND=KPIV
      LEND=K-1
      TOL=DABS(EPS*A(KPIV))
      DO 11 I=K,N
      DSUM=0.D0
      IF (LEND) 2,4,2
    2 DO 3 L=1,LEND
      LANF=KPIV-L
      LIND=IND-L
    3 DSUM=DSUM+A(LANF)*A(LIND)
    4 DSUM=A(IND)-DSUM
      IF(I-K) 10,5,10
    5 IF (DSUM-TOL) 6,6,9
    6 IF (DSUM) 12,12,7
    7 IF (IER) 8,8,9
    8 IER=K-1
    9 DPIV=DSQRT(DSUM)
      A(KPIV)=DPIV
      DPIV=1.D0/DPIV
      GO TO 11
   10 A(IND)=DSUM*DPIV
   11 IND=IND+I
      RETURN
   12 IER=-1
      RETURN
      END
```

```
      SUBROUTINE SINV(A,N,EPS,IER)
      IMPLICIT REAL*8 (A-H,O-Z)
      DIMENSION A(91)
      CALL MFSD(A,N,EPS,IER)
      IF (IER) 9,1,1
    1 IPIV=N*(N+1)/2
      IND=IPIV
      DO 6 I=1,N
      DIN=1D0/A(IPIV)
      A(IPIV)=DIN
      MIN=N
      KEND=I-1
      LANF=N-KEND
      IF (KEND) 5,5,2
    2 J=IND
      DO 4 K=1,KEND
      WORK=0D0
      MIN=MIN-1
      LHOR=IPIV
      LVER=J
      DO 3 L=LANF,MIN
      LVER=LVER+1
      LHOR=LHOR+L
    3 WORK=WORK+A(LVER)*A(LHOR)
      A(J)=-WORK*DIN
    4 J=J-MIN
    5 IPIV=IPIV-MIN
    6 IND=IND-1
      DO 8 I=1,N
      IPIV=IPIV+I
      J=IPIV
      DO 8 K=I,N
      WORK=0D0
      LHOR=J
      DO 7 L=K,N
      LVER=LHOR+K-I
      WORK=WORK+A(LHOR)*A(LVER)
    7 LHOR=LHOR+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 probablities 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 satisfacfory. 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 Dirchlet distribution would be a natural extension for describing the bivariate case. However, a theoretical requirement of the Dirchet 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 $\underline{x}=(x_1,x_2)$ denote an arbitrary two dimensional vector and $f(\underline{x})$ a real valued function of $\underline{x}$. Consider the integral

$$\theta = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\underline{x})g(\underline{x})dx_1dx_2 \tag{2.1}$$

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

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} f(\underline{x}_i)$$

where $\underline{x}_i$, $i=1,\ldots,N$ are random samples from the pdf $g(\underline{x})$. The variance of $\hat{\theta}$, is given by

$$\mathrm{var}\ (\hat{\theta}) = \frac{1}{N}\ \mathrm{var}\ (f(\underline{x})) = \frac{1}{N} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (f(\underline{x})-\theta)^2 g(\underline{x})\ dx_1 dx_2$$

which can be estimated by

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (f(\underline{x}_i) - \hat{\theta})^2.$$

The estimated standard error is given by, $\hat{e} = s/\sqrt{n}$.

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

$$\chi = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\underline{x})\ g(\underline{x}) dx_1 dx_2$$

is known. Then

$$\theta = \chi + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (f(\underline{x}) - h(\underline{x})) g(\underline{x}) dx_1 dx_2.$$

The variance of $f(\underline{x}) - h(\underline{x})$, is given by

$$\mathrm{var}\ (f(\underline{x})-h(\underline{x})\ ) = \mathrm{var}\ (f(\underline{x}))+\mathrm{var}(h(\underline{x}))-2\ \mathrm{cov}(f(\underline{x}),h(\underline{x})).$$

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

$$\mathrm{var}\ (f(\underline{x}) - h(\underline{x})) < \mathrm{var}\ (f(\underline{x})).$$

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

$$\begin{aligned} \mathrm{var}(f) &= \mathrm{var}[\ h+f-h\ ] \\ &= \mathrm{var}(h) + \mathrm{var}(f-h) + 2\ \mathrm{cov}(h,f-h) \end{aligned}$$

Thus we have

$$\mathrm{var}(f-h) = \mathrm{var}(f) - \mathrm{var}(h) - 2\ \mathrm{cov}\ (h,f-h).$$

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

$$\theta = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(\underline{x}|\underline{\mu}, \Sigma) dx_1 dx_2.$$

where $\underline{\mu}' = (\mu_1, \mu_2)$; $\Sigma = \begin{pmatrix} \sigma_1 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2 \end{pmatrix}$ and

$f(\underline{x}|\ \underline{\mu}, \Sigma\ )$ = BVN distribution =

$$\frac{1}{2\pi\sigma_1\sigma_2\ (1-\rho^2)^{1/2}} \exp\{- \frac{1}{2(1-\rho^2)} [(\frac{x_1-\mu_1}{\sigma_1})^2 - 2\rho \frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\ \sigma_2} +$$

$$(\frac{x_2-\mu_2}{\sigma_2})^2]\ \} \qquad (2.2)$$

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

$$\phi = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(\underline{x}|\ \underline{\mu}, \Sigma) \frac{dx_1 dx_2}{(b_1-a_1)(b_2-a_2)}. \qquad (2.3)$$

It follows that

$$\theta = \phi(b_1-a_1)(b_2-a_2)$$

and the estimate

$$\hat{\theta} = \hat{\phi}\,(b_1-a_1)(b_2-a_2)$$

where

$$\hat{\phi} = \frac{1}{N}\sum_{i=1}^{N} f(\underline{x}_i \mid \underline{\mu}, \Sigma)$$

when $\underline{x}_i$ is a random vector from the pdf

$$g(\underline{x}) = \begin{cases} \dfrac{1}{(b_1-a_1)(b_2-a_2)} & ; \; a_j \le x_j \le b_j \;, \quad j=1,2 \\ 0 & ; \; \text{Otherwise} \end{cases}$$

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 BVN program the regular part $h(x)$ is defined to be all the terms up to the coeffecient 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

$$\begin{aligned} f(x_1,x_2) &= f(a_1,a_2) + (x_1-a_1)\frac{\partial f}{\partial x_1}(a_1,a_2) \\ &+ (x_2-a_2)\frac{\partial f}{\partial x_2}(a_1,a_2) + \frac{1}{2!}\Big[(x_1-a_1)^2\frac{\partial^2 f}{\partial x_1^2}(a_1,a_2) \\ &+ 2(x_1-a_1)(x_2-a_2)\frac{\partial^2 f}{\partial x_1\partial x_2}(a_1,a_2) + (x_2-a_2)\frac{\partial^2 f}{\partial x_2^2}(a_1,a_2)\Big] \\ &+ \frac{1}{3!}\Big[(x_1-a_1)^3\frac{\partial^3 f}{\partial x_1^3}(a_1,a_2) + 3(x_1-a_1)^2(x_2-a_2)\frac{\partial^3 f}{\partial x_1^2\partial x_2}(a_1,a_2) \\ &+ 3(x_1-a_1)(x_2-a_2)^2\frac{\partial^3 f}{\partial x_1\partial x_2^2}(a_1,a_2) + (x_2-a_2)^3\frac{\partial^3 f}{\partial x_2^3}(a_1,a_2)\Big] + \cdots \end{aligned} \tag{2.4}$$

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

$$\frac{\partial f}{\partial x_1}(\mu_1,\mu_2)=f(x_1,x_2)\left[\frac{1}{-2(1-\rho^2)}\left\{\frac{2}{\sigma_1^2}(x_1-\mu_1) - \frac{2\rho}{\sigma_1\sigma_2}(x_2-\mu_2)\right\}\right]\Bigg|_{\substack{x_1=\mu_1\\ x_2=\mu_2}} = 0$$

$$\frac{\partial^2 f}{\partial x_1^2}(\mu_1,\mu_2) = \frac{-1}{2\pi\sigma_1^3\sigma_2(1-\rho^2)^{3/2}}$$

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

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

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

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

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

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

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

$$\frac{\partial^6 f}{\partial x_1^6} = \frac{-15}{2\pi\sigma_1^7\sigma_2(1-\rho^2)^{7/2}}$$

$$\frac{\partial^6 f}{\partial x_1^5 \partial x_2} = \frac{15\rho}{2\pi\sigma_1^6\sigma_2^2(1-\rho^2)^{7/2}}$$

$$\frac{\partial^6 f}{\partial x_1^4 \partial x_2^2} = \frac{-3-12\rho^2}{2\pi\sigma_1^5\sigma_2^3(1-\rho^2)^{7/2}}$$

$$\frac{\partial^6 f}{\partial x_1^3 \partial x_2^3} = \frac{9\rho+6\rho^3}{2\pi\sigma_1^4\sigma_2^4(1-\rho^2)^{7/2}}$$

$$\frac{\partial^6 f}{\partial x_1^2 \partial x_2^4} = \frac{-3-12\rho^2}{2\pi\sigma_1^3\sigma_2^5(1-\rho^2)^{7/2}}$$

$$\frac{\partial^6 f}{\partial x_1 \partial x_2^5} = \frac{15\rho}{2\pi\sigma_1^2\sigma_2^6(1-\rho^2)^{7/2}}$$

$$\frac{\partial^6 f}{\partial x_2^{5}} = \frac{-15}{2\pi\sigma_1\sigma_2^{7}(1-\rho^2)^{7/2}}$$

$$\frac{\partial^8 f}{\partial x_1^{8}} = \frac{105}{2\pi\sigma_1^{9}\sigma_2\,(1-\rho^2)^{9/2}}$$

$$\frac{\partial^8 f}{\partial x_1^{7}\partial x_2} = \frac{-105\,\rho}{2\pi\sigma_1^{8}\sigma_2^{2}(1-\rho^2)^{9/2}}$$

$$\frac{\partial^8 f}{\partial x_1^{6}\partial x_2^{2}} = \frac{15+90\rho^2}{2\pi\sigma_1^{7}\sigma_2^{3}(1-\rho^2)^{9/2}}$$

$$\frac{\partial^8 f}{\partial x_1^{5}\partial x_2^{3}} = \frac{-45\rho-60\rho^3}{2\pi\sigma_1^{6}\sigma_2^{4}(1-\rho^2)^{9/2}}$$

$$\frac{\partial^8 f}{\partial x_1^{4}\partial x_2^{4}} = \frac{72\rho^2+9+24\rho^4}{2\pi\sigma_1^{5}\sigma_2^{5}(1-\rho^2)^{9/2}}$$

$$\frac{\partial^8 f}{\partial x_1^{3}\partial x_2^{5}} = \frac{-45\,\rho-60\rho^3}{2\pi\sigma_1^{4}\sigma_2^{6}(1-\rho^2)^{9/2}}$$

$$\frac{\partial^8 f}{\partial x_1^{2}\partial x_2^{6}} = \frac{15+90\rho^2}{2\pi\sigma_1^{3}\sigma_2^{7}\,(1-\rho^2)^{9/2}}$$

$$\frac{\partial^8 f}{\partial x_1\partial x_2^{7}} = \frac{-105\,\rho}{2\pi\sigma_1^{2}\sigma_2^{8}(1-\rho^2)^{9/2}}$$

$$\frac{\partial^8 f}{\partial x_2^8} = \frac{105}{2\pi\sigma_1\sigma_2^9(1-\rho^2)^{9/2}}$$

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

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

$$+(x_1-\mu_1)(x_2-\mu_2)\frac{\rho}{2\pi\sigma_1^2\sigma_2^2(1-\rho^2)^{3/2}} - \frac{1}{2}(x_2-\mu_2)^2 \frac{1}{2\pi\sigma_1\sigma_2^3(1-\rho^2)^{3/2}}$$

$$+ \frac{1}{24}(x_1-\mu_1)^4 \frac{3}{2\pi\sigma_1^5\sigma_2(1-\rho^2)^{5/2}}$$

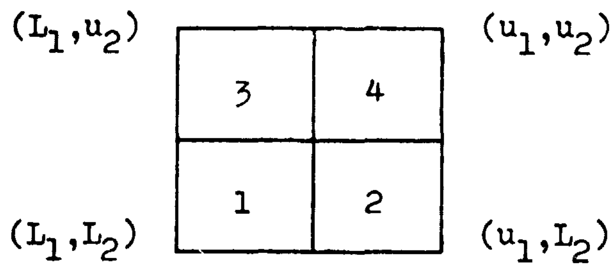
$$- \frac{1}{6}(x_1-\mu_1)^3(x_2-\mu_2)\frac{3\rho}{2\pi\sigma_1^4\sigma_2^2(1-\rho^2)^{5/2}} + - \cdots \qquad (2.5)$$

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 (μ_1,μ_2), where $h(x_1,x_2)$ are the first 25 terms in (2.5) and $\|\cdot\|$ is some distance function. For this reason

$$\int_A f(\underline{x}) - h(\underline{x})\; g(\underline{x})\; d\underline{x}$$

may not be bounded, especially for large region A. However, if the regular part $h(\underline{x})$ 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 $L_1 \le x_1 \le u_1$ and $L_2 \le x_2 \le u_2$ be the integration region. When divided into the four desired regions this becomes



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

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

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

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

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 $\mathrm{var}(h) < 2\ \mathrm{cov}(f,h)$, 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, (μ_1, μ_2), the standard deviations, σ_1, σ_2, the correlation ρ, the maximum standard error, starting value for random number generation (odd integer I5), and the limits of integration. The estimates for each of the four regions are outputted 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. $$\int_0^\infty \int_0^\infty f(\underline{x} \mid \underline{o}, \Sigma)dx_1dx_2$$

   where $\Sigma = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$

2. $$\int_{1/2}^\infty \int_1^\infty f(\underline{x} \mid \underline{o}, \Sigma)dx_1dx_2$$

   where $\Sigma = \begin{bmatrix} 1 & -.5 \\ -.5 & 1 \end{bmatrix}$

3. $$\int_0^{\infty} \int_0^{\infty} f(\underline{x} \mid \underline{o}, \Sigma)dx_1dx_2$$

where $$\Sigma = \begin{bmatrix} 1 & -.75 \\ -.75 & 1 \end{bmatrix}$$

4. $$\int_{1/2}^{\infty} \int_1^{\infty} f(\underline{x} \mid \underline{o}, \Sigma)dx_1dx_2$$

where $$\Sigma = \begin{bmatrix} 1 & .75 \\ .75 & 1 \end{bmatrix}$$

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 0.0025139714
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 0.0040517887 WITH STANDARD ERROR OF 0.0002146261
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 0.0000000995 WITH STANDARD ERROR OF 0.0000000122
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.

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 0.0 0.0

AN APPROXIMATION FOR THE 1 REGION

THE VALUE IS 0.1118712551 WITH STANDARD ERROR OF 0.0028780424
THE REGULAR PART IS NOT REMOVED

AN APPROXIMATION FOR THE 2 REGION

THE VALUE IS 0.0001379567 WITH STANDARD ERROR OF 0.0000210493
THE REGULAR PART IS NOT REMOVED

AN APPROXIMATION FOR THE 3 REGION

THE VALUE IS 0.0001607447 WITH STANDARD ERROR OF 0.0000219862
THE REGULAR PART IS NOT REMOVED

AN APPROXIMATION FOR THE 4 REGION

THE VALUE IS 0.0000000005 WITH STANDARD ERROR OF 0.0000000001
THE REGULAR PART IS NOT REMOVED

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

$$h(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\; x^{\alpha-1}(1-x)^{\beta-1};\quad 0\le x\le 1,\alpha,\quad \beta > 0. \qquad (3.1)$$

To approximate the probability that $0\le x\le x_o$, i.e.

$$\Pr\{x\le x_o\} = \int_o^{x_o} h(x:\alpha,\beta)dx$$

calculate the quantities

$$d_1 = (\alpha+\beta - 2/3)\,x_o - (\alpha - 1/3)$$

$$d_2 = d_1 + .02\left(\frac{x_o}{\beta} - \frac{1-x_o}{\alpha} + \frac{x_o-.5}{\alpha+\beta}\right),$$

and

$$z = \frac{d_z}{|\beta-.5-(\alpha+\beta-1)(1-x_o)|}\left\{\frac{12(\alpha+\beta-1)}{6(\alpha+\beta-1)-1}\left[(\beta-.5)\text{Log}\frac{\beta-.5}{[\alpha+\beta-1][1-x_o]} + (\alpha-.5)\text{Log}\frac{\alpha-.5}{[\alpha+\beta-1]\,x_o}\right]\right\}^{1/2} \qquad (3.2)$$

The approximate probability is given by

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

Of course, should you desire to have a right tail probability. The approximate value for the right tail probability is

$$P = \int_z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy.$$

The error in these approximations is less than .01 if $\alpha, \beta > 1$ and less than .001 if $\alpha, \beta > 2$. It also follows that $P_r \{x_o \leq x \leq x_1\}$ can be approximated as

$$P_r \{x_o \leq x \leq x_1\} = \int_{x_o}^{x} h(x;\alpha,\beta)dx = 1 - \int_0^{x_o} h(x;\alpha,\beta)dx - \int_{x_1}^{1} h(x;\alpha,\beta)dx$$

or

$$1 - \int_{-\infty}^{z_o} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy - \int_{z_1}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy = \int_{z_o}^{z_1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy.$$

However, the error is potentially doubled for this case.

The approximation is not valid for $\alpha, \beta \leq .5$ 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, $x_o \leq x \leq x_1$, $y_o \leq y \leq y_1$ can be approximated by

$$\int_{z_{x_o}}^{z_{x_1}} \int_{z_{y_o}}^{z_{y_1}} f(z_x,z_y)dz_x z_y . \qquad (3.3)$$

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

Category	Tenths
1	0
2	1,2,3
3	4,5
4	6,7,8,9
5	10

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

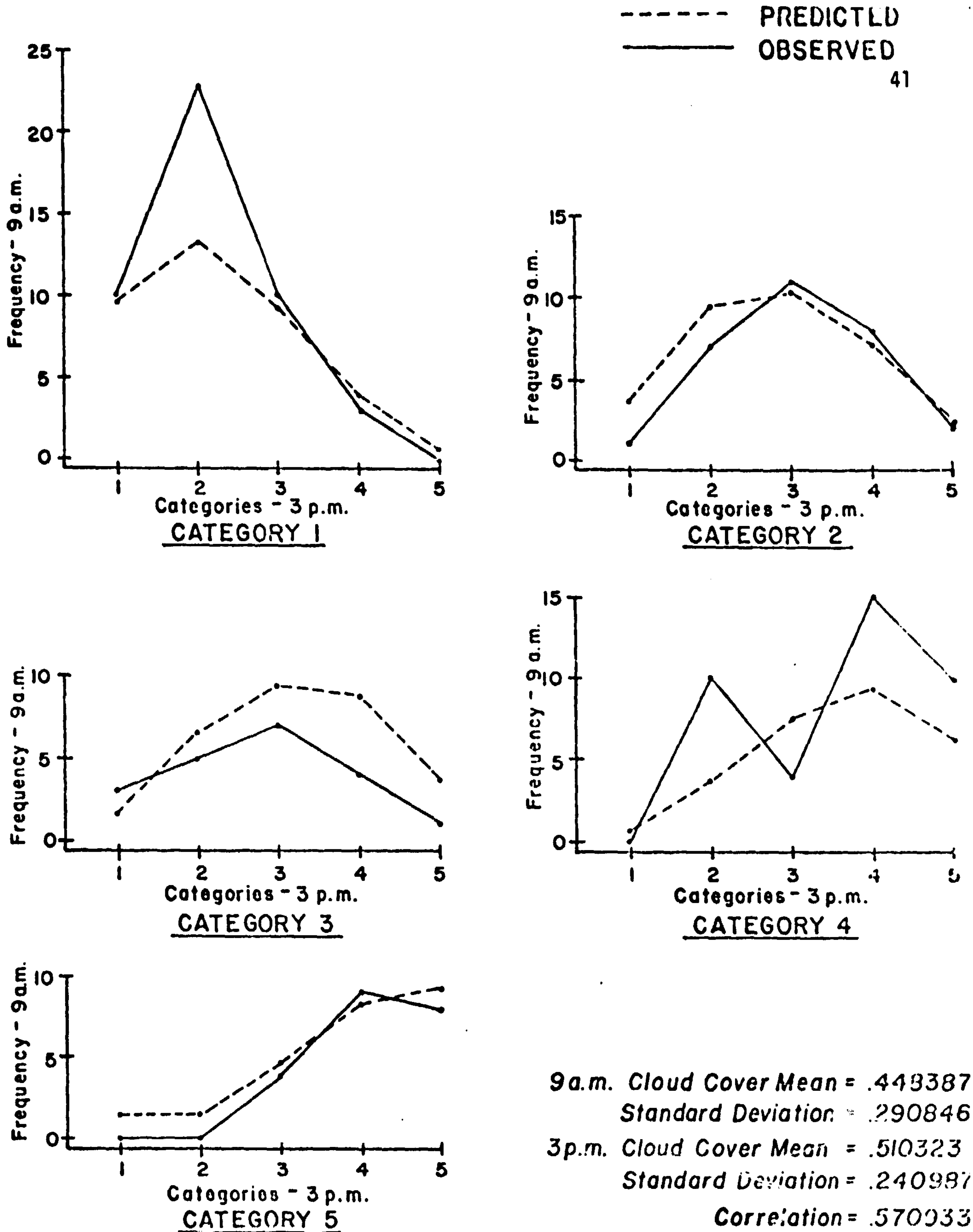
VI. EXAMPLES

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 discrepencies 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 $x \rightarrow 1$. Consequently, the fitted distribution does not reflect the variation in these



9a.m. Cloud Cover Mean = .448387
Standard Deviation = .290846
3p.m. Cloud Cover Mean = .510323
Standard Deviation = .240987
Correlation = .570033

Beta Parameters for 9a.m. are .862646 1.061241
Beta Parameters for 3p.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 6

Integration Limits

Category	Integration Limits	Midpoint
1	.01 to .2	.1
2	.2 to .4	.3
3	.4 to .6	.5
4	.6 to .8	.7
5	.8 to .99	.9

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.

JULY 71-75 FT.WORTH 9 A.M. AND 3 P.M.

***** RESULTS USING ORIGINAL DATA *****

FREQUENCY TABLE: 155 VALUES

10	3	9	11	9	1	0	3	0	0	0
1	0	1	2	4	2	0	2	0	0	1
0	0	0	1	2	2	0	2	1	0	1
0	0	1	2	0	1	0	0	1	2	0
2	2	0	2	4	2	1	0	2	0	1
1	0	0	1	0	1	0	1	0	0	0
0	0	0	2	0	1	0	2	0	0	1
0	0	2	5	1	2	0	0	2	1	0
0	0	0	1	0	0	0	2	1	0	4
0	0	0	0	0	0	2	1	1	3	5
0	0	0	0	1	3	0	0	4	5	8

MEANS=	0.4151	0.5065
ST. DEV=	0.3798	0.3162
CORR =	0.6411	

ESTIMATED BETA PARAMETERS

SITE I	0.2849	0.3998
SITE II	0.7600	0.7406

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

FREQUENCY TABLE: 155 VALUES

10	23	10	3	0
1	7	11	8	2
3	5	7	4	1
0	10	4	15	10
0	0	4	9	8

MEANS=	0.4434	0.5103
ST. DEV=	0.2908	0.2410
CORR =	0.5709	

SITE I	0.8626	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		SECOND SITE	
	UPPER	LOWER	UPPER	LOWER
INTEGRAL LIMITS	0.80000000	0.20000000	0.80000000	0.20000000
CATEGORY DATA Z'S	1.02408124	-0.64939553	1.08935735	-1.17765974
ORIGINAL DATA Z'S	0.69792524	-0.69792529	0.71430073	-0.75874548

TABLE 7

INPUT PARAMETERS

```
MEANS=                              0.0         0.0
ST. DEV=                            1.0000      1.0000
CORRELATION=                        0.5709
MAX ERROR=                          0.0030
UPPER BOUNDS=                       1.0241      1.0894
LOWER BOUNDS=                      -0.6494     -1.1777


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

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


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

+++THE REGULAR PART HAS BEEN REMOVED
THE VALUE IS    0.07902
ST. ERROR =     0.00065
CORR  =         0.55691


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

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


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

+++THE REGULAR PART HAS BEEN REMOVED
THE VALUE IS    0.13429
ST. ERROR =     0.00000
CORR  =         0.74074


THE TOTAL PROBABILITY IS   0.47424 WITH A STANDARD ERROR OF      0.00018
```

TABLE 8

REFERENCES

Falls, L.W. (1973). The Beta Distribution: A Statistical Model for World Cloud Cover. Nasa TM X-64714.

Falls, L.W. (1974). The Beta Distribution: A Statistical Model for World Cloud Cover. Journal of Geophysical Research, 79, pp. 1261-1264.

Fulks, W. (1969). Advanced Calculus, second edition. John Wiley & Sons, N.Y.

Hahn, G.J. and Shapiro, S.S. (1967). Statistical Models in Engineering. John Wiley & Sons, N.Y.

Johnson, N.L. and Kotz, S. (1972). Distributions in Statistics: Continuous Multivariate Distributions, Vol. IV. John Wiley & Sons, N.Y.

Newman, T.G. and Odell, P.L. (1971). The Generation of Random Variates. Hafner, New York.

Pearson, K. (1931). Biometrika Tables for Statisticians and Biometricians, Vol. II. Cambridge University Press.

Peizer, D.B. and Pratt, J.W. (1968). A Normal Approximation for Binomial, F, Beta, and Other Common, Related Tail Probabilities, I. Journal of the American Statistical Association, 63, 1416-1456.

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

Card 1	1-14	mean of the first variable of the BVN distribution
	15-28	mean of the second variable
Card 2	1-14	standard deviation of the first variable
	15-28	standard deviation of the second variable
	29-42	correlation between the two variables
Card 3	1-14	maximum standard error allowed
	21-25	odd, five digit random integer
Card 4	1-14	upper integration limit for the first variable
	15-28	lower integration limit for the first variable
Card 5	1-14	upper integration limit for the second variable
	15-28	lower integration limit for the second variable

```
C************************************CARD INPUT***************************************
C
C   CARD  COLS.   VARIABLE
C
C
C     1    1-14    MEAN AT FIRST SITE,REAL NUMBER
C     1    15-28   MEAN AT SECOND SITE, REAL NUMBER
C     2    1-14    STD.DEV. AT FIRST SITE, REAL NUMBER
C     2    15-28   STD. DEV. AT SECOND SITE, REAL NUMBER
C     2    29-42   CORRELATION, REAL NUMBER
C     3    1-14    ERROR BOUND
C     3    21-25   INTEGER RANDOM NUMBER
C     4    1-14    UPPER BOUND AT FIRST SITE
C     4    15-28   LOWER BOUND AT FIRST SITE
C     5    1-14    UPPER BOUND AT SECOND SITE
C     5    15-28   LOWER BOUND AT SECOND SITE
C
      IMPLICIT REAL*8(A-H,O-Z)
      COMMON  B,M1,M2,RO,SIG1,SIG2
      REAL*8 C(8)/8*0./,X(4),NT,M1,M2,B(8,9),Q(4)/4*0./,LO1(4),UP1(4),LO
     C2(4),UP2(4),L1,L2
      INTEGER RAND
   15 READ(5,1,END=100) M1,M2,SIG1,SIG2,RO,ERROR,RAND,U1,L1,U2,L2
    1 FORMAT(2F14.8/3F14.8/F14.8,6X,I5/2F14.8/2F14.8)
      WRITE(6,200)
  200 FORMAT(' INPUT PARAMETERS',//)
      WRITE(6,2) M1,M2,SIG1,SIG2,RO
    2 FORMAT(' MEANS=',20X,2F10.4,/,' ST. DEV=',13X,2F10.4,/,
     * ' CORRELATION=',14X,F10.4)
      WRITE(6,3) ERROR
    3 FORMAT(' MAX ERROR=',16X,F10.4)
      WRITE(6,4) U1,U2,L1,L2
    4 FORMAT(' UPPER BOUNDS=',13X,2F10.4,/,
     *  ' LOWER BOUNDS=',13X,2F10.4,//)
      UP1(1)=(L1+U1)/2D0
      UP1(2)=U1
      UP1(3)=UP1(1)
      UP1(4)=U1
      UP2(1)=(L2+U2)/2D0
      UP2(2)=UP2(1)
      UP2(3)=U2
      UP2(4)=U2
      LO1(1)=L1
      LO1(2)=(L1+U1)/2D0
      LO1(3)=L1
      LO1(4)=LO1(2)
      LO2(1)=L2
      LO2(2)=L2
      LO2(3)=(L2+U2)/2D0
      LO2(4)=LO2(3)
      B(2,1)=-1.0D0
      B(2,3)=-1.0D0
      B(2,2)=RO
      B(4,1)=3D0
      B(4,5)=3D0
      B(4,2)=-3D0*RO
      B(4,4)=-3D0*RO
      B(4,3)=2D0*RO**2+1D0
      B(6,1)=-15D0
```

```
      B(6,7)=-15D0
      B(6,2)=15D0*RO
      B(6,6)=15D0*RO
      B(6,3)=-3D0-12D0*RO**2
      B(6,4)=9D0*RO+6D0*RO**3
      B(6,5)=-3D0-12D0*RO**2
      B(8,1)=105D0
      B(8,2)=(-105D0)*RO
      B(8,3)=15D0+90D0*RO**2
      B(8,4)=-45D0*RO-60D0*RO**3
      B(8,5)=72D0*RO**2+24D0*RO**4+9D0
      B(8,6)=B(8,4)
      B(8,7)=B(8,3)
      B(8,8)=B(8,2)
      B(8,9)=B(8,1)
      ROSQ=1.0D0-RO**2
      C(1)=(1.0D0)/(2D0*3.14159265359D0*SIG1*SIG2*ROSQ**(.5D0))
      DO 8 J=1,4
      JJ=2*J
    8 C(JJ)=C(1)/ROSQ**J
      SU=0D0
      SE=0D0
      DO 16 I=1,4
   16 CALL TAYLOR(Q(I),C,UP1(I),UP2(I),LO1(I),LO2(I),1)
      DO 17 I=1,4
      WRITE(6,18) I
   18 FORMAT(//,' ***** APPROX. FOR REGION NO. =',I4,'   *****',//)
      REGP=Q(I)
      NT=1000
      COFG=0D0
      GSUM=0D0
      FSUM=0D0
      FGSQ=0D0
      FSQ=0D0
      PRO=(UP1(I)-LO1(I))*(UP2(I)-LO2(I))
    5 DO 6 II=1,1000
      DO 7 J=1,2
      CALL RANDU(RAND,IY,YFL)
      RAND=IY
    7 X(J)=YFL
      X(1)=LO1(I)+X(1)*(UP1(I)-LO1(I))
      X(2)=LO2(I)+X(2)*(UP2(I)-LO2(I))
      HOL=(-1.0D0/(2D0*ROSQ))*((X(1)-M1)**2/SIG1**2-2D0*RO*(X(1)-M1)*
     C(X(2)-M2)/(SIG1*SIG2)+(X(2)-M2)**2/SIG2**2)
      F=C(1)*DEXP(HOL)
      F=F*PRO
      FSQ=FSQ+(F**2)
      FSUM=FSUM+F
      CALL TAYLOR(G,C,X(1),X(2),LO1(I),LO2(I),2)
      G=G*PRO
      GSUM=GSUM+G
      FGSQ=FGSQ+((F-G)**2)
      COFG=COFG+F*(F-G)
    6 CONTINUE
      FGSUM=FSUM-GSUM
      FGM=FGSUM/NT
      FVAR=(FSQ-(FSUM**2/NT))/(NT-1.0D0)
      VVAR=(FGSQ-FGSUM**2/NT)/(NT-1.0D0)
```

```
      SFG=DSQRT(VVAR/NT)
      COF=(COFG-FSUM*FGSUM/NT)/(NT-1.0D0)
      SF=DSQRT(FVAR/NT)
      FM=FSUM/NT
      IF (SFG-ERROR) 9,9,10
   10 IF(SF-ERROR) 12,12,13
   13 NT=NT+1000
      GO TO 5
   12 CONTINUE
      WRITE(6,114)
  114 FORMAT(' +++THE REGULAR PART IS NOT REMOVED')
      WRITE(6,14) FM,SF
      SE=SE+SF
      SU=SU+FM
      GO TO 17
    9 FG=FGM+REGP
      COF=COF/DSQRT(VVAR*FVAR)
      SE=SE+SFG
      SU=SU+FG
      WRITE(6,111)
  111 FORMAT(' +++THE REGULAR PART HAS BEEN REMOVED')
      WRITE(6,11) FG,SFG,COF
   14 FORMAT(' THE VALUE IS',F10.5,/,' ST. ERROR =',F11.5)
   11 FORMAT(' THE VALUE IS',F10.5,/,' ST. ERROR =',F11.5,/,
     *  ' CORR   = ', 3X,F11.5,/)
   17 CONTINUE
      SE=SE/400
      WRITE(6,19) SU,SE
   19 FORMAT('0'/'0THE TOTAL PROBABILITY IS',F10.5,' WITH A STANDARD ERR
     COR OF',F12.5)
      GO TO 15
  100 STOP
      END


      SUBROUTINE TAYLOR (Q,C,U1,U2,L1,L2,INC)
      IMPLICIT REAL*8(A-H,O-Z)
      COMMON   B,M1,M2,RO,SIG1,SIG2
      REAL*8 C(8),L1,L2,M1,M2,B(8,9)
      Q=C(1)
      IF (INC.EQ.1) Q=Q*(U1-L1)*(U2-L2)
      DO 12 K=2,8,2
      NK=K+1
      VAR2=1.0D0
      DO 12 J=1,NK
      JJ=J-1
      IF (J.LE.2) GO TO 17
      J3=J-2
      NVAR2=JJ
      DO 15 L=1,J3
   15 NVAR2=NVAR2*(JJ-L)
      VAR2=NVAR2
   17 IF (J-K) 18,19,19
   18 NVAR3=(K-JJ)
      KH=K-JJ-1
      DO 16 L=1,KH
   16 NVAR3=NVAR3*(K-JJ-L)
      VAR3=NVAR3
      GO TO 20
   19 VAR3=1.0D0
   20 VAR=1.0D0/(VAR2*VAR3)
      IF (INC.EQ.2) GO TO 14
      Q=Q+(VAR*(C(K)/(SIG1**(K-JJ)*SIG2**JJ))
     C*((U2-M2)**J-(L2-M2)**J)*(1.0D0/((NK-JJ)*J))*((U1-M1)**(NK-JJ)-
     C(L1-M1)**(NK-JJ))*B(K,J))
      GO TO 12
   14 Q=Q+(VAR*(C(K)/(SIG1**(K-JJ)*SIG2**JJ))
     C*((U2-M2)**JJ)*((U1-M1)**(NK-J))*B(K,J))
   12 CONTINUE
      RETURN
      END
```

APPENDIX B

Approximation Program - Card Input

	Cols.	
Card 1	1-4	number of data pairs
	5-80	19 pairs of data with each element of each pair right justified in a two column space; no decimal points
Card 2+	1-76	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.
Last Card	1-10	lower integration limit for the first site
	11-20	upper integration limit for the first site
	21-30	lower integration limit for the second site
	31-40	upper integration limit for the second site

```
C*************************************CARD INPUT****************************************
C
C   CARD  COLS.   VARIABLE
C
C     1    1-80    TITLE
C     2    1-4     NUMBER OF DATA FROM EACH SITE
C     2    5-80    ALTERNATING DATA; FIRST SITE CLOUD COVER THEN SECOND SITE CLOUD
C                  COVER BOTH AN INTEGER BETWEEN 1 AND 10 IN CONSECUTIVE TWO COLUMN
C                  SPACES.
C     3+   1-76   CONTINUE DATA INPUT AS ABOVE
C     4    1-10    LOWER INTEGRATION LIMIT FOR SITE ONE
C     4    11-20   UPPER INTEGRATION LIMIT FOR SITE ONE
C     4    21-30   LOWER INTEGRATION LIMIT FOR SITE TWO
C     4    31-40   UPPER INTEGRATION LIMIT FOR SITE TWO
C    ALL DATA ON CARD 4 MUST BE LESS THAN 1 SINCE WE ARE DEALING WITH THE BETA
C    DISTRIBUTION (AND GREATER THAN ZERO)
C
      IMPLICIT REAL*8 (A-H,M,O-Z)
      REAL*8 X(155),Y(155),CX(155),CY(155),MX,AA(10)
      INTEGER F(5,5)/25*0/,FO(11,11)/121*0/
      READ(5,76) (AA(I),I=1,10)
      WRITE(6,76) (AA(I),I=1,10)
   76 FORMAT(10A8)
   59 READ(5,37,END=100) N,(X(I),Y(I),I=1,N)
   37 FORMAT (I4,(38F2.1))
      DO 60 I=1,5
      DO 60 J=1,5
   60 F(I,J)=0
      DO 71 I=1,11
      DO 71 J=1,11
   71 FO(I,J)=0
      DO 72 I=1,N
      N1=X(I)*10D0+1.1D0
      N2=Y(I)*10D0+1.1D0
   72 FO(N1,N2)=FO(N1,N2)+1
      CALL CAT (X,CX,N)
      CALL CAT(Y,CY,N)
      DO 51 I=1,N
      N1=CX(I)
      N2=CY(I)
   51 F(N1,N2)=F(N1,N2)+1D0
      WRITE(6,200)
  200 FORMAT(//,' ***** RESULTS USING ORIGINAL DATA *****',//)
      WRITE(6,78) N
   78 FORMAT ('0',22X,' FREQUENCY TABLE;',I5,2X,'VALUES',//)
      DO 73 I=1,11
   73 WRITE(6,52) (FO(I,J),J=1,11)
   52 FORMAT (10X,11I5)
      DO 55 I=1,N
      CX(I)=(CX(I)-.5D0)/5D0
   55 CY(I)=(CY(I)-.5D0)/5D0
      CALL STAT(X,Y,MX,MY,SVX,SVY,RO1,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. DEV=',8X,2F10.4,/,
     *' CORR =',10X,F10.4,//)
      B1=((1D0-MX)/SVX)*(MX*(1D0-MX)-SVX)
      A1=(MX*B1)/(1D0-MX)
      B2=((1D0-MY)/SVY)*(MY*(1D0-MY)-SVY)
      A2=(MY*B2)/(1D0-MY)
      WRITE(6,201)
  201 FORMAT(' ESTIMATED BETA PARAMETERS',//)
      WRITE(6,50) A1,B1,A2,B2
   50 FORMAT(' SITE I',10X,2F10.4,/,' SITE II',9X,2F10.4,//)
      WRITE(6,202)
  202 FORMAT(//,' ***** RESULTS USING CATEGORICAL DATA *****',//)
   54 FORMAT('0',10X,'FREQUENCY TABLE:',I5,2X,'VALUES',//)
      WRITE(6,54) N
      DO 53 I=1,5
   53 WRITE(6,52) (F(I,J),J=1,5)
      WRITE(6,48) MCX,MCY,SIGC1,SIGC2,RO2
      HB1=((1D0-MCX)/SVCX)*(MCX*(1D0-MCX)-SVCX)
      HA1=(MCX*HB1)/(1D0-MCX)
      HB2=((1D0-MCY)/SVCY)*(MCY*(1D0-MCY)-SVCY)
      HA2=(MCY*HB2)/(1D0-MCY)
      WRITE(6,50) HA1,HB1,HA2,HB2
      WRITE(6,203)
  203 FORMAT(//,' ***** NOTE *****')
      WRITE (6,75)
   75 FORMAT(' ','IF A PARAMETER OR PARAMETERS IS LESS THAN OR EQUAL TO
     C .5, THE Z-VALUE IS'/' UNDEFINED. FOR FURTHER COMPUTATION THE PARA
     CMETER IS RESET TO .51')
   58 READ(5,56,END=100) Y1,Y3,Y2,Y4
   56 FORMAT (8F10.0)
      CALL CALZ(HB1,HA1,Y1,N,ZC1)
      CALL CALZ(HB2,HA2,Y2,N,ZC2)
      CALL CALZ(HB1,HA1,Y3,N,ZL3)
      CALL CALZ(HB2,HA2,Y4,N,ZL4)
      CALL CALZ(B1,A1,Y1,N,Z1)
      CALL CALZ(B2,A2,Y2,N,Z2)
      CALL CALZ (B1,A1,Y3,N,ZL1)
      CALL CALZ(B2,A2,Y4,N,ZL2)
      WRITE(6,74)
   74 FORMAT('0',31X,'FIRST SITE',20X,'SECOND SITE'/26X,'UPPER',10X,'LOW
     CER',10X,'UPPER',10X,'LOWER')
   70 FORMAT('0',2X,'INTEGRAL LIMITS',1X,2(F14.8,1X,F14.8,1X)/1X,'CATEGO
     CRY DATA Z''S',1X,2(F14.8,1X,F14.8,1X)/1X,'ORIGINAL DATA Z''S',1X,
     C2(F14.8,1X,F14.8,1X))
      WRITE(6,70) Y3,Y1,Y4,Y2,ZL3,ZC1,ZL4,ZC2,ZL1,Z1,ZL2,Z2
   68 GO TO 58
  100 STOP
      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=0D0
      SY=0D0
      SXS=0D0
      SYS=0D0
      SXY=0D0
      DO 47 I=1,N
      SX=SX+X(I)
      SY=SY+Y(I)
      SXS=SXS+(X(I))**2
      SYS=SYS+(Y(I))**2
   47 SXY=SXY+X(I)*Y(I)
      M1=SX/N
      M2=SY/N
      SVARX=(SXS-N*M1**2)/(N-1D0)
      SVARY=(SYS-(SY**2/N))/(N-1D0)
      RO=(SXY-N*M1*M2)/(N-1D0)
      RO=RO/DSQRT(SVARX*SVARY)
      RETURN
      END
```

```
      SUBROUTINE CALZ(B1,A1,Y,N,Z)
      IMPLICIT REAL*8 (A-H,M,O-Z)
      IF (A1.LE.5D-1) A1=51D-2
      IF (B1.LE.5D-1) B1=51D-2
      SX1=B1-.5D0
      SX2=A1-.5D0
      SXN=A1+B1-1D0
      P=1D0-Y
      D2=(SXN+.33333333D0)*Y-(A1-.33333333D0)+2D-2*(Y/B1-P/A1+(Y-.5D0)/
     C(A1+B1))
      DA=DABS(SX1-SXN*P)
      DLS=DLOG(SX1/(SXN*P))
      DLT=DLOG(SX2/(SXN*Y))
      Z=D2/DA*DSQRT(12D0*SXN/(6D0*SXN+1D0)*(SX1*DLS+SX2*DLT))
      RETURN
      END
```

```
      SUBROUTINE CAT (X,H,N)
      REAL*8 X(155),H(155)
      DO 38 I=1,N
      IF (X(I)-1D0) 39,40,40
   40 H(I)=5D0
      GO TO 38
   39 IF (X(I)-.6D0)41,42,42
   42 H(I)=4D0
      GO TO 38
   41 IF (X(I)-.4D0) 43,44,44
   44 H(I)=3D0
      GO TO 38
   43 IF (X(I)-.1D0) 45,46,46
   46. H(I)=2D0
      GO TO 38
   45 H(I)=1D0
   38 CONTINUE
      RETURN
      END
```

A PROGRAM TO COMPUTE CONDITIONAL BIVARIATE NORMAL PARAMETERS

Summary

This report derives the conditional bivariate normal parameters from an original quadravariate distribution. The paper presents the theory and appended is a computor 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 quadravariate 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

$$f\,(x_1,\, x_2, \ldots,\, x_k) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\,\{-\tfrac{1}{2}(\underline{x}-\underline{\mu})'\,\Sigma^{-1}(\underline{x}-\underline{\mu})\} \qquad (1)$$

where $\underline{\mu}^1 = (\mu_1,\, \mu_2, \ldots,\, \mu_k)$, the vector of mean values and

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots\sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots\sigma_{2k} \\ \sigma_{k1} & \sigma_{k2} & \cdots\sigma_{kk} \end{bmatrix} ,$$

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

$$\underline{\mu} = \left[\begin{array}{c} \underline{\mu}_1 \\ \hdashline \underline{\mu}_2 \end{array}\right] = \left[\begin{array}{c} \mu_1 \\ \mu_2 \\ \hdashline \mu_3 \\ \mu_4 \end{array}\right] \quad \text{and} \quad = \left[\begin{array}{c:c} \Sigma_{11} & \Sigma_{12} \\ \hdashline \Sigma_{21} & \Sigma_{22} \end{array}\right] = \left[\begin{array}{cc:cc} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \hdashline \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{array}\right],$$

then letting

$$\underline{x} = \left[\begin{array}{c} \underline{x}_1 \\ \hdashline \underline{x}_2 \end{array}\right] = \left[\begin{array}{c} x_1 \\ x_2 \\ \hdashline x_3 \\ x_4 \end{array}\right], \text{ we have}$$

$$f\ (\underline{x}_1|\underline{x}_2) = \frac{1}{2\pi|\Sigma^*|^{\frac{1}{2}}} \exp\ \{-\tfrac{1}{2}\ (\underline{x}_1 - \underline{\mu}^*)'\ (\Sigma^*)^{-1}(\underline{x}_1 - \underline{\mu}^*)\}\ . \tag{2}$$

Where

$$\Sigma^* = \Sigma_{11} - \Sigma_{12}\, \Sigma_{22}^{-1}\, \Sigma_{21} \tag{3}$$

and

$$\underline{\mu}^* = \underline{\mu}_1 + \Sigma_{12}\ \Sigma_{22}^{-1}\ (\underline{x}_2 - \underline{\mu}_2). \tag{4}$$

Computation of the parameters for this conditional distribution really reduces to computation of the quantities Σ^* and $\underline{\mu}^*$. Carefully note that the value of $\underline{\mu}^*$ includes values of $\underline{x}_2 = [x_3, x_4]'$ that must be specified before numerical values for $\underline{\mu}^*$ can be calculated.

Even for this rather easy case the actual expressions for Σ^* and $\underline{\mu}^*$ 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 $\underline{\mu}^*$ and the actual form is given below (letting $\sigma_{34} = \sigma_{43}$ for convenience).

$$\underline{\mu}^* = \begin{bmatrix} \mu_1 + \{(\sigma_{13}\sigma_{44} - \sigma_{14}\sigma_{33})(x_3-\mu_3) + (\sigma_{14}\sigma_{33} - \sigma_{13}\sigma_{34})(x_4-\mu_4)\} \,/\, (\sigma_{33}\sigma_{44} - \sigma_{34}^2) \\ \mu_2 + \{(\sigma_{23}\sigma_{44} - \sigma_{24}\sigma_{33})(x_3-\mu_3) + (\sigma_{24}\sigma_{33} - \sigma_{23}\sigma_{34})(x_4-\mu_4)\} \,/\, (\sigma_{33}\sigma_{44} - \sigma_{34}^2) \end{bmatrix} .$$

The matrix triple product $\Sigma_{12}\, \Sigma_{22}^{-1}\, \Sigma_{21}$ makes Σ^* 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 quadravariate 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 $\underline{\mu}^*$.

Example: The following data was input to the program

$\underline{\mu} = [21.58, -.04, 43.35, 1.25]'$

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

$\sqrt{\sigma_{22}} = 11.52$, $\rho_{23} = 1614$, $\rho_{24} = .8134$

$\sqrt{\sigma_{33}} = 15.47$, $\rho_{34} = .1524$

$\sqrt{\sigma_{44}} = 14.59$

$[x_3, x_4] = [43.35, 0]$

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 quadravariate normal in 4F10.4 Format.
Card 2	Standard deviation for variable 1 followed by correlations for variables 1&2, 1&3, and 1&3 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 I2 Format.
Card 7	1st set of x_3, x_4 values to be conditioned on
Card 8	2nd set of " " " " " " "
"	3rd " " " " " " " " " "
"	etc.

The source deck listing is given in Appendix II.

References

Morrison, D. F. (1967). Multivariate Statistical Methods, Wiley, N. York.

APPENDIX I

MEANS VECTOR = 21.5800 -0.4000 43.3500 1.2500

VARIANCE-COVARIANCE MATRIX

121.6609	6.3914	125.9621	-3.2025
6.3914	132.7104	28.7638	136.7336
125.9621	28.7638	239.3209	34.3978
-3.2025	136.7336	34.3978	212.8631

COND. VAR. COV.MATRIX

53.17973	4.83824
4.83823	44.71625

SD&CORR.MATRIX

7.29244	0.09922
0.09922	6.68762

VALUES CONDITIONED ON 43.3500 -0.8870

CONDITIONAL MEANS 21.7081 -0.8870

APPENDIX II

```
      DIMENSION SIGMA(2,2),XO(2),VX(2),US(2),COR(2,2),V(2,2),
     1 V2(2,2),V3(2,2),V4(2,2),U1(2),U2(2),X(2),VV24(2,2),VVV(2,2),SD(4),
     2 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.EQ.5) GO TO 4
      DO 4 J=L,4
      S(I,J)=RHO(I,J)*SD(I)*SD(I)
      S(J,I)=S(I,J)
    4 CONTINUE
      WRITE(6,5) (U(I),I=1,4)
    5 FORMAT('1'///' MEANS VECTOR = ',4(F10.4,2X))
      WRITE(6,6)
    6 FORMAT(' VARIANCE - COVARIANCE MATRIX'/)
      DO 7 I=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(1,2)=-V4(1,2)/D
      V4(2,1)=-V4(2,1)/D
      V4(1,1)=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 I=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
   13 SIGMA(I,J)=V1(I,J)-VVV(I,J)
      COR(1,1)=SQRT(SIGMA(1,1))
      COR(2,2)=SQRT(SIGMA(2,2))
      COR(2,1)=SIGMA(1,2)/(COR(1,1)*COR(2,2))
      COR(1,2)=COR(2,1)
      WRITE(6,14) SIGMA
      WRITE(6,15) COR
   14 FORMAT('OCOND. VAR. COV.MATRIX'//2(2X,F10.5)/)
   15 FORMAT('OSD&CORR.MATRIX'//2(2X,F10.5)/)
      READ(5,16) N
   16 FORMAT(I2)
      DO 23 M=1,N
      READ(5,17) (X(I),I=1,2)
   17 FORMAT(2F10.5)
      DO 18 I=1,2
   18 XU(I)=X(I)-U2(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('OVALUES 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 computor 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 satisfacory 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

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & \text{for } \lambda \neq 0. \\ \\ \log(y) & \text{for } \lambda \neq 0. \end{cases} \tag{1}$$

Andrews et. al. (1971) extended this transformation to the bivariate case. In their paper, they were able to find approximate maximum likelihood estimates for λ, by examining the contures 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 λ 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 $Y_{i_1}, \ldots Y_{i_{n_i}}$ denote a random sample of n_i p - dimensional observations from a population with finite mean μ_i and finite covariance Σ_i, for i = 1, 2, ..., 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 μ_i, and common covarince Σ, where

$$Y_{ij}^{(\lambda)} = (y_{ij}^{(\lambda_1)},\ldots,y_{ij}^{(\lambda_p)})^T \tag{2}$$

$$y_{ij}^{(\lambda_k)} = \begin{cases} (y_{ijk}^{(\lambda_k)} - 1)/\lambda_k & \text{for } \lambda_k \neq 0 \\ \log(y_{ijk}) & \text{for } \lambda_k = 0 \end{cases} \tag{3}$$

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

$$Y_{ij}^{(\lambda)} = D^{-1}(Y_{ij}^{\lambda} - J) \tag{4}$$

where

$D = \mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_p)$

J is a px1 vector of 1's

$Y_{ij}^{\lambda} = (y_{ij1}^{\lambda_1}, y_{ij2}^{\lambda_2},\ldots,y_{ijp}^{\lambda_p})^T.$

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

$$f(z) = \exp\{-1/2(z-\mu)^T\Sigma^{-1}(z-\mu)\}(2\pi)^{-p/2}|\Sigma|^{-1/2} \tag{5}$$

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,

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

Hence the joint likelihood function becomes

$$L(\lambda) = \left(\prod_{i=1}^{m} \prod_{j=1}^{ni} K_{ij} \right) \quad (2\Pi)^{-np/2} \; |\Sigma|^{-n/2}$$

$$\cdot \exp \left\{ -\tfrac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij}^{(\lambda)} - \mu)^T \; \Sigma^{-1} \, (Y_{ij}^{(\lambda)} - \mu) \right\} \tag{7}$$

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

$$L(\lambda) = K(2\Pi)^{-np/2} \; |\hat{\Sigma}|^{-n/2} \; \exp\{-np/2\} \tag{8}$$

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

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}^{(\lambda)} = \bar{Y}_i^{(\lambda)}$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij}^{(\lambda)} - \bar{Y}_i^{(\lambda)}) \; (Y_{ij}^{(\lambda)} - \bar{Y}_i^{(\lambda)})^T . \tag{9}$$

Equation (8) follows from equation (7) since

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

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\mathrm{tr}\, \hat{\Sigma}^{-1} (Y_{ij}^{(\lambda)} - \mu_i)\, (Y_{ij}^{(\lambda)} - \mu_i)^T\,)$$

$$= \mathrm{tr}\, \hat{\Sigma}^{-1} \Big(\sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij}^{(\lambda)} - \ _i)\, (Y_{ij}^{(\lambda)} - \hat{\mu}_i)^T\, \Big)$$

$$= n\, \mathrm{tr}\, (\hat{\Sigma}^{-1}\, \hat{\Sigma}) = np.$$

Equation (8) can be further simplified as

$$L(\lambda) = C \cdot h(\lambda) \tag{10}$$

where $C = (2\Pi)^{-np/2} \ \exp \ \{-np/2\}$

$$h(\lambda) = |K^{-2/n}\, \hat{\Sigma}\, |^{-n/2} \tag{11}$$

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

$$K^{2/n} = \Big(\prod_{i=1}^{m} \prod_{j=1}^{n_i} k_{ij} \Big)^{2/n}$$

$$= \prod_{k=1}^{p} \Big(\prod_{i=1}^{m} \prod_{j=1}^{n_i} (y_{ijk})^{\lambda_k - 1} \Big)^{1/n} \Big)^2$$

$$= \prod_{k=1}^{p} (\dot{y}_k)^{\lambda_k - 1})^2 \tag{12}$$

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

$$\hat{\Sigma} = \sum_{i=1} \sum_{j=1} (Y_{ij}^{(\lambda)} - \bar{Y}_i^{(\lambda)}) \; (Y_{ij}^{(\lambda)} - \bar{Y}^{(\lambda)})^T$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} D^{-1}(Y_{ij}^{\lambda} - \bar{Y}_i^{\lambda}) \; (Y_{ij}^{\lambda} - \bar{Y}_i^{\lambda})^T \; D^{-1} \tag{13}$$

Hence $|\Sigma|$, becomes

$$|\hat{\Sigma}| = |D^{-1}| \; \left|\sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij}^{\lambda} - \bar{Y}_i^{\lambda}) (Y_{ij}^{\lambda} - \bar{Y}_i^{\lambda})\right| \; |D^{-1}|$$

$$= |D^{-2}| \; |G| \tag{14}$$

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

$$\phi(\lambda) = \frac{|G|}{|K^{2/N} \;\; D^2|}$$

$$= \frac{|G|}{\left(\prod_{k=1}^{p} \lambda_k \dot{y}_k^{\lambda_k - 1}\right)^2}. \tag{15}$$

Note that equation (15) reduces to

$$\frac{\sum_{i=1}^{n} (y_i^{\lambda} - y^{\lambda})^2}{(\lambda \dot{y}^{\lambda-1})^2} \tag{16}$$

which was proposed by Box and Cox (1964) for the univariate case.

The function $\emptyset(\lambda)$ 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

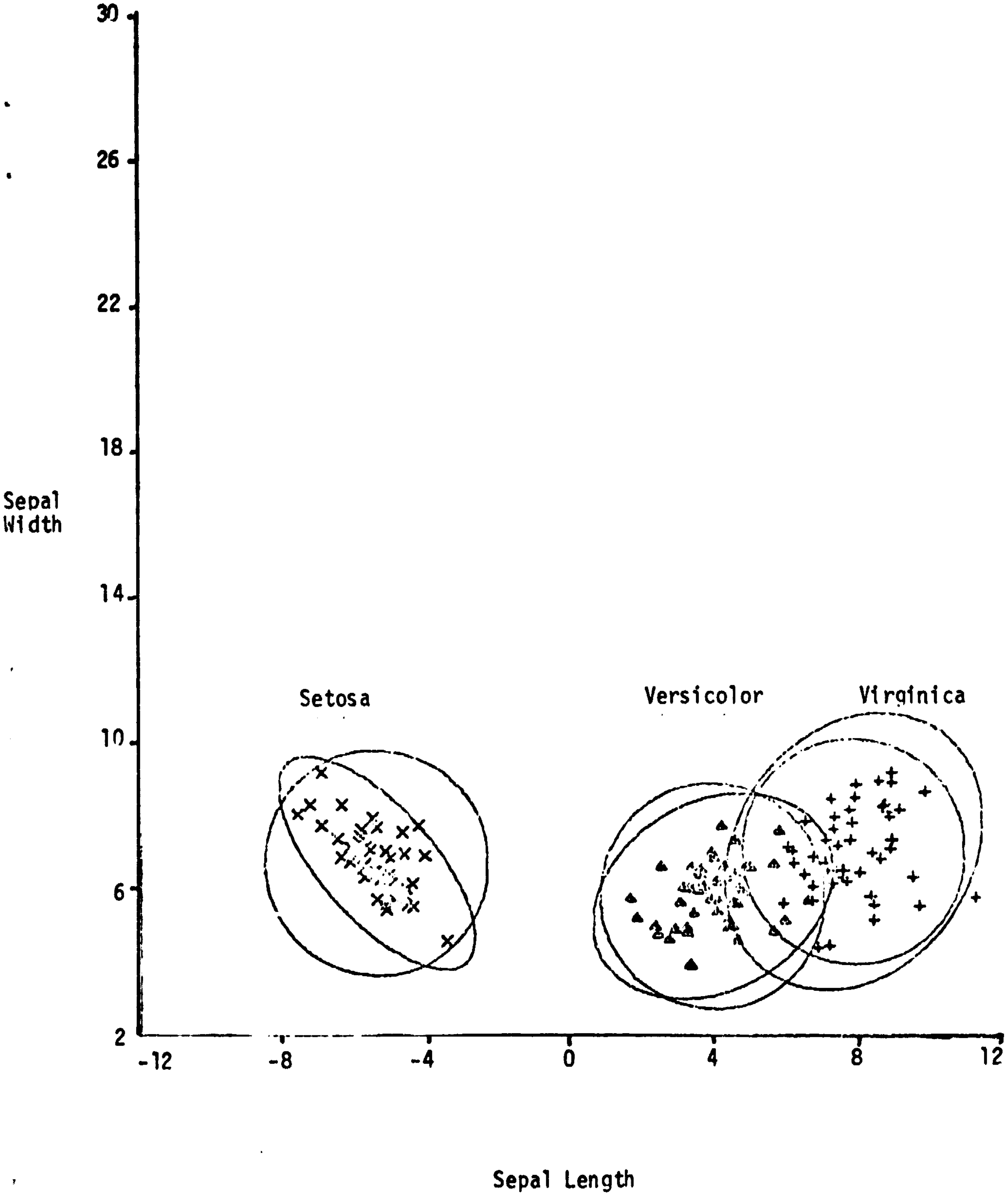
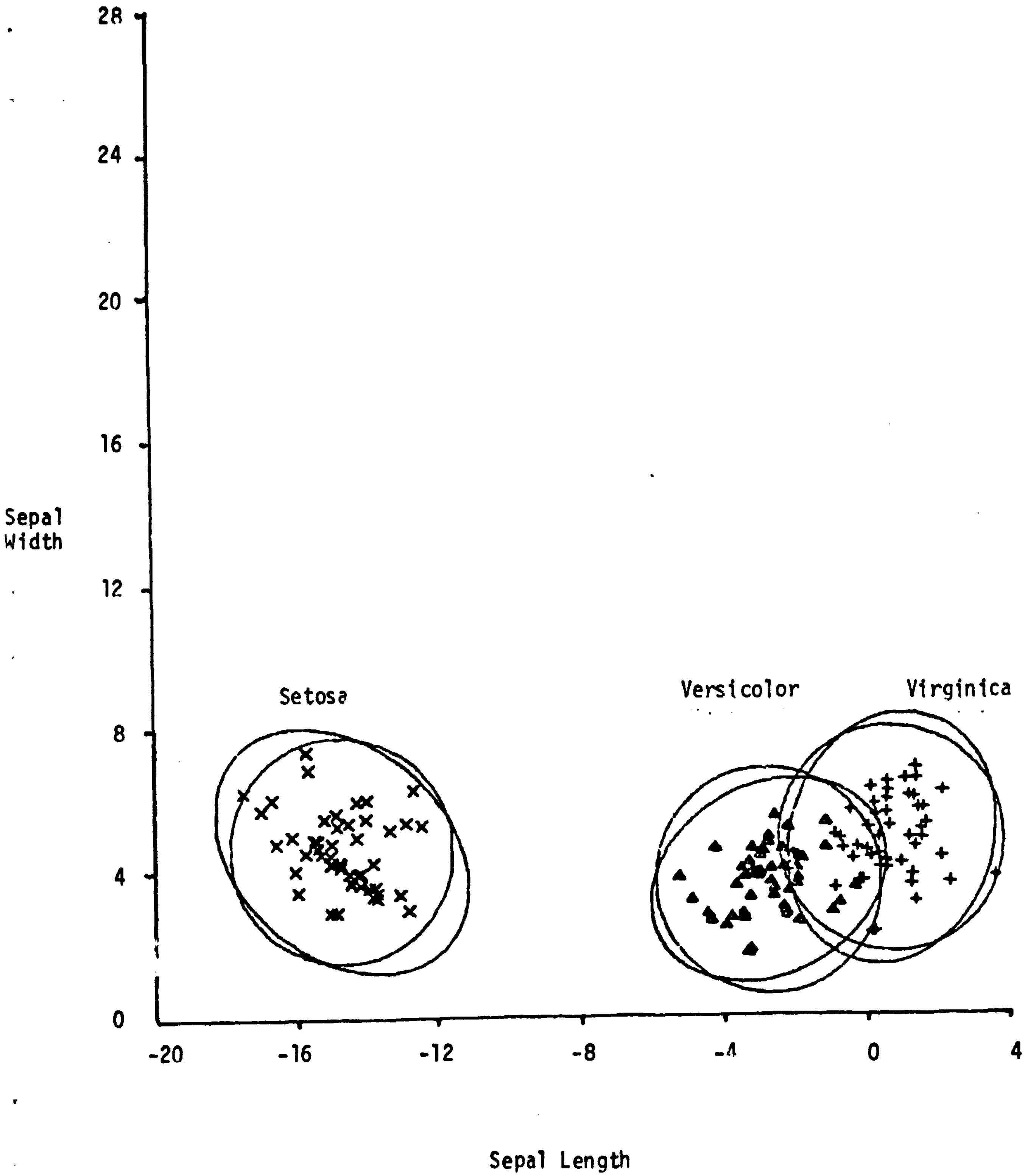


Figure 2. Transformed 95% Confidence Ellipses



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 objectes 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

Precipitation in inches	Data Set*							
	A		B		C		D	
	I**	II	I	II	I	II	I	II
1	.978	.966	.993	.999	.981	.971	.995	.997
2	.913	.903	.959	.999	.932	.924	.971	.976
3	.821	.819	.893	.962	.865	.864	.926	.931
4	.717	.723	.806	.809	.789	.794	.866	.866
5	.613	.624	.706	.625	.710	.719	.794	.788
6	.515	.528	.605	.472	.631	.644	.717	.706
7	.427	.439	.507	.361	.556	.571	.639	.623
8	.349	.359	.418	.283	.486	.500	.562	.544
9	.283	.291	.340	.227	.423	.436	.489	.471
10	.227	.232	.273	.186	.365	.376	.422	.405
15	.070	.066	.079	.090	.165	.166	.182	.174
20	.019	.016	.020	.057	.070	.066	.070	.076
25	.005	.003	.002	.042	.028	.025	.025	.032
30	.001	.001	.001	.033	.011	.009	.008	.023

* 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

1. Andrews,D.F.,Gnandesikan,R.,and Warner,J.L. (1971). Transformations of multivariate data. Biometrics vol. 27,p.825-840.

2. Box,G.E.P. and Cox,D.R. (1964). An analysis of transformations. J.R.Statistical Soc. series B vol.26,p.211-252.

Appendix A

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

$$\emptyset(\lambda) = \begin{bmatrix} \frac{\partial \emptyset}{\partial \lambda_1} \\ \frac{\partial \emptyset}{\partial \lambda_2} \\ \frac{\partial \emptyset}{\partial \lambda_p} \end{bmatrix} \tag{A.1}$$

where

$$\emptyset(\lambda) = \frac{|G|}{\left(\prod_{k=1}^{p} \lambda_k \dot{y}_k^{\lambda_k - 1}\right)^2} . \tag{A.2}$$

$$\frac{\partial \emptyset(\lambda)}{\partial \lambda_h} = \frac{\partial |G|}{\partial \lambda_h} \left(\prod_{k=1}^{p} \lambda_k \dot{y}_k^{\lambda_k - 1}\right)^{-2} + \frac{\partial \left(\prod_{k=1}^{p} \lambda_k \dot{y}_k^{\lambda_k - 1}\right)^{-2} |G|}{\partial \lambda_h} \tag{A.3}$$

$$\frac{\partial \left(\prod_{k=1}^{p} \lambda_k \dot{y}_k^{\lambda_k - 1}\right)^{-2}}{\partial \lambda_h} = -2\left(\prod_{k=1}^{p} \lambda_k \dot{y}_k^{\lambda_k - 1}\right)^{-2} (\lambda_h + \ln \dot{y}_h)\, \lambda_n^{-1} . \tag{A.4}$$

Since $|G| = \sum_{j=1}^{p} g_{ij}\, a_{ij}$ where

$$G = (g_{ij}) \tag{A.5}$$

a_{ij} is the cofactor of g_{ij}

Also, since g_{ij} only depends upon λ_i, λ_j using the chain rule we have

$$\frac{\partial|G|}{\partial\lambda_h} = \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial|G|}{\partial g_{ij}} \frac{\partial g_{ij}}{\partial\lambda_h} \tag{A.6}$$

where

$$\frac{\partial|G|}{\partial g_{ij}} = a_{ij} \tag{A.7}$$

and

$$\frac{\partial g_{uv}}{\partial\lambda_h} = \begin{cases} 0 & \text{if } u, v \neq h \\ b_2 & \text{if } u \text{ or } v = h \\ b_3 & \text{if } u = v = h \end{cases} \tag{A.8}$$

and

$$b_2 = \frac{\partial}{\partial\lambda_h} \left(\sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{iju}^{\lambda_u} - \bar{y}_{iu}^{\lambda_u}) (y_{ijh}^{\lambda_h} - \bar{y}_{ih}^{\lambda_h}) \right)$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{iju}^{\lambda_u} - \bar{y}_{iu}^{\lambda_u}) (y_{ijh}^{\lambda_h} \ln y_{ijh} - \bar{y}_{ih}^{\lambda_h} \ln \bar{y}_{ijh})$$

$$b_3 = 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ijh}^{\lambda_h} - \bar{y}_{ih}^{\lambda_h}) (y_{ijh}^{\lambda_h} \ln y_{ijh} - \bar{y}_{ih}^{\lambda_h} \ln \bar{y}_{ijh})$$

(A.9)

From this, equation (A.3) becomes

$$\frac{\partial\emptyset(\lambda)}{\partial\lambda_h} = \frac{2}{\left(\prod_{k=1}^{p} \lambda_k \dot{y}_k^{\lambda_k - 1}\right)^2} \left[\sum_{\substack{k=1 \\ k\neq h}}^{p} \alpha_{kh} \frac{\partial g_{kh}}{\partial\lambda_h} + \frac{\alpha_{hh}}{2} \frac{\partial g_{hh}}{\partial\lambda_h} - |G| \left(\lambda_h^{-1} + \ln \dot{y}_h\right) \right] . \tag{A.10}$$

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

$$H_0\colon X_1, X_2, \ldots, X_n \sim F_X(x;\theta) \tag{1}$$

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,

$$F_X(x;\theta) = \exp\{-\exp(-(x-\alpha)/\beta)\} \qquad -\infty < \alpha < \infty \qquad \beta > 0. \tag{2}$$

Let T denote a transformation from the population raw moments to ξ, which can be written as a linear function of the parameters θ where

$$\begin{aligned} \eta' &= (\eta_1', \eta_2', \ldots, \eta_s')^T \\ \xi &= (\xi_1, \xi_2, \ldots, \xi_s)^T \end{aligned} \tag{3}$$

and η_j' is the j^{th} raw population moment for F_X and $\xi = W\theta$, W is a known sx2 matrix, and $\theta = (\alpha,\beta)^T$. That is,

$$T\colon \eta \rightarrow \xi = W\theta. \tag{4}$$

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

$$T\colon m' \rightarrow h. \tag{5}$$

By the central limit theorem, we know

$$n(m' - \eta') \sim n(\emptyset, G) \tag{6}$$

where the ij^{th} element of the matrix G is

$$g_{ij} = \eta'_{i+j} - \eta'_i \eta'_j \tag{7}$$

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

$$n(h - \xi) \sim n(\emptyset, \Sigma) \tag{8}$$

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

$$Q^* = n(h - \xi)^T \hat{\Sigma}^{-1} (h - \xi) \tag{9}$$

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

$$\theta = (W^T \hat{\Sigma}^{-1} W)^{-1} W^T \hat{\Sigma}^{-1} h. \tag{10}$$

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

$$\hat{Q} = nh^T\hat{A}h \tag{11}$$

where

$$\begin{aligned} \hat{A} &= \hat{\Sigma}^{-1} (I - \hat{R}) \\ \hat{R} &= W(W^T \hat{\Sigma}^{-1} W)^{-1} W^T \hat{\Sigma}^{-1}. \end{aligned} \tag{12}$$

Again by the distributional properties of the quadratic forms, $\hat{Q}$ has a non-central chi-square distribution with degrees of freedom $=$ tr $\hat{\Sigma}\hat{A}$ and non centrality parameter $\lambda = \xi^T A \xi$ if and only if $\hat{\Sigma}\hat{A}$ is an idempotent matrix. It is easy to verify that $(\hat{\Sigma}\hat{A})^2 = \hat{\Sigma}\hat{A}$, 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 $\hat{Q} > \chi_\alpha^2(s-q)$, where

$$\Pr(X \geq \chi_\alpha^2(s-q)) = \alpha. \tag{13}$$

Dahiya and Gurland (1970) developed the non-null distribution for $\hat{Q}$, 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

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

where

$$\psi_{(1)}^{(n)} = (-1)^{n+1} n!\,\delta(n+1)$$

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

By letting $\xi = (\kappa_3\kappa_2^{-1}, \kappa_4\kappa_3^{-1}, \ldots, \kappa_{s+2}\kappa_{s+1}^{-1})^T$ and $W = (\psi_{(1)}^{(2)}/\psi_{(1)}^{(1)}, \ldots, \psi_{(1)}^{(s+1)}/\psi_{(1)}^{(s)})^T$, and $\theta = \beta$ it is possible to map $\eta' \to \xi$ where $s = 4$ and $q = 1$. By letting $h = (h_1, h_2, h_3, h_4)^T$, where $h_j = k_{j+2}/k_{j+1}$, for $j=1,2,..,4$, and k_j is the j^{th} sample cumulant. We are now able to compute Q, where

$$\Sigma = JGJ^T|_{\beta=\hat{\beta}} \tag{16}$$

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

and $\hat{\beta}$ is the maximum likelihood estimate for β.

The values in equation (15) can be found in Abrahomovich, hence J becomes

$$J = 1/\beta \begin{bmatrix} 1 & 0 & 0 & 0 \\ -.885 & .6079 & 0 & 0 \\ 0 & -1.131 & .4174 & 0 \\ 0 & 0 & -.5901 & .154 \end{bmatrix} \quad (17)$$

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

Method of Moments				Generalized minimum χ^2			
$\hat{\alpha}$	$\hat{\beta}$	X*	$F_X(x)$	$\hat{\alpha}$	$\hat{\beta}$	X*	$G_X(x)$
102.49	1.39	100.90	.0433	102.53	1.25	100.90	.0251
		101.6C	.1625			101.66	.1346
		102.61	.3994			102.61	.3914
		103.24	.5582			103.24	.5674
		104.22	.7497			104.22	.7720
		105.72	.9067			105.72	.9250
		106.50	.9457			106.50	.9591

* 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

Method of Moments				Generalized minimum χ^2			
$\hat{\alpha}$	$\hat{\beta}$	X*	$F_X(x)$	$\hat{\alpha}$	$\hat{\beta}$	X*	$G_X(x)$
103.83	1.25	102.54	.0604	103.33	1.57	102.54	.2118
		103.31	.2196			103.31	.3866
		103.94	.4002			103.94	.5293
		104.52	.5623			104.52	.6442
		106.15	.8553			106.15	.8558
		106.50	.8889			106.50	.8829

* same as in Table 1

TABLE 3

Extreme Monthly Temperatures

Site	Extreme Value Distribution $\hat{\alpha}$	$\hat{\beta}$	Q*
New Orleans	83.8	.98	.001
Orlando	84.8	.88	.003
Daytona Beach	81.7	.67	.002

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

TABLE 4

Extreme Monthly Winds

Site	Extreme Value Distribution $\hat{\alpha}$	$\hat{\beta}$	$\hat{Q}$*
New Orleans	15.4	2.9	.9
Orlando	13.6	2.5	.6
Daytona Beach	13.0	2.2	.4

* 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 meteorogical interpretation.

V. References

1. Dahiya,R.C.,and Gurland,J. (1970). A test of fit for continous distributions based on generalized minimum chi-square. Statistical Papers in Honor of George W. Snedecor, T.A. Bancroft, editor

2. Dahiya,R.C.,and Gurland,J. (1972). Goodness of fit tests for the gamma and exponential distributions, Technometrics, vol. 14,p.791-801.

3. Gumbel,E.J.,and Goldstein,N. (1964). Analysis of empirical bivariate extremal distribution, JASA vol. 59,p.794-816.

Test of Fit for Continuoas 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-sqare 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 severally 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

$$Ho: \; x_1, x_2, \ldots x_n \sim F_X(x;\theta) \in \mathcal{F}(x;\Theta) \tag{1}$$

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

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

$$T: \; \mu \rightarrow \xi \tag{2}$$

where $\mu' = (\mu'_1, \mu'_2, \ldots \mu'_s)^T$, μ'_j is the j^{th} raw population moment and $\xi = (\xi_1, \xi_2, \ldots \xi_s)^T$ can be expressed as $\xi = W\theta$ (3)

for a known s x q matrix w and s > q. Let $m' = (m'_1, m'_2, \ldots m'_s)^T$ denote a s x 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

$$n(m' - \mu') \rightarrow n(\emptyset, G) \tag{4}$$

where $G = (g_{ij})$, $g_{ij} = \mu'_{i+j} - \mu'_i \mu'_j$, i, j = 1, 2, ...,s.

From this, it can be shown that

$$n(h - \xi) \rightarrow N(\emptyset, \Sigma) \quad (5)$$

where $\Sigma = JGJ^T$, J the jacobian matrix for the transformation T. Now using the properties of quadratic forms, we know that

$$Q = n(h - \xi)^T \ \Sigma^{-1}(h - \xi) \quad (6)$$

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

$$\hat{\theta} = (W^T \Sigma^{-1} W)^{-1} \ W^T \ \hat{\Sigma}^{-1} h. \quad (7)$$

By letting $\xi = W\theta$, the minimal Q is

$$\hat{Q} = n(h - \hat{\xi})^T \ {}^{-1}(h - \xi) = nh^T Ah \quad (8)$$

where

$$\hat{A} = \hat{\Sigma}^{-1}(I - \hat{R}) \quad (9)$$

$$\hat{R} = W(W^T \ \hat{\Sigma}^{-1} \ W)^{-1} \ W^T.$$

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

Gurland and Dahiya (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.

<u>Normal Distribution</u>

Suppose one would like to test the following hypothesis

$$H_0: X_1, X_2, \ldots, X_n \sim F_X(x;\theta) \in N(\mu, \sigma^2) \tag{10}$$

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

$$\xi = (\xi_1 = \mu_1', \xi_2 = \log\theta_2, \xi_3 = \mu_3, \xi_4 = \log(\tfrac{1}{3}\mu_4))^T$$

we have

$$\xi = W\theta^* \tag{11}$$

where

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

$$W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 2 \end{bmatrix} \tag{12}$$

The transforamtion T from μ' to ξ can be achieved in two steps; $T_1: \mu' \to \mu$ $T_2: \mu \to \xi$. Hence, Σ in equation (5) becomes

$$\Sigma = J_2 J_1 G J_1^T J_2^T \tag{13}$$

where

$$J_1 = (j_{mn}); \quad j = \frac{\partial \mu'_m}{\partial \mu_n} \qquad m, n = 1, 2, \ldots, s$$

$$J_2 = (j_{uv}); \quad j = \frac{\partial \mu_u}{\partial \xi_v} \qquad u, v = 1, 2, \ldots, s.$$

By assuming that $\mu'_1 = 0$, J_1 and J_2 become

$$J_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -3\theta_2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{15}$$

$$J_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/\theta_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{1}{3\theta_2^2} \end{bmatrix} \tag{16}$$

and equation (14), becomes

$$\Sigma = \begin{bmatrix} \theta_2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 4 \\ 0 & 0 & 6\theta_3^2 & 0 \\ 0 & 4 & 0 & 32/3 \end{bmatrix} \tag{17}$$

Since θ_2 is unknown, let $\hat{\theta}_2$ denote the usual maximum likelihood estimate. Then $\hat{\Sigma} = \Sigma \Big|_{\theta_2 = \hat{\theta}_2}$. Now by computing $\hat{\theta}$ and $\hat{Q}$ in equation (7) and (8), one can test the hypothesis (10).

Gamma Distribution

Test the hypothesis

$$Ho: \; X_1, X_2, \ldots, X_n \sim F_X(x;\theta) \sim \Gamma(\theta_1, \theta_2) \tag{18}$$

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

$$f_X(x;\theta_1,\theta_2) = \frac{e^{-y}\, y^{\theta_1 - 1}}{\theta_2\, \Gamma(\theta_1)} \; ; \; y = x/\theta_2 \tag{19}$$

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

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

$$\xi = (\xi_1 = \kappa_1, \xi_2 = \kappa_2 \kappa_1^{-1}, \xi_3 = \kappa_3 \kappa_2^{-1}, \; \xi_4 = \kappa_4 \kappa_3^{-1})^T \tag{20}$$

$$W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 2 \\ 0 & 3 \end{bmatrix} \qquad \theta^* = (\theta_1^* = \theta_1 \theta_2, \theta_2^* = \theta_2) \tag{21}$$

The transformation T from η' to ξ can be obtained in two steps

$$\begin{aligned} T_1 &: \; \eta' \to K \\ T_2 &: \; K \to \xi \end{aligned} \tag{22}$$

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

$$\Sigma = J_2 J_1 G J_1^T J_2^T \tag{23}$$

where

$$J_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ j_{12} & 1 & 0 & 0 \\ j_{13} & j_{23} & 1 & 0 \\ j_{14} & j_{24} & j_{34} & 1 \end{bmatrix} \tag{24}$$

$$j_{12} = -2\,\eta_1' \qquad\qquad j_{13} = -3\eta_1'$$

$$j_{23} = -3\eta_2 + 6\eta_1' \qquad\qquad j_{14} = -4\eta_3' + 12\eta_3'\,\eta_1' - 24(\eta_1')^3$$

$$j_{24} = -6\ {}_2' + 12(\eta_1')^2 \qquad\qquad j_{34} = -4\eta_1'$$

$$\eta_j' = \frac{\Gamma(\theta_1 + j)}{\Gamma(\theta_1)}\,\theta_2^j; \quad j = 1, 2, 3, 4, \ldots \tag{25}$$

$$J_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\kappa_2\kappa_1^{-1} & \kappa_1^{-1} & 0 & 0 \\ 0 & -\kappa_3\kappa_2^{-1} & \kappa_2^{-1} & 0 \\ 0 & 0 & -\kappa_4\kappa_3^{-1} & \kappa_3^{-1} \end{bmatrix} \tag{26}$$

Since θ_1, θ_2 are unknown, they can be estimated by $\hat{\theta}_1, \hat{\theta}_2$ where

$$\hat{\theta}_2 = X/\hat{\theta}_1$$

$$\hat{\theta}_1 = y^{-1}/4\ (1 + (1 + 4y/3)^{\frac{1}{2}})$$

$$y = \log\ (\overline{X}/GM)$$

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{27}$$

$$GM = (\prod_{i=1}^{n} X_i)^{1/n}$$

By replacing $\hat{\theta}_1 \hat{\theta}_2$ in Σ, 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 1

Evaluation GMCS procedures using Simulated Data

True Distribution	Parameters		Sample Size	Estimated Parameters		$\hat{Q}$
$\Gamma(\gamma, \beta)$	γ	β		$\hat{\gamma}$	$\hat{\beta}$	
	3	1	10	1.2	1.06	6.600*
	"	"	25	1.0	.91	.001
	"	"	50	1.1	.86	1.200
	"	"	100	1.1	.90	4.900*
	2	1	10	.97	.98	5.600*
	"	"	85	.88	.88	14.900*
	"	"	50	1.18	.96	12.700
	"	"	100	.83	.72	3.300
	.5	1	10	1.99	1.6	.420
	"	"	25	.80	.63	1.000
	"	"	50	1.02	.77	1.200
	"	"	100	1.17	.91	15.100*
$N(\mu, \sigma^2)$	μ	σ^2	NØB	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{Q}$
	10	25	10	12.1	11.8	.008
			25	9.5	31.5	.091
			50	8.9	20.0	.041
			100	10.2	23.9	.001
Extreme value, α, β	α	β	NOB	$\hat{\alpha}$	$\hat{\beta}$	$\hat{Q}$
	5.	1.	10	5.01	1.68	.001
			25	5.04	1.15	.008
			50	5.04	.85	.003
			100	4.82	.85	.006
	2.	2.	10	2.90	.98	.002
			25	2.69	1.50	.004
			50	1.74	2.08	.017
			100	2.09	1.95	.033
Exponential λ		λ	NOB	λ		$\hat{Q}$
		.5	10	.69		9.04 *
			25	.56		3.2
			50	.53		1.3
			100	.42		4.15 *
		1.0	10	1.04		.33
			25	.83		.75
			50	1.28		.29
			100	1.1		2.90
		2.0	10	2.60		4.9 *
			25	1.89		1.54
			50	1.97		1.04
			100	1.92		.35

* 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 perceptation 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

Site**	Year	Normal			Exp		Gamma			Extreme		
		$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{Q}$	$\hat{\lambda}$	$\hat{Q}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{Q}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{Q}$
I	1936-40	4.8	24	3.3	.02	7.6*	1.9	.04	.0	12.9	3.0	3.5*
	41-45	4.5	12	1.3	"	8.2*	2.2	.05	"	8.3	2.4	1.8
	46-50	5.4	20	.8	"	5.3*	1.8	.03	"	9.4	3.1	3.3*
	51-55	4.6	9	1.4	"	7.8*	---	---	---	6.7	2.2	1.4
	56-60	4.7	8	.3	"	10.6*	2.7	.06	"	7.3	2.1	1.1
	61-65	4.6	7	.0	"	9.3*	2.3	.05	"	6.3	2.1	1.1
	66-70	4.4	8	.3	"	8.6*	2.3	.05	"	6.7	2.2	1.2
	71-75	5.8	10	.3	"	13.1*	3.5	.06	"	8.8	2.4	1.3
II	1936-40	4.2	13	.7	.02	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	.02	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 3

Monthly Mean Temperature

Site**	Year	Normal			Exp		Gamma			Extreme		
		$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{Q}$	$\hat{\lambda}$	$\hat{Q}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{Q}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{Q}$
I	1936-40	69.7	116	.0	.01	24.1*	26	.36	.0	75.2	8	.0
	41-45	69.4	120	"	"	"	25	.39	"	75.1	9	"
	46-50	69.4	106	"	"	"	29	.39	"	74.6	8	"
	51-55	69.2	111	"	"	"	25	.42	"	74.9	"	"
	56-60	68.5	121	"	"	"	25	.36	"	74.2	"	"
	61-65	67.5	121	"	"	"	22	.32	"	73.0	"	"
	66-70	67.0	130	"	"	"	23	.34	"	72.9	9	"
	71-75	68.6	99	"	"	"	23	.49	"	73.8	8	"
II	1936-40	71.0	69	.0	.01	24.6*	40	.57	.0	75.2	6	.0
	41-45	72.0	80	"	"	"	41	"	"	76.7	7	"
	46-50	73.4	58	"	"	"	43	"	"	77.0	6	"
	51-55	71.8	72	"	"	"	57	.80	"	76.0	7	"
	56-60	71.8	78	"	"	"	34	.48	"	76.1	7	"
	61-65	72.4	73	"	"	"	40	.55	"	76.6	"	"
	66-70	71.8	83	"	"	"	36	.51	"	76.3	"	"
	71-75	73.6	56	"	"	"	53	.72	"	77.4	6	"
III	1936-40	69.7	63	.0	.01	24.7*	41	.54	.0	73.7	6	.0
	41-45	70.1	86	"	"	"	33	.47	"	74.8	7	"
	46-50	71.5	61	"	"	"	40	.56	"	75.3	6	"
	51-55	70.4	75	"	"	"	55	.78	"	74.9	7	"
	56-60	70.0	82	"	"	"	32	.46	"	74.5	7	"
	61-65	69.8	76	"	"	"	39	.56	"	74.3	7	"
	66-70	70.0	89	"	"	"	34	.49	"	74.7	7	"
	71-75	71.3	60	"	"	"	34	.50	"	75.2	7	"

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

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

TABLE 4

Daily Maximum Temperature

Site**	Date***	Normal			Exp		Gamma			Extreme		
		$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{Q}$	$\hat{\lambda}$	$\hat{Q}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{Q}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{Q}$
I	1/70	47.5	141	.0	.02	11.6*	18.9	.39	.0	55.8	9.	.0
	3/70	60.0	44	"	.01	12.8*	49.2	.73	"	63.8	5.	.0
	6/70	79.7	20	"	"	"	116.0	1.4	"	81.8	3.	.0
	10/70	69.0	26	"	"	"	99.8	1.4	"	71.7	4.	.1
	1/71	55.3	112	"	"	"	23.5	.42	"	61.0	8.	.0
	3/71	59.4	75	"	"	"	23.3	.38	"	64.6	7.	.0
	6/71	80.2	5	"	"	"	------	------	------	81.9	2.	.0
	10/71	71.8	23	"	"	"	70.5	.98	"	74.1	4.	.0
II	1/70	55.	94	.0	.01	12.2*	18	.32	.0	60.	7.	.0
	3/70	76.	22	"	"	"	11	1.5	"	78.4	4.	"
	6/70	83.9	1	"	"	"	67.8	8.	"	84.4	1.	"
	10/70	63.5	66	"	"	"	22.3	.35	"	67.1	6.	"
	1/71	64.3	87	"	"	"	22.2	.34	"	68.8	7.	"
	3/71	72.0	53	"	"	"	54.1	.74	"	76.0	6.	"
	6/71	83.4	3	"	"	"	45.8	5.5	"	84.3	1.	"
	10/71	71.6	17	"	"	"	61.9	.8	"	73.3	3.	"
III	1/70	54.7	94	.0	.01	12.5*	17.4	.3	.0	59.6	7.7	.0
	3/70	65.6	53	"	"	"	50.9	.7	"	70.0	5.6	"
	6/70	80.9	8	"	"	"	221.	2.7	"	82.4	1.3	"
	10/70	82.7	12	"	"	"	166.	2.1	"	78.7	2.7	"
	1/71	58.8	92	"	"	"	19.	.3	"	63.4	7.6	"
	3/71	60.1	65	"	"	"	51.	.8	"	65.2	6.2	"
	6/71	71.0	8	"	"	"	843.	10.6	"	80.8	2.5	"
	10/71	76.0	9	"	"	"	99.	1.2	"	77.5	2.4	"

* 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 5

Daily Maximum Wind

Site	Date***	Normal			Exp		Gamma			Extremely		
		$\hat{\mu}$	$\hat{\sigma}$	$\hat{Q}$	$\hat{\lambda}$	$\hat{Q}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{Q}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{Q}$
I	1/70	9.6	7	.0	.10	11.2*	14.8	1.5	.0	11.8	2.1	.0
	3/70	9.8	6	"	"	"	---	---	---	11.2	2.0	"
	6/70	6.8	6	"	.14	9.7*	7.7	1.1	.0	8.7	1.8	"
	10/70	7.6	9	1.3	.13	9.3*	5.6	.7	"	9.6	2.3	"
	1/71	8.4	12	.0	.11	8.7*	4.7	.5	"	10.6	2.7	"
	3/71	9.7	7	"	.10	11.1*	11.6	1.2	"	11.6	2.1	"
	6/71	5.3	2	"	.18	10.4*	8.6	1.6	"	6.1	1.2	"
	10/71	4.7	5	.3	.20	9.1*	6.5	1.3	"	7.1	1.6	"
II	1/70	9.6	10	.0	.10	10.4*	7.8	.8	.0	11.7	2.4	2.4
	3/70	10.3	10	"	.04	10.6*	8.3	.8	"	12.3	2.4	.0
	6/70	8.4	4	"	.12	11.1*	14.1	1.6	"	9.8	1.6	"
	10/70	8.8	6	"	.11	11.1*	10.9	1.2	"	10.5	1.9	"
	1/71	8.8	7	"	.11	10.7*	8.7	.9	"	10.4	2.0	"
	3/71	10.1	11	"	.11	10.7*	10.9	1.	"	12.7	2.5	"
	6/71	7.4	3	"	.13	11.3*	15.6	2.	"	8.5	1.3	"
	10/71	6.8	5	"	.14	11.0*	7.1	1.	"	8.2	1.7	"
III	1/70	9.2	5	.0	.10	11.3*	---	---	---	10.5	1.8	.0
	3/70	8.8	6	"	"	11.2*	11.8	1.3	.0	10.5	1.9	"
	6/70	9.0	7	"	"	10.9*	15.6	1.7	"	11.1	1.9	"
	10/70	10.3	13	"	"	10.3*	8.6	.8	"	12.8	2.7	"
	1/71	8.0	7	"	"	"	8.8	1.1	"	4.4	2.	"
	3/71	9.5	11	"	"	"	10.5	1.	"	12.0	2.4	"
	6/71	7.3	3	"	"	11.5*	21.9	2.9	"	8.5	1.2	"
	10/71	7.5	6	"	"	10.7*	9.7	1.2	"	9.3	1.8	"

* 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 6

Extreme Monthly Temperatures

Site	Normal			Exponential		Gamma			Extreme		
	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{Q}$	$\hat{\lambda}$	$\hat{Q}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{Q}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{Q}$
I	82.5	1.5	.0	.012	16.9*	-----	-----	-----	83.3	.98	.00
II	82.9	1.3	.0	.012	16.9*	-----	-----	-----	84.8	.88	.00
III	81.1	.8	.0	.012	16.9*	-----	-----	-----	81.7	.67	.00

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

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

TABLE 7

Extreme Monthly Winds

Site	Normal			Exponential		Gamma			Extreme		
	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{Q}$	$\hat{\lambda}$	$\hat{Q}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{Q}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{Q}$
I	11.1	17	.26	.09	13.6*	1.4	.13	.0	15.4	2.9	.9
II	11.2	11	.0	.08	14.1*	1.1	.10	.0	13.6	2.5	.6
III	10.3	9	.0	.09	14.5*	1.6	.16	.0	13.0	2.2	.4

* null hypothesis can be rejected at α = .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 satisfactorly 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 similiar 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

Data Set**	Normal			Exp		Gamma			Extreme			Haggard et. al. Result	
	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{Q}$	$\hat{\lambda}$	$\hat{Q}$	$\hat{\gamma}$	$\hat{\beta}^{-1}$	$\hat{Q}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{Q}$	$\hat{\gamma}$	$\hat{\beta}^{-1}$
A	7.29	50.3	1.75	.14	5.14*	1.9	3.85	.14	16.3	4.4	.04	2.2	3.33
B	8.08	53.5	1.24	.12	4.70*	2.2	3.85	.09	16.6	9.6	.03	2.8	2.88
C	9.37	55.6	.42	.10	3.40*	1.9	5.07	.00	15.9	5.2	.05	1.9	4.73
D	10.18	55.3	.32	.09	3.90*	2.2	4.56	.00	16.6	5.2	.03	2.6	3.87
A'	7.18	39.7	1.23	.13	5.05*	2.1	3.4	.06	14.2	4.0	.02	2.2	3.1
B'	7.94	41.8	.86	.12	4.78*	2.4	3.4	.04	14.7	4.2	.02	2.9	2.6
C'	9.2	47.9	.26	.10	3.73*	1.9	4.8	.02	14.5	4.9	.04	2.0	4.5
D'	10.0	46.5	.18	.09	4.24*	2.3	4.3	.00	15.2	4.9	.02	2.7	3.6

* null hypothesis can be rejected at α = .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

Precipitation in inches	Data Sets**							
	A		B		C		D	
	I*	II	I	II	I	II	I	II
1	.976	.966	.992	.980	.980	.980	.994	.987
2	.909	.890	.954	.926	.926	.930	.968	.950
3	.817	.797	.887	.850	.850	.863	.923	.894
4	.716	.698	.801	.764	.764	.788	.826	.827
5	.615	.602	.705	.674	.674	.705	.792	.754
6	.519	.513	.607	.587	.587	.632	.716	.680
7	.433	.432	.513	.505	.505	.559	.639	.607
8	.357	.362	.427	.431	.431	.490	.565	.537
9	.292	.301	.351	.364	.364	.427	.494	.471
10	.237	.248	.286	.306	.306	.371	.429	.412
15	.077	.090	.091	.118	.118	.172	.191	.195
20	.023	.030	.024	.042	.042	.075	.077	.085
25	.006	.009	.006	.014	.014	.031	.029	.036
30	.002	.003	.001	.005	.005	.013	.010	.014
	A'		B'		C'		D'	
	I	II	I	II	I	II	I	II
1	.978	.972	.993	.985	.981	.977	.995	.980
2	.913	.900	.959	.934	.932	.924	.971	.956
3	.821	.806	.893	.858	.865	.855	.926	.903
4	.717	.704	.806	.768	.789	.779	.866	.838
5	.613	.603	.706	.673	.710	.700	.794	.765
6	.515	.510	.605	.580	.631	.623	.717	.690
7	.427	.425	.507	.492	.556	.500	.639	.615
8	.349	.352	.418	.413	.486	.482	.562	.544
9	.283	.288	.340	.344	.423	.419	.489	.477
10	.227	.235	.273	.284	.365	.364	.422	.416
15	.070	.078	.079	.098	.165	.169	.182	.192
20	.019	.023	.020	.031	.070	.073	.070	.082
25	.005	.006	.002	.004	.028	.031	.025	.033
30	.001	.002	.001	.003	.011	.015	.008	.013

* I- 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

Dahiya, R.C. and Gurland, J. (1970a). A test of fit for continuous distributions based on generalized minimum chi-square. Statistical Papers in Honor of George W. Snedecor, T.A. Bancroft, editor.

Dahiya, R.C. and Gurland, J. (1970b). Estimating the parameters of a gamma distribution. MRC-TR #1067.

Dahiya, R.C. and Gurland, J. (1972). Goodness of fit tests for the gamma and exponential distributions. Technometrics, vol. 14, no. 3, pp 791-801.

Appendix A

User setup for Gurland's (GMCS) procedure

JOB CONTROL PARAMETERS

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

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*y*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.
GREXPO	-	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

```
                IMPLICIT REAL*8   (A-H , O-Z)
                DIMENSION  RAW(8),CUML(8),CENRL(8),G(4,4),X(1000),IDIST(10),
         $             NFORMT(20),XJ1(4,4),EXOB(1000)
                DIMENSION LINE(33)
                COMMON /MOMENT/ RAW,CUML,CENRL,G,NOB
                COMMON /NUMBER/ XDIV,XMEAN,XVAR,XGEOM,IUNIT,ICOR,PI,STD
         9000   READ(5,1,END=9999) IUNIT,NOB,(IDIST(I),I=1,5)
         1      FORMAT (9I5)
                READ(5,2) (NFORMT(I),I=1,20)
         2      FORMAT (20A4)
                IF(IDIST(1) .EQ. 5)  GO TO 9004
                READ(IUNIT,NFORMT) (X(J), J = 1,NOB)
                DO 152  I=1,NOB,12
                XMAX = X(I)
                XMIN = X(I)
                K = I+1
                L = I+11
                DO 151  J = K,L
                IF (X(J) .GT. XMAX)  XMAX=X(J)
                IF (X(J) .LT. XMIN) XMIN=X(J)
         151    CONTINUE
         152    WRITE(6,153) XMAX, XMIN
         153    FORMAT( T25,5F5.1)
                ICHECK = 0
                DO 111 J = 1,NOB
         111    IF( X(J) .LE. 0.)  ICHECK = 1
                WRITE(6,125)
                WRITE(6,123)  (X(J),J=1,NOB)
                XDIV  = DFLOAT(NOB)
                XMEAN = 0.0
                XVAR  = 0.0
                SVAR  = 0.0
                SUM = 0.0
                XM3 = 0.0
                XM4 = 0.0
                SM2 = 0.0
                SM3 = 0.0
                SM4 = 0.0
                PI    = 3.1415926
                SDIV  = XDIV - 1.
                DO 9001 I = 1,NOB
                XMEAN = XMEAN + X(I) / XDIV
                SUM = SUM + DABS(X(I))
                IF (SUM .LE. 0.)  SUM=0.1
                SD  = DLOG(SUM) / XDIV
                XGEOM = DEXP(SD)
         9001   CONTINUE
                DO 9002 I = 1,NOB
                XVAR = XVAR + ( X(I) - XMEAN )**2 / XDIV
                XM3    = XM3    + ( X(I) - XMEAN ) ** 3 / XDIV
                XM4    = XM4    + ( X(I) - XMEAN ) ** 4 / XDIV
                SM2 = SM2 + X(I)**2 / XDIV
                SM3 = SM3 + X(I)**3 / XDIV
                SM4 = SM4 + X(I)**4 / XDIV
                STD = DSQRT(XVAR)
         9002   CONTINUE
         C
         C
```

```
C          LOOP TILL ALL DISTRIBUTION REQUESTS HAVE BEEN SATISFIED
C
C
           DO 9003 I = 1,4
           IF  ((IDIST(I) .LE. 0) .OR. (IDIST(I) .GT. 4)) GO TO 9003
           IDUM = IDIST(I)
           GO TO ( 11,12,13,14), IDUM
C
C            NORMAL                IDIST = 1
C
      11   CALL GRNORM(XM3)
           GO TO 9003
C
C            EXPONENTIAL           IDIST = 2
C
      12   CALL GREXPO(SM2,SM3,SM4,X)
           GO TO 9003
C
C            GAMMA                 IDIST = 3
C
      13   IF( ICHECK .EQ. 0 )   CALL GRGAMM(X,SM2,SM3,SM4)
           WRITE (6,121)   ICHECK
     121   FORMAT( 10X, 25( I2,1X))
           GO TO 9003
C
C            EXTREME VALUE           IDIST = 4
C
      14   CALL GREXTR(X)
    9003   CONTINUE
           GO TO 9000
C
C            BIVARIATE NORMAL     IDIST(1) = 5
C
    9004   READ(IUNIT,NFORMT) ((X((J-1)*2+1), X((J-1)*2+2)), J=1,NOB)
           CALL BIVAR(X,NOB,IUNIT)
     123   FORMAT(T25, 5F12.5)
     125   FORMAT(1H1///1H0,T51,'THE OBSERVATIONS',//)
C
    9999   WRITE(6,25)
      25   FORMAT('1')
           REWIND 9
      10   READ(9,15,END=20)  LINE
      15   FORMAT(33A4)
           WRITE(6,15)  LINE
           GO TO  10
      20   STOP
           END
```

```
FORTRAN IV G LEVEL  21                  GCALC               DATE = 78192

              SUBROUTINE GCALC(ICOR)
        C
        C     CALCULATE G FOR FIRST FOUR DISTRIBUTIONS
        C
              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,G,NOB
              XN = DFLOAT(NOB)
              DO 100 I = 1,4
              DO 100 J = 1,4
              G(I,J) = RAW(I+J) - RAW(I)*RAW(J)
          100 CONTINUE
              RETURN
              END
```

```
FORTRAN IV G LEVEL  21                  RHAT1               DATE = 78192

              SUBROUTINE RHAT1(W,SIGI,R)
        C
        C     CALCULATE VECTOR R HAT FOR EXPONENTIAL DISTRIBUTION
        C
              IMPLICIT REAL*8  (A-H , O-Z)
              DIMENSION W(4),SIGI(4,4),R(4,4),DUM(4),X(1),FOUR(4,4)
              CALL DGMPRD(W,SIGI,DUM,1,4,4)
              CALL DGMPRD(DUM,W,X,1,4,1)
              X(1) = 1.0 / X(1)
              CALL DGMPRD(X,W,DUM,1,1,4)
              CALL DGMPRD(W,DUM,FOUR,4,1,4)
              CALL DGMPRD(FOUR,SIGI,R,4,4,4)
              RETURN
              END
```

```
FORTRAN IV G LEVEL  21                  RHAT2               DATE = 78192

              SUBROUTINE RHAT2(W,SIGI,R)
        C
        C     CALCULATE R HAT MATRIX(4X2) FOR GAMMA, NEG BIN, NORMAL
        C
              IMPLICIT REAL*8  (A-H , O-Z)
              DIMENSION  W(4,2),SIGI(4,4),R(4,4),WT(2,4),DUM(2,4),X(2,2),
             &           FOUR(4,4),M(2),L(2)
              DO 9000 I = 1,2
              DO 9000 J = 1,4
         9000 WT(I,J) = W(J,I)
              CALL DGMPRD(WT,SIGI,DUM,2,4,4)
              CALL DGMPRD(DUM,W,X,2,4,2)
              CALL DMINV(X,2,DET,L,M)
              CALL DGMPRD(X,WT,DUM,2,2,4)
              CALL DGMPRD(W,DUM,FOUR,4,2,4)
              CALL DGMPRD(FOUR,SIGI,R,4,4,4)
              RETURN
              END
```

```
      SUBROUTINE AHAT(SIGI,R,A)
C
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
```

FORTRAN IV G LEVEL 21 TRIPLE DATE = 78192

```
      SUBROUTINE TRIPLE(X,Y,Z)
C
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 I = 1,4
      DO 1 J = 1,4
    1 XT(I,J) = X(J,I)
      CALL DGMPRD(X,Y,DUM,4,4,4)
      CALL DGMPRD(DUM,XT,Z,4,4,4)
      RETURN
      END
```

FORTRAN IV G LEVEL 21 QHAT

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

ORIGINAL PAGE IS
OF P

```
0001                SUBROUTINE GREXTR(X)
           C
           C        GURLAND ROUTINE FOR EXTREME VALUE DISTRIBUTION
           C
0002                IMPLICIT REAL*8 (A-H,O-Z)
0003                DIMENSION   XJ1(4,4),RAW(8),CUML(8),CENRL(8),G(4,4),W(4),H(4),
                &            DUM(4,4),SIGI(4,4),L(4),M(4),THETA(4),R(4,4),A(4,4),
                &            X(1000),BHAT(100),TDENOM(100)
0004                COMMON /MOMENT/ RAW,CUML,CENRL,G,NOB
0005                COMMON /NUMBER/ XDIV,XMEAN,XVAR,XGEOM,IUNIT,ICOR,PI,STD
           C
0006                XN = DFLOAT(NOB)
0007                ZERO=0.0
0008                ONE=1.0
           C
           C        CALCULATE EXTREME CUMULANT MOMENTS
           C
           C        PUT CUML(I-1) IN PLACE OF CUML(I) IN ORDER TO MAKE THE SAME
           C        SUBSCRIPTS OF H-VECTOR AS THAT OF THE JACOBIAN MATRIX
           C
0009                ESUM = 0.0
0010                SIX=6.
0011                BETA = DSQRT(SIX) * STD / PI
0012                B = BETA
0013                DO 1  I = 1,NOB
0014       1        ESUM = ESUM + DEXP( X(I)/B)
0015                ALPHA = B * DLOG(ESUM) - B * DLOG(XDIV)
0016                EMEAN = ALPHA - 0.577216*B
0017                EMODE = ALPHA
0018                EVAR = PI ** 2 * B ** 2 / 6.
0019                CUML(1) = 1.645*B**2.
0020                CUML(2) = 2.396*B**3.
0021                CUML(3) = 6.494*B**4.
0022                CUML(4) = 24.886*B**5.
0023                CUML(5) = 122.078*B**6.
0024                CUML(6) = 726.01*B**7.
0025                CUML(7) = 5060.545*B**8.
           C
0026                C1 = CUML(1)
0027                C2 = CUML(2)
0028                C3 = CUML(3)
0029                C4 = CUML(4)
0030                C5 = CUML(5)
0031                C6 = CUML(6)
0032                C7 = CUML(7)
           C
0033                RAW(1) = XMEAN
0034                RAW(2) = C1 + XMEAN**2
0035                RAW(3) = C2 + 3.*C1*XMEAN + XMEAN**3
0036                RAW(4) = C3 + 4.*C2*XMEAN + 3.*C1**2  + 6.*C1*XMEAN**2
                &           + XMEAN**4
0037                RAW(5) = C4 + 5.*C3*XMEAN + 10.*C2*C1 + 10.*C2*XMEAN**2
                &           + 15.*C1**2 *XMEAN + 10.*C1*XMEAN**3  + XMEAN**5
           C
0038                RAW(6) = C5 + 6.*C4*XMEAN + 15.*C3*C1 + 15.*C3*XMEAN**2  +
                &           10.*C2**2  + 60.*C2*C1*XMEAN + 20.*C2*XMEAN**3
                &           + 15.*C1**3  + 45.*C1**2 *XMEAN**2  + 15.*C1*XMEAN**4  +
                &           XMEAN**6
```

ORIGINAL PAGE IS
OF POOR QUALITY

```
            C
0039              RAW(7) = C6 + 7.*C5*XMEAN + 21.*C4*C1 + 21.*C4*XMEAN**2
                 &         + 35.*C3*C2 + 105.*C3*C1*XMEAN + 35.*C3*XMEAN**3
                 &         + 70.*C2**2 *XMEAN + 105.*C2*C1**2  + 210.*C2
                 &         *C1*XMEAN**2  + 35.*C2*XMEAN**4  + 105.*C1**3 *XMEAN
                 &         + 105.*C1**2 *XMEAN**3  + 21.*C1*XMEAN**5  + XMEAN**7
            C
0040              RAW(8) = C7 + 8.*C6*XMEAN + 28.*C5*C1 + 28.*C5*XMEAN**2  +
                 &         56.*C4*C2 + 168.*C4*C1*XMEAN + 56.*C4*XMEAN**3  +
                 &         35.*C3**2  + 280.*C3*C2*XMEAN + 210.*C3*C1**2  +
                 &         420.*C3*C1*XMEAN**2  + 70.*C3*XMEAN**4  + 280.*C2**2 *C1
                 &         280.*C2**2 *XMEAN**2  + 840.*C2*C1**2 *XMEAN + 560.*C2*C
                 &         *XMEAN**3  + 56.*C2*XMEAN**5  + 105.*C1**4
                 &         + 420.*C1**3 *XMEAN**2  + 210.*C1**2 *XMEAN**4
                 &         + 28.*C1*XMEAN**6  + XMEAN**8
            C
0041              CALL  GCALC(ICOR)
            C
            C        INITIALIZE   W
            C
0042              W(1) = 1.456
0043              W(2) = 2.710
0044              W(3) = 3.850
0045              W(4) =4.906
            C
            C        INITIALIZE   H
            C
0046              H(1) = CUML(2)/CUML(1)
0047              H(2) = CUML(3)/CUML(2)
0048              H(3) = CUML(4)/CUML(3)
0049              H(4) = CUML(5)/CUML(4)
            C
            C        INITIALIZE   J1
            C
0050              DO 120 I=1,4
0051              DO 120 J=1,4
0052              IF(((I.EQ.J).OR.((I-1).EQ.J)) GO TO 120
0053              XJ1(I,J)=ZERO
0054        120   CONTINUE
0055              XJ1(1,1) = ONE
0056              XJ1(2,2) = 1./CUML(1)
0057              XJ1(2,1) = -CUML(2)/CUML(1)**2.
0058              XJ1(3,3) = 1./CUML(2)
0059              XJ1(4,3) = -CUML(4)/CUML(3)**2.
0060              XJ1(3,2) = -CUML(3)/CUML(2)**2.
0061              XJ1(4,4) = 1./CUML(3)
            C
            C        CALCULATE CHI-SQUARE TEST AND EXTREME PARAMETER
            C
0062              CALL TRIPLE(XJ1,G,SIGI)
0063              CALL DMINV(SIGI,4,DET,L,M)
0064              CALL RHAT1(W,SIGI,R)
0065              CALL AHAT(SIGI,R,A)
0066              CALL QHAT(XN,H,A,Q)
0067              CALL DGMPRD(R,H,THETA,4,4,1)
            C
0068              WRITE(6,125)
0069              WRITE(6,122)  (X(J),J=1,NOB)
```

```
                 WRITE(6,123)  EMEAN
                 WRITE(6,124)  STD
                 WRITE(6,126)  EMODE
                 WRITE(6,127)  EVAR
                 WRITE(6,128)  XVAR
                 WRITE(6,129)  XMEAN
                   WRITE(9,130) ALPHA,B,Q
                 WRITE (5,121)  ALPHA,B,Q
         121     FORMAT(//,T25,' PARAMETERS :  ALPHA= ',E15.5,10X,' BETA= ',E15.5,
                &       //,T39,' ***( CHI-SQUARE  VALUE )*** ',E15.5)
         122     FORMAT(T35, 5F10.5)
         123     FORMAT(///,T37,'THE MEAN OF THE EXTREME VALUES IS',F15.7,/)
         127     FORMAT(///,T37,'VARIANCE OF THE EXTREME VALUES IS',F15.7,/)
         124     FORMAT(///,T37,'THE STANDARD DEVIATION IS',8X,F15.7,/)
         125     FORMAT(1H1///1H0,T51,'THE OBSERVATIONS'/1H,
                &T41,'GURLANDS PROCEDURE FOR EXTREME VALUES',//)
         126     FORMAT(///,T37,'THE MODE OF THE EXTREME VALUES IS',F15.7,/)
         128     FORMAT(///,T37,'THE SAMPLE VARIANCE IS',11X,F15.7)
         C          INITIALIZE W
         129     FORMAT(///,T37,'THE SAMPLE MEAN IS',15X,F15.7)
         130     FORMAT(//,T25,' EXTREME PARAMETERS: ALPHA= ',F6.2,5X,' BETA= ',
                &       F6.2,//,T39,' ***( CHI-SQUARE VALUE )*** ',F10.3)
                 RETURN
                 END
```

```
0001              SUBROUTINE GREXPO(SM2,SM3,SM4,X)
            C
            C     GURLAND ROUTINE FOR EXPONENTIAL DISTRIBUTION
            C
0002              IMPLICIT REAL*8  (A-H , O-Z)
0003              DIMENSION  XJ1(4,4),RAW(8),CUML(8),CENRL(8),G(4,4),W(4),
                &           H(4),DUM(4,4),SIGI(4,4),L(4),M(4),THETA(4),
                &           R(4,4),A(4,4),X(1000)
0004              COMMON /NUMBER/ XDIV,XMEAN,XVAR,XGEOM,IUNIT,ICOR,PI,STD
0005              COMMON /MOMENT/ RAW,CUML,CENRL,G,NOB
0006              WRITE(6,1000)IUNIT
0007        1000  FORMAT(///,' EXPONENTIAL DISTRIBUTION   WITH DATA FROM UNIT ',
                &       I3,///
0008              XN = DFLOAT(NOB)
0009              ZERO = 0.0
0010              ONE = 1.0
            C
            C        CALCULATE EXPONENTIAL MOMENTS
            C
0011              RAW(1) = XMEAN
0012              RAW(2) = 2 * XMEAN**2
0013              RAW(3) = 6 * XMEAN**3
0014              RAW(4) = 24 * XMEAN**4
0015              RAW(5) = 120 * XMEAN**5
0016              RAW(6) = 720 * XMEAN**6
0017              RAW(7) = 5040 * XMEAN**7
0018              RAW(8) = 40320 * XMEAN**8
0019              CALL GCALC(ICOR)
            C
            C        INITIALIZE W
            C
0020              DO 9000 I = 1,4
0021        9000  W(I) = I
            C
            C        INITIALIZE H
            C
0022              H(1) = RAW(1)
0023              H(2) = SM2 / RAW(1)
0024              H(3) = SM3 / SM2
0025              H(4) = SM4 / SM:
            C
            C        INITIALIZE J1
            C
0026              DO 9001 I = 1,4
0027              DO 9001 J = 1,4
0028              IF ( (I .EQ. J) .OR. ((I-1) .EQ. J) ) GO TO 9001
0029              XJ1(I,J) = ZERO
0030        9001  CONTINUE
0031              XJ1(1,1) = ONE
0032              XJ1(2,2) = 1.0 / RAW(1)
0033              XJ1(3,3) = 1.0 / RAW(2)
0034              XJ1(4,4) = 1.0 / RAW(3)
0035              XJ1(2,1) = -RAW(2) / RAW(1)**2
0036              XJ1(3,2) = -RAW(3) / RAW(2)**2
0037              XJ1(4,3) = -RAW(4) / RAW(3)**2
            C
            C        CALCULATE CHI-SQUARE TEST AND EXPONENTIAL PARAMETER
            C
```

ORIGINAL PAGE IS OF POOR QUALITY

```
      SUBROUTINE GRGAMM(X,SM2,SM3,SM4)
C
C     GURLAND ROUTINE FOR GAMMA DISTRIBUTION
C
      IMPLICIT REAL*8 (A-H , O-Z)
      DIMENSION  XJ1(4,4),XJ2(4,4),RAW(8),CUML(8),CENRL(8),
     &           G(4,4),W(4,2),H(4),DUM(4,4),SIGI(4,4),L(4),M(4),
     &           THETA(4),R(4,4),A(4,4),X(1000),SCUML(100)
      COMMON /MOMENT/ RAW,CUML,CENRL,G,NOB
      COMMON /NUMBER/ XDIV,XMEAN,XVAR,XGEOM,IUNIT,ICOR,PI,STD
      WRITE(6,1000)IUNIT
 1000 FORMAT(///,' GAMMA DISTRIBUTION   WITH DATA FROM UNIT ',I3,///)
      XN = DFLOAT(NOB)
      ZERO = 0.0
      ONE = 1.0
      IF(XMEAN .LE. 0.)  WRITE(6,21)
   21 FORMAT(' ***NEGATIVE VALUES WRONG DISTRIBUTION*** ')
C
C        CALCULATE GAMMA MOMENTS
C
      AX = DABS(XMEAN)
      YY = DLOG10(AX / XGEOM)
      YYY = DABS(YY)
      T1 = .25D0 * (1.0 /YYY) * (1.0 + DSQRT(1. + 1.3333333333D0*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
      TLIMIT = 2. **(-251)
      XX = DFLOAT(I)
      IF ((T1 .GE. 57.57) .OR. (T1 .LE. TLIMIT))  GO TO 98
      RAW(I) = DGAMMA(T1 + XX) / DGAMMA(T1) * T2**XX
      GO TO  351
   98 RAW(I) = (T1+XX-1.) * 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.*SM2*RAW(1) + 2.*RAW(1)**3
      SCUML(4) = SM4 - 4.*SM3*RAW(1) - 3.*SM2**2 + 12.*SM2*RAW(1)**2
     &           -6.*RAW(1)**4
C
      CALL GCALC(ICOR)
C
C
      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
```

```
              C          INITIALIZE H
              C
0043                H(1) = SCUML(1)
0044                H(2) = SCUML(2) / SCUML(1)
0045                H(3) = SCUML(3) / SCUML(2)
0046                H(4) = SCUML(4) / SCUML(3)
              C
              C          INITIALIZE J1
              C
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(2,1) = -2 * RAW(1)
0054                XJ1(3,1) = -3*RAW(2) + 6*RAW(1)
0055                XJ1(4,1) = -4*RAW(3) + 12*RAW(3)*RAW(1) - 24*RAW(1)**3
0056                XJ1(3,2) = -3*RAW(1)
0057                XJ1(4,2) = -6*RAW(2) + 12*RAW(1)**2
0058                XJ1(4,3) = -4*RAW(1)
              C
              C          INITIALIZE J2
              C
0059                DO 101 I = 1,4
0060                DO 101 J = 1,4
0061                IF ( (I .EQ. J) .OR. ((I-1) .EQ. J) ) GO TO 101
0062                XJ2(I,J) = ZERO
0063          101   CONTINUE
0064                XJ2(1,1) = ONE
0065                XJ2(2,2) = 1.0 / CUML(1)
0066                XJ2(3,3) = 1.0 / CUML(2)
0067                XJ2(4,4) = 1.0 / CUML(3)
0068                XJ2(2,1) = -CUML(2) / CUML(1)**2
0069                XJ2(3,2) = -CUML(3) / CUML(2)**2
0070                XJ2(4,3) = -CUML(4) / CUML(3)**2
              C
              C          CALCULATE CHI-SQUARE TEST AND GAMMA PARAMETERS
              C
0071                CALL TRIPLE(XJ1,G,DUM)
0072                CALL TRIPLE(XJ2,DUM,SIGI)
0073                CALL DMINV(SIGI,4,DET,L,M)
0074                CALL RHAT2(W,SIGI,R)
0075                CALL AHAT(SIGI,R,A)
0076                CALL QHAT(XN,H,A,Q)
0077                CALL DGMPRD(R,H,THETA,4,4,1)
0078                XR = THETA(1) / THETA(2)
0079                XL = 1.0 / THETA(2)
0080                WRITE(6,123) XR,XL,Q
0081                WRITE(9,124) XR,XL,Q
0082          123   FORMAT(//,T25,' PARAMETERS :      R = ',E15.5,10X,' LAMDA=',E15.5,
                   &      //,T39,' ***( CHI-SQUARE  VALUE )*** ',E15.5)
0083          124   FORMAT(//,T25,' GAMMA PARAMETERS: R= ',F6.2,5X,' LAMDA= ',F6.2,
                   &      //,T39,' ***( CHI-SQUARE  VALUE )*** ',F10.3)
0084                RETURN
0085                END
```

```
0001             SUBROUTINE GRNORM (XM3,X)
           C
           C     GURLAND NORMAL DISTRIBUTION ROUTINE
           C
0002             IMPLICIT REAL*8  (A-H , O-Z)
0003             DIMENSION  XJ1(4,4),XJ2(4,4),RAW(8),CUML(8),CENRL(8),
          &                 G(4,4),W(4,2),H(4),DUM(4,4),SIGI(4,4),L(4),M(4),
          &                 THETA(4),R(4,4),A(4,4),X(1000)
0004             COMMON /NUMBER/ XDIV,XMEAN,XVAR,XGEOM,IUNIT,ICOR,PI,STD
0005             COMMON /MOMENT/ RAW,CUML,CENRL,G,NOB
0006             WRITE(6,1000)IUNIT
0007       1000  FORMAT(///,' NORMAL DISTRIBUTION   WITH DATA FROM UNIT ',I3,//)
0008             XN = DFLOAT(NOB)
0009             ZERO = 0.0
0010             ONE = 1.0
           C
           C        CALCULATE NORMAL MOMENTS
           C
0011             CENRL(1) = ZERO
0012             CENRL(2) = XVAR
0013             CENRL(3) = ZERO
0014             CENRL(4) = 3 * XVAR**2
0015             RAW(1) = XMEAN
0016             RAW(2) = XVAR + XMEAN**2
0017             RAW(3) = 3 * XMEAN * XVAR + XMEAN**3
0018             RAW(4) = 3 * XVAR**2 + 6 * XMEAN**2 * XVAR + XMEAN**4
0019             RAW(5) = 15 * XVAR**2 * XMEAN + 10 * XVAR * XMEAN**3 + XMEAN**5
0020             RAW(6) = 15*XVAR**3 + 45*XVAR*XMEAN**2 + 15*XVAR*XMEAN**4 +
          &               XMEAN**6
0021             RAW(7) = 105*XMEAN*XVAR**3 + 84*XVAR**2*XMEAN**3 +
          &               21*XVAR*XMEAN**5 + XMEAN**7
0022             RAW(8) = 105*XVAR**4 + 420*XVAR**3*XMEAN**2 +
          &               210*XVAR**2*XMEAN**4 + 28*XVAR*XMEAN**6 + XMEAN**8
0023             CALL GCALC(ICOR)
           C
           C        INITIALIZE W
           C
0024             W(1,1) = ONE
0025             W(2,1) = ZERO
0026             W(3,1) = ZERO
0027             W(4,1) = ZERO
0028             W(1,2) = ZERO
0029             W(2,2) = ONE
0030             W(3,2) = ZERO
0031             W(4,2) = 2.0
           C
           C        INITIALIZE H
           C
0032             H(1) = XMEAN
0033             H(2) = DLOG  (CENRL(2))
0034             H(3) = XM3
0035             H(4) = DLOG (CENRL(4) / 3.0)
           C
           C        INITIALIZE J1
           C
0036             DO 100 I = 1,4
0037             DO 100 J = 1,4
0038             IF (I .GT. J) GO TO 100
```

```
0039              IF (I .EQ. J) XJ1(I,J) = ONE
0040              IF (I .NE. J) XJ1(I,J) = ZERO
0041      100     CONTINUE
0042              XJ1(2,1) = -2 * RAW(1)
0043              XJ1(3,1) = -3*RAW(2) + 6*RAW(1)**2
0044              XJ1(3,2) = -3 * RAW(1)
0045              XJ1(4,1) = -4*RAW(3) + 12*RAW(2)*RAW(1) - 12*RAW(1)**3
0046              XJ1(4,2) = 5*RAW(1)**2
0047              XJ1(4,3) = -4 * RAW(1)
          C
          C          INITIALIZE J2
          C
0048              DO 101 I = 1,4
0049              DO 101 J = 1,4
0050              IF (I .EQ. J) GO TO 101
0051              XJ2(I,J) = ZERO
0052      101     CONTINUE
0053              XJ2(1,1) = ONE
0054              IF (XN .EQ. 1.)   GO TO 8
0055      7       XJ2(2,2) = 1. / (CENRL(2) * XN / (XN - 1.0) )
0056              XJ2(3,3) = ONE
0057              XJ2(4,4) = XJ2(2,2)**2 / 3.0
0058              GO TO 9
          C
          C          CALCULATE CHI-SQUARE TEST AND PARAMETERS
          C
0059      8       XN = XN + 1.
0060              GO TO 7
0061      9       CALL TRIPLE(XJ1,G,DUM)
0062              CALL TRIPLE(XJ2,DUM,SIGI)
0063              CALL DMINV(SIGI,4,DET,L,M)
0064              CALL RHAT2(W,SIGI,R)
0065              CALL AHAT(SIGI,R,A)
0066              CALL QHAT(XN,H,A,Q)
0067              CALL DGMPRD(R,H,THETA,4,4,1)
0068              TVAR = DEXP(THETA(2))
0069              WRITE(6,123) THETA(1),TVAR,Q
0070              WRITE(9,124) THETA(1), TVAR, Q
0071      123     FORMAT(//,T25,' PARAMETERS :   MJ    = ',E15.5,10X,' SIGMA=',E15.5
                &         //,T39,' ***( CHI-SQUARE  VALUE )*** ',E15.5)
0072      124     FORMAT(//,T25,' NORMAL PARAMETERS: MU= ',F6.2,5X,' SIGMA= ',F6.2
                &         //,T39,' ***( CHI-SQUARE  VALUE )*** ',F10.3)
0073              RETURN
0074              END
```

```
0038          CALL TRIPLE(XJ1,G,SIGI)
0039          CALL DMINV(SIGI,4,DET,L,M)
0040          CALL RHAT1(W,SIGI,R)
0041          CALL AHAT(SIGI,R,A)
0042          CALL QHAT(XN,H,A,Q)
0043          CALL DGMPRD(R,H,THETA,4,4,1)
0044          XLAMDA = 1. / THETA(1)
0045          WRITE(6,123) XLAMDA,Q
0046          WRITE(9,124) XLAMDA,Q
0047      123 FORMAT(//,T25,' PARAMETERS :   LAMDA = ',E15.5 ,/
             &      //,T39,' ***( CHI-SQUARE  VALUE )*** ',E15.5)
0048      124 FORMAT(//,T25,' EXPONENTIAL PARAMETERS: LAMDA= ',F6.2,/
             &      //,T39,' ***( CHI-SQUARE  VALUE )*** ',F10.3)
0049          RETURN
0050          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 struction in the LANDSAT data is similiar 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 μ and common positive definite covariance matrix Σ. Suppose that $X = [X_1, X_2 \ldots, X_n]^T$ and that

$$E[(X - E(X))(X - E(X)^T] = \Gamma_n \otimes \Sigma \tag{1}$$

where Γ_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 Γ_n denotes the autocorrelation matrix for n lags.
That is,

$$\begin{aligned} \Gamma_n &= (\rho_{ij}) \quad i,j = 1,2\ldots,n \\ \rho_{ij} &= \mathrm{corr}(X_i, X_j). \end{aligned} \tag{2}$$

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

$$U^*\Gamma_n U \simeq 2\Pi\, D_X \tag{3}$$

where

$$\begin{aligned} D_X &= \mathrm{diag}\,(d_1, d_2, \ldots, d_n) \\ d_1 &= f_X(0) \\ d_n &= f_X(\Pi) \\ d_{2k} &= d_{2k+1} = f_X\left(\frac{2\Pi k}{n}\right); \quad k = 1,2,\ldots,(n-1)/2. \end{aligned}$$

and

$$n^{\frac{1}{2}}2^{-\frac{1}{2}}U^* = \begin{bmatrix} 2^{-\frac{1}{2}} & 2^{-\frac{1}{2}} & \cdots & 2^{-\frac{1}{2}} \\ 1 & \cos(2\Pi/n) & \cdots & \cos(2\Pi \frac{n-1}{n}) \\ 0 & \sin(2\Pi/n) & \cdots & \sin(2\Pi \frac{n-1}{n}) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 1 & \cos(\frac{n-1}{n} 2\Pi/n) & \cdots & \cos(\frac{n-1}{n} 2\Pi \frac{n-1}{n}) \\ 0 & \sin(\frac{n-1}{n} 2\Pi/n) & \cdots & \sin(\frac{n-1}{n} 2\Pi \frac{n-1}{n}) \end{bmatrix} \tag{4}$$

By letting

$$Z = U^*X \tag{5}$$

it follows that

$$E[\,(Z - E(Z))\,(Z - E(Z))^T] = D_x \otimes \Sigma. \tag{6}$$

Furthermore, it follows that

$$Z_1 = n^{\frac{1}{2}}\,\overline{X}; \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{7}$$

where $Z = [Z_1 \ldots Z_n]^T$. The distribution for Z_j is

$$\begin{aligned} Z_1 &\sim N(n^{\frac{1}{2}}\mu, d_1\Sigma) \\ Z_j &\sim N(\emptyset, d_j\Sigma); \quad j = 2,3,\ldots,n \end{aligned} \tag{8}$$

where the symbol $\sim$ means "is distributed as". The expectation of Z_j = zero since

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

or

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

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

$$= 0.$$

Now let

$$\begin{aligned} Q_1(\mu) &= n(\overline{X} - \mu)^T \Sigma^{-1} (\overline{X} - \mu) \\ Q_2 &= \sum_{j=1}^{n} (X_j - \overline{X})^T \Sigma^{-1}(X_j - \overline{X}). \end{aligned} \tag{10}$$

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

$$\begin{aligned} Q_1(\mu) &\sim \chi^2(p) \\ Q_2 &\sim \chi^2(n-1)p \end{aligned} \tag{11}$$

where $\chi^2(v)$ denotes a chi-square distribution with v degrees of freedom.

However, if Γ_n is given by (1) we have

$$\begin{aligned} Q_1(\mu) &= n(\overline{X}-\mu)^T \Sigma^{-1} (\overline{X}-\mu) \\ &= (n^{\frac{1}{2}}\overline{X}-n^{\frac{1}{2}}\mu)^T \Sigma^{-1} (n^{\frac{1}{2}}\overline{X}-n^{\frac{1}{2}}\mu) \\ &= (Z_1-E(Z_1))^T \Sigma^{-1} (Z_1-E(Z_1)) \\ &= d_1(Z_1-E(Z_1))^T (d_1\Sigma)^{-1} (Z_1-E(Z_1)). \end{aligned} \tag{12}$$

Hence

$$Q_1(\mu)/d_1 \sim \chi^2(p). \tag{13}$$

Now consider $Q_2 = \sum_{j=1}^{n} (X_j-\overline{X})^T \Sigma^{-1} (X_j-\overline{X})$

$$= \operatorname{tr} \Sigma^{-1}[\sum_{j=1}^{n} (X_j-\overline{X})(X_j-\overline{X})^T]$$

$$= \operatorname{tr} \Sigma^{-1}[\sum_{j=1}^{n} X_jX_j^T - n\overline{X}\overline{X}^T] \tag{14}$$

However, since U is orthogonal (14) becomes

$$\begin{aligned} Q_2 &= \operatorname{tr} \Sigma^{-1}[\sum_{j=1}^{n} Z_jZ_j^T - n\overline{X}\overline{X}^T] \\ &= \operatorname{tr} \Sigma^{-1}[\sum_{j=2}^{n} Z_jZ_j^T] \\ &= \sum_{j=2}^{n} Z_j^T \Sigma^{-1} Z_j \\ &= \sum_{j=2}^{n} d_jW_j \end{aligned} \tag{15}$$

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 μ. That is, let

$$I_\varepsilon = \{\mu:\ Q_1(\mu) \leq \chi_\varepsilon^2(p)\} \tag{16}$$

where $\chi_\varepsilon^2(p)$ is the 100 ε 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 α such that

$$P[\mu \in I_\varepsilon \mid H_1 \text{ true}] = \alpha. \tag{18}$$

From equation (13), we know that α must satisfy the following relationship

$$\chi_\alpha^2(p) = \chi_\varepsilon^2(p)/d_1 \tag{19}$$

IV. Confidence Interval for the Dispersion Scalar

Let $X_1 \ldots X_n$ denote a sample from a normal distribution with mean μ and covariance matrix $\sigma^2\Sigma$, where Σ is a known positive definite matrix. Let H_o denote the hypothesis that the sample is random and H_1 denote the hypothesis that the sample satisfies equation (1). If H_o is true, then

$$Q_2/\sigma^2 \sim \chi^2_{p(n-1)} \tag{20}$$

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

$$0 \leq \sigma^2 \leq Q_2/\chi^2_{\varepsilon,p(n-1)} \tag{21}$$

is a 100 ε confidence interval for σ^2. However, to find the confidence interval for σ^2 when H_1 is true, it is necessary to determine the distribution of Q_2. From equation (15) we obtain

$$Q_2/\sigma^2 = \sum_{j=2}^{n} d_j W_j \tag{22}$$

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)].

$$P[Q_2/\sigma^2 \leq y] = \sum_{k=0}^{\infty} c_k\, G(v + 2k;\, y/\beta) \qquad (23)$$

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, β are known functions of the d_j 's, for $j=2,3,\ldots,n$. Hence, whenever H_1 is true, the confidence interval for σ^2 in equation (21) is given by α where α is the value which satisfied the following relationship

$$\alpha = \sum_{k=0}^{\infty} c_k\, G(p(n-1) + 2k;\, y_\varepsilon/\beta). \qquad (24)$$

where

$$y_\varepsilon = \chi^2_{\varepsilon,p(n-1)}.$$

V. Examples

Suppose that $X_1 \ldots X_n$ are a realization from a stationary autoregressive process of order one with parameter ϕ. Then the spectral density function is

$$f_X(w) = \frac{1}{2\,\Pi\,(1+\phi^2-2\phi\cos w)} \qquad (25)$$

Hence

$$\begin{aligned} d_{2k} &= (1+\phi^2-2\phi\cos(2k\Pi/n))^{-1} \quad k=1,],\ldots,n-1/2 \\ d_1 &= (1-\phi)^{-2} \end{aligned} \qquad (26)$$

The α-values which satisfy equation (19) are given in Table 1 for $\varepsilon = .99, 95$.

TABLE 1

α-Values for AR(1) Process

$P\backslash\phi$	.0	.1	.2	.3	.4	.5
1	.9900	.9795	.9606	.9285	.8776	.8021
	.9500	.9222	.8830	.8298	.7603	.6728
2	.9900	.9760	.9475	.8953	.8094	.6838
	.9500	.9116	.8529	.7695	.6598	.5270
5	.9900	.9681	.9145	.8071	.6346	.4174
	.9500	.8896	.2856	.6337	.4485	.2642
10	.9900	.9570	.8623	.6704	.4055	.1682
	.9500	.8614	.6952	.4648	.2363	.0823

From Table 1, we observe that a 95% confidence elipse is a 65.98% confidence elipse if the sample $X_1 \ldots X_n$ is a bivariate sample from an autoregressive process of order 1 with parameter $\phi = .4$

TABLE 2

α-Values for AR(1) Process

N	p\ϕ	.0	.1	.2	.3	.4	.5	.8
13	1	.9500	.9326	.8759	.7901	.6913	.5938	.3896
	2	.9143*	.8817	.7768	.6317	.4822	.3539	.1518
	5	.9144*	.8211	.5822	.3365	.1666	.0754	.0089
25	1	.9143*	.8742	.7577	.5996	.4386	.3020	.0902
	2	1.0000*	.8935	.6452	.3869	.1998	.0927	.0081
	5	1.0000*	.7547	.3344	.0934	.0178	.0026	.0000
51	1	1.0000*	.8859	.6223	.3550	.1702	.0712	.0036
	2	1.0000*	.7850	.3872	.1260	.0286	.0050	.0000
	5	1.0000*	.5460	.0933	.0056	.0001	.0000	.0000
101	1	1.0000*	.7822	.3811	.1209	.0266	.0043	.0000
	2	1.0000*	.6123	.1453	.0146	.0007	.0000	.0000
	5	1.0000*	.2932	.0080	.0000	.0000	.0000	.0000

* the specified level $\epsilon = .9500$

From Table 2, a 99% confidence interval for σ^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

1. Basu, R.,Basu,J.P.,and Lewis,T.O.(1976). Effect of intraclass correlation on confidence coefficients of confidence sets based on chi-square statistics. IEEE Trans. on Systems,Man,and Cybernetics June 1976,p.445-448.

2. Fuller, W.A. (1976) Introduction to Statistical Time Series. Wiley & Sons.

3. Kotz,S.,Johnson,N.L.,and Boyd,D.W. (1967). Series representations of distributions of quadratic forms in normal variables; central case. Annals of Math. Stat. vol 38,p.823-837.

4. Tubbs,J.D. and Coberly,W.A. (1978). Spatial correlation and its effect upon classification results in landsat. Proceedings of the 12th International Symposium on Remote Sensing of Environment, Manila, Philippines.

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 explantion 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,

$$\begin{aligned} x &= (-2 \ln u)^{1/2} \cos 2\pi v \\ y &= (-2 \ln v)^{1/2} \sin 2\pi v \end{aligned} \qquad (1)$$

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 σ, σ is a fixed known constant, then $y_i = \sigma x_i + \mu$, i=1,2,...,n is a sequence of independent normal with mean = μ, variance = σ^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 $\emptyset$ and covariance matrix I_p (pxp identity matrix). However, if $x \sim N_p(\emptyset, I_p)$ then $y = Bx + \mu$ has a multivariate normal distribution with mean$= \mu$ and covariance matrix Σ, where $\Sigma = BB^T$. From x we can find y for any specified real positive definite symmetric matrix Σ. This follows from the following result.

Result: 2.2 Let Σ 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 Σ.

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

$$x = -1/\lambda \ \ln \prod_{i=1}^{k} u_i \tag{2}$$

is a gamma with parameters λ and k. Note the chi-square distribution with n degrees of freedom can be obtained by letting $k=n/2$ and $\lambda=½$. Also, if n is odd then $y = x+w^2$ is chi-square with d.f.=n if $x \sim \Gamma(k = n-½, \lambda=½)$ with $w \sim N(0,1)$. The exponential distribution with parameter λ 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 $F_x(x-1) \leq u < F_x(x)$ 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 $P(\lambda)$

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

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

2.6 Negative Binomial Distribution NB(p,N)

The negative binomial distribution can be generated from a mixture of a Poisson and a Gamma distribution. That is, let $\underline{X}$ be distributed as a Poisson with parameter θ, where θ is a random variable from a Gamma distribution with parameters λ, R. Then $\underline{X}$ 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 (μ, Σ, A)

Let $Z_o, Z_i, .., Z_N$ be a sequence of N+1 p-dimensional independent multivariate normals with common null mean vector $\emptyset$ and pxp covariance matrix Σ. Then

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

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

$$\underset{n\times n}{A} \otimes \underset{p\times p}{\Sigma} = \underset{(np \times np)}{\begin{bmatrix} a_{11}\Sigma & a_{12}\Sigma & \cdots & a_{1n}\Sigma \\ a_{21}\Sigma & a_{22}\Sigma & \cdots & a_2 n\Sigma \\ \vdots & & & \\ a_{n1}\Sigma & \cdots & \cdots & a_{nn}\Sigma \end{bmatrix}}$$

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

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

From the dispersion matrix $A \otimes \Sigma$ we have that

$$\mathrm{COV}\,(x_i, x_j) = \alpha_i \alpha_j \Sigma \qquad i \neq j$$
$$= \Sigma \qquad i = j$$

Hence the correlation matrix between vector X_i, X_j is

$$\mathrm{CORR}(X_i, X_j) = \alpha_i \alpha_j \qquad i \neq j$$
$$= I_p \qquad i = j$$

where I_p is a pxp identity matrix. When p is 1 we have the univariate case.

3.2 Correlated Univariate Gamma Distribution $\Gamma(\lambda, R, A)$

Let $Z_o, Z_1, \ldots. Z_n$ denote a sequence of independent variables having the following Gamma distributions

$$Z_o \sim \Gamma(\lambda, R_o)$$
$$Z_i \sim \Gamma(\lambda, R_i - R_o)$$

Let $X_i = Z_o + Z_i, i = 1, 2, \ldots, n$, 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

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

where a_{ij} is the ij^{th} element of the nxn matrix A and

$$a_{ij} = \begin{cases} 1 & \text{if } i = j \\ \left(\dfrac{R_o^2}{R_i R_j} \right)^{1/2} & \text{if } i \neq j \end{cases}$$

3.3 Correlated Beta Distribution $\beta(p,q,A)$

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

$$X_i = Z_i \Big/ \left(\sum_{j=0}^{n} Z_j \right) \qquad i=1,2,\ldots,n$$

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

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

Discrete Distributions

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

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

$$X_i = Z_0 + Z_i$$

is a sequence of correlated Poissons with $X_i \sim P(\lambda_i)$ $\lambda_i = C_i + C_0, i=1,2,\ldots,n$ and the correlation between X_i X_j is given by

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

and

$$a_{ij} = \begin{cases} 1 & i=j \\ \left(\dfrac{C_o^2}{\lambda_i \lambda_j} \right)^{1/2} & i \neq j \end{cases} .$$

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

1. Box and Muller (1958). A note on the generation of random normal deviates, Amer. Math. Stat. 29, 610-11.

2. Falls, L.W. (1971). A Computer Program for Standard Statistical Distributions, NAS TM X-64588.

3. Johnson and Kotz (1972). Distribution in Statistics: Continuous Multivariate Distributions, Vol. 4, John Wiley & Sons, N.Y.

4. Newman and Odell (1971). The Generation of Random Variates, Hafner Publ. Co., N.Y.

APPENDIX A

JOB CONTROL PARAMETERS

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

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

CARD	COL	DESCRIPTION
2	1-5	NOB - Number of observations to be generated (I5)
	6-10	ITYPE - Type distribution to be generated (I5) 1 - Normal 4 - Poisson 2 - Gamma 5 - Negative Binomial 3 - Beta 6 - Binomial
	11	ICOR = 1 correlated data (I1) = 0 uncorrelated data (I1)
	12	ISTAT= 1 Print Statistics (I1) = 0 No print
	12-13	IUNIT= 0 Do not output generated data (I2) ≠ 0 Generated data output on external device # IUNIT

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

- NORMAL -

CARD	COL	DESCRIPTION
3	1-5	NV = Number of variates (NV=2=bivariate normal (I5)
	6-10	KEY= 0 Standardized normal mean = 0 variance = 1
		KEY =1 Read Mean, Variance (I5)

IF KEY = 1 Read following cards

CARD	COL	DESCRIPTION	
4	(16F5.0)	Y(I),I=1,.NV	Mean vector
5	(16F5.0)	S(I),I=1,NV**2	Covariance matrix

** 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 -

CARD	COL	DESCRIPTION		
3	1-5	P	parameter	(F5.0)
	6-10	N	parameter	(I5)

**IF ICOR = 1 Read following

4+	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	(16I5)	NOL(I),I=1,NG Length of each group NOL(1)+ NOL(2)+...+ NOL(NG) = NOB
3	(16F5.0)	VALUE (I),I=1,NG, A value for each group

example 1

NOB=25 NG=1 NOL(1)=25
VALUE(1)=.8
the $CORR(X_i,X_j)=(.8)x(.8) = .64$

CARD

1	ƀƀ1
2	ƀƀƀ25
3	ƀƀƀ.8

example 2

NOB=25 NG=2 NOL(1)=10 NOL(2)=15
VALUE (1)=.5 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	ƀƀ2
2	ƀƀƀ10ƀƀƀ15
3	ƀƀƀ.5ƀƀƀ.8

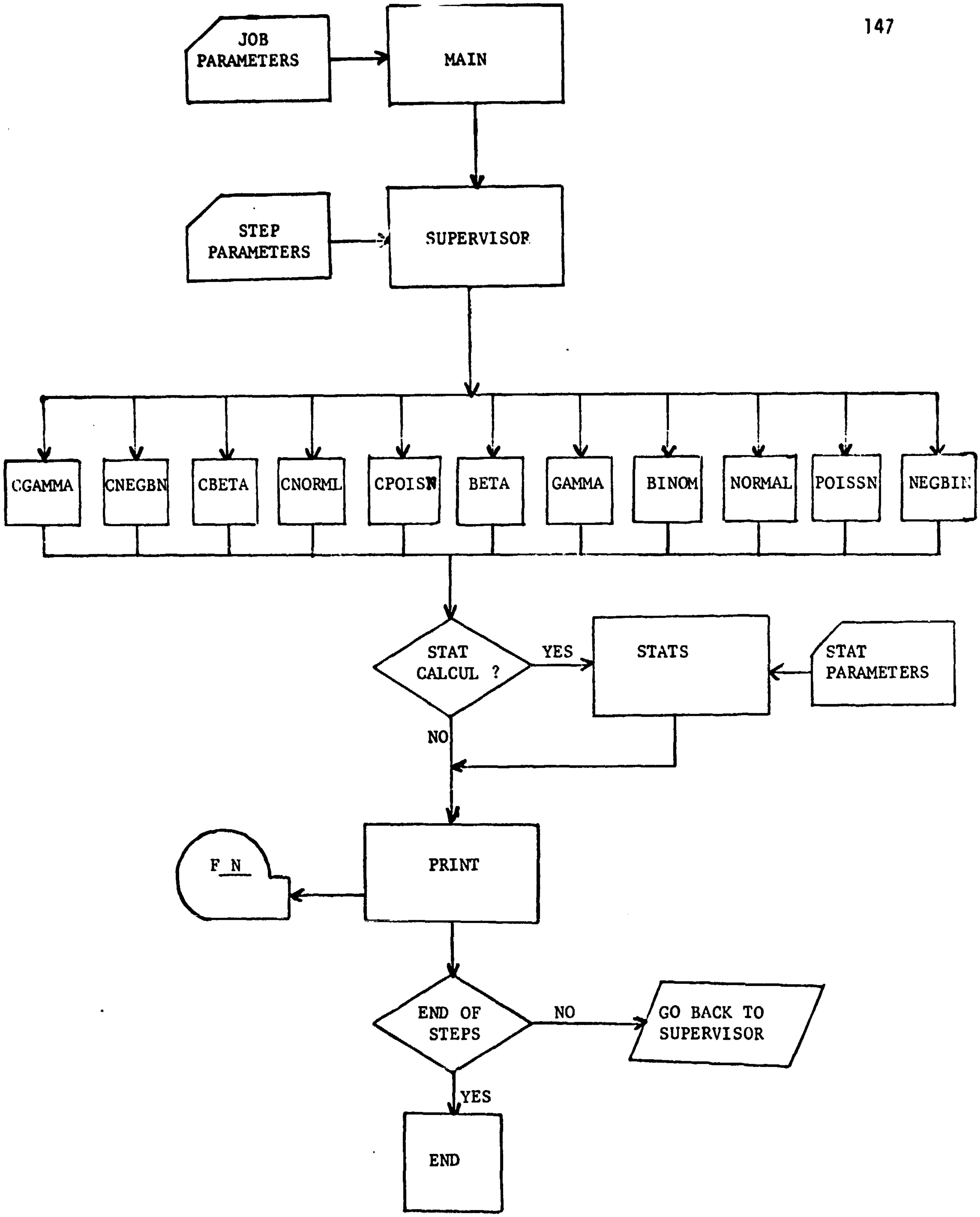
(Note: ƀ denote blank column)

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

CNORML	-	RANDA NORMAL
CNEGBN	-	GAMMA POISSN
CBETA	-	GAMMA
CGAMMA	-	GAMMA
CPOISN	-	POISSN
NORMAL	-	RANDU
BETA	-	GAMMA
GAMMA	-	RANDU NORMAL
BINOM	-	RANDU
GMETRC	-	RANDU
POISSN	-	GAMMA
NEGBIN	-	GMETRC
RANDA	-	RANDU
RANDU	-	
STATS	-	
PRINT	-	
MAIN	-	SUPER
SUPER	-	CNORML CBETA CGAMMA CNEGBN CPOISN NORMAL GAMMA BETA NEGBIN BINOM POISSN STATS PRINT


JOB PARAMETERS
MAIN
STEP PARAMETERS
SUPERVISOR
CGAMMA
CNEGBN
CBETA
CNORML
CPOISN
BETA
GAMMA
BINOM
NORMAL
POISSN
NEGBIN
STAT CALCUL ?
YES
NO
STATS
STAT PARAMETERS
PRINT
F N
END OF STEPS
GO BACK TO SUPERVISOR
END

```
      DIMENSION X(200),Y(100),S(100),Z(100)
      COMMON/B/ Y,S,Z
       COMMON/A/ IX,NV,R1,XLAMDA,R2,P,N
      IX=51487735
      READ(5,100) NREPS
      DO 99 II=1,NREPS
      READ(5,100) NO,ITYPE,IUNIT
      CALL TYPE(X,NO,ITYPE)
      CALL PRINT(X,NO,IUNIT,ITYPE,II)
   99 CONTINUE
  100 FORMAT(3I5)
      STOP
      END


      SUBROUTINE BETA(X,NO)
      DIMENSION X(200),Y(100),S(100),Z(100)
      COMMON/A/ IX,NV,R1,XLAMDA,R2,P,N
      COMMON/B/ Y,S,Z
      XLAMDA=1.
      CALL GAMMA(X,NO)
      R=R1
      R1=R2
      CALL GAMMA(Y,NO)
      R1=R
      DO 1 I=1,NO
      XX=X(I)/(X(I)+Y(I))
    1 X(I)=XX
       RETURN
      END


      SUBROUTINE GAMMA(X,NO)
      DIMENSION X(NO),Y(100)
      COMMON/B/ Y,S,Z
      COMMON/A/IX,NV,R1,XLAMDA,R2,P,N
       XL=-1.*(1./XLAMDA)
      K=R1+.5
      R=R1-K
      DO 99 II=1,NO
      XX=1.
      DO 1 I=1,K
      IN=IX-(IX/10)*10+1
      DO 2 J=1,IN
    2 CALL RANDU(IX,IX,YFL)
    1 XX=XX*YFL
   99 X(II)=XL*ALOG(XX)
      IF(R.LE.0) RETURN
      NV=1
      CALL NORMAL(Y,NO,NO,0)
      DO 98 I=1,NO
   98 X(I)=X(I)+Y(I)*Y(I)
      RETURN
      END
```

ORIGINAL PAGE IS OF POOR QUALITY

```
      SUBROUTINE BINOM(X,NO)
      DIMENSION X(NO)
      COMMON/A/ IX,NV,R1,XLAMDA,R2,P,N
      DO 1 I=1,NO
      IN=IX-(IX/10)*10+1
      DO 2 K=1,IN
    2 CALL RANDU(IX,IX,YFL)
    1 X(I)=YFL
      RETURN
      END
```

```
      SUBROUTINE NORMAL(X,NO,NVO,KEY)
      DIMENSION X(200),Y(100),S(100),Z(100)
      COMMON/B/Y,S,Z
      COMMON/A/IX,NV,R1,XLAMDA,R2
     *   ,P,M
      NVV=NV*NV
      P2=6.283185
C     GENERATE NVO INDEPENDENT UNIFORM (0,1)
      DO 2 I=1,NVO
      IN=IX-(IX/10)*10+1
      DO 3 K=1,IN
    3 CALL RANDU(IX,IX,YFL)
    2 X(I)=YFL
C     TRANSFORM TO NVO NORMAL (0,1)
      J=NVO/2.
      DO 4 I=1,J
      A=SQRT(-2.*ALOG(X(2* I-1)))
      X(2*I-1)=A*COS(P2*X(I*2))
    4 X(2*I)=A*SIN(P2*X(I*2))
      IF(KEY.EQ.0) RETURN
      WRITE(6,200)
  200 FORMAT(' MEAN AND COVARIANCE',//)
      READ(5,100) (Y(I),I=1,NV)
      WRITE(6,201) (Y(I),I=1,NV)
      READ(5,100)(S(I),I=1,NVV)
      WRITE(6,201) (S(I),I=1,NVV)
  201 FORMAT(10X,10F10.3)
      IF(NV.GT.1) GO TO 6
      DO 5 I=1,NO
    5 X(I)=S(1   )*X(I)+Y(1)
      RETURN
    6  N=NV
      DO 11 J=1,NV
      DO11 I=1,N
   11 Z((J-1)*N+I)=0.
      Z(1)=SQRT(S(1))
      DO13 I=2,N
   13 Z(I)=S(I)/Z(1)
      DO19 J=2,N
      DO19 I=1,N
      SUM=0.
      IF(I-J)19,15,17
   15 M=J-1
      DO16 K=1,M
   16 SUM=SUM+Z((K-1)*N+I)**2
      Z((J-1)*N+I)=SQRT(S((J-1)*N+I)-SUM)
      GO TO19
   17 M=J-1
      DO18 K=1,M
   18 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)
   19 CONTINUE
      DO 7 I=1,NO
      DO 7 K=1,NV
      S((I-1)*NV+K)=Y(K)
      DO 7 J=1,K
    7 S((I-1)*NV+K)=S((I-1)*NV+K)+Z((J-1)*NV+K)
     *    *X(NV*I-(NV-J))
      DO 8 I=1,NO
      DO 8 J=1,NV
    8 X((J-1)*NO+I)=S((I-1)*NV+J)
      RETURN
  100 FORMAT(16F5.0)
      END
```

```
      SUBROUTINE GMETRC(XX)
      COMMON/A/ IX,NV,R1,XLAMDA,R2,P,N
      XX=0.
      P=.5
      ONE=1.
      Q=ONE-P
      DO 1 I=1,N
      CALL RANDU(IX,IX,U)
      SUM=0.
      J=0
      QQ=P
    3 J=J+1
      QQ=QQ*Q
      SUM=SUM+QQ
      IF(SUM-U)2,1,1
    2 IF(J.LT.10) GO TO 3
    1 XX=XX+J
      RETURN
      END
```

```
      SUBROUTINE POISSN(X,NO)
      DIMENSION X(NO)
      COMMON/A/IX,NV,R1,XLAMDA,R2,P,N
      NSUM=0
      R1=1.
    1 SUM=0.
      J=0
    2 CALL GAMMA(XX,1)
      J=J+1
      SUM=SUM+XX
      IF(SUM.LE.1.) GO TO 2
      X(NSUM)=J-1
      NSUM=NSUM+1
      IF(NSUM.LE.NO) GO TO 1
      RETURN
      END
```

```
      SUBROUTINE TYPE(X,NO,ITYPE)
      DIMENSION X(NO)
      COMMON/A/ IX,NV,R1,XLAMDA,R2,P,N
      GO TO (1,2,3,4,5,6),ITYPE
    1 READ(5,100) NV,KEY
      NVV=NV*NV
      NVO=NV*NO
      CALL NORMAL(X,NO,NVO,KEY)
      RETURN
    2 READ(5,101) R1,XLAMDA
      CALL GAMMA(X,NO)
      RETURN
    3 READ(5,101) R1,R2
      CALL BETA(X,NO)
      RETURN
    4 READ(5,101) XLAMDA
      CALL POISSN(X,NO)
      RETURN
    5 CALL BINOM(X,NO)
      RETURN
    6 READ(5,102) P,N
      CALL NEGBIN(X,NO)
      RETURN
  100 FORMAT(2I5)
  101 FORMAT(2F5.0)
  102 FORMAT(F5.0,I5)
      END
```

```
      SUBROUTINE NEGBIN(X,NO)
      DIMENSION X(NO)
      COMMON/A/ IX,NV,R1,XLAMDA,R2,P,N
      DO 1 I=1,NO
      CALL GMETRC(XX)
    1 X(I)=XX
      RETURN
      END
```

```
      SUBROUTINE RANDU(IX,IY,YFL)
      IY=IX*65539
      IF(IY)5,6,6
    5 IY=IY+2147483647+1
    6 YFL=IY
      YFL=YFL*.4656613E-9
      RETURN
      END
```

```
      SUBROUTINE PRINT    (X,NO,IU,IT,II)
      DIMENSION X(NO)
      COMMON/A/IX,NV,R1,XLAMDA,R2,P,N
      NVO=NV*NO
      GO TO(1,2,3,4,5,6),IT
    1 WRITE(6,100) II,NV
      WRITE(6,200)  (X(I),I=1,NVO)
      IF(IU.EQ.2) WRITE(9,300)(X(I),I=1,NVO)
      RETURN
    2 WRITE(6,101) II,R1,XLAMDA
      WRITE(6,200)X
      IF(IU.EQ.2) WRITE(9,300) X
      RETURN
    3 WRITE(6,102) II,R1,R2
      WRITE(6,200)X
      IF(IU.EQ.2) WRITE(9,300) X
      RETURN
    4 WRITE(6,103) II,XLAMDA
      WRITE(6,200)X
      IF(IU.EQ.2) WRITE(9,300) X
      RETURN
    5 WRITE(6,104) II
      WRITE(6,200)X
      IF(IU.EQ.2) WRITE(9,300) X
      RETURN
    6 WRITE(6,105) II,P,N
      WRITE(6,200)X
      IF(IU.EQ.2) WRITE(9,300) X
      RETURN
  100 FORMAT(//,' NORMAL DATA FOR REP.=',I6,//,' NO. OF VARIATES=',I6,
     *   //)
  101 FORMAT(//,' GAMMA DATA FOR REP.=',I6,//,' N=',F10.3,
     *   ' LAMDA=',F10.3,//)
  102 FORMAT(//,' BETA  DATA FOR REP.=',I6,//,' ALPHA=',F10.2,' BETA=',
     *   F10.2,//)
  103 FORMAT(//,' POISSN DATA FOR REP.=',I6,//,' LAMDA=',F10.2,//)
  104 FORMAT(//,' UNIFORM (0,1) DATA FOR INTEGRAL TRANSFORM REP=',I6,//)
  105 FORMAT(//,' NEG. BINOMIAL DATA FOR REP.=',I6,//,
     *   ' P=',F10.3,' N=',I10,//)
  200 FORMAT(10X,10F10.3)
  300 FORMAT(50F10.3)
      END
```