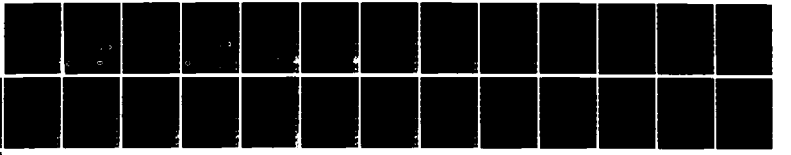
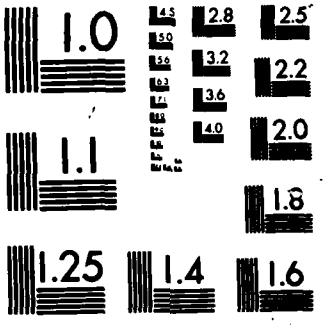
AD-A165 113 A METHOD FOR THE SIMULATION OF ENVIRONMENTAL DATA SETS 1/1
(U) OLD DOMINION UNIV NORFOLK VA APPLIED MARINE
RESEARCH LAB R W ALDEN JUL 84 DACW65-81-C-0051
UNCLASSIFIED F/G 9/2 NL



END
FILMED
DTIC



OLD DOMINION UNIVERSITY RESEARCH FOUNDATION

Applied Marine Research Laboratory
Old Dominion University
Norfolk, Virginia 23508

A METHOD FOR THE SIMULATION OF ENVIRONMENTAL DATA SETS

By

Raymond W. Alden III, Principal Investigator

Supplemental Contract Report

Prepared for the
Department of the Army
Norfolk District, Corps of Engineers
Fort Norfolk, 803 Front St.
Norfolk, Virginia 23510

Under
Contract DACW65-81-C-0051

US Army Corps Of Engineers
Norfolk District

July 1984

Report B-19

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release, distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) B-19
6a. NAME OF PERFORMING ORGANIZATION Old Dominion University, Applied Marine Research Laboratory 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION U.S. Army Corps of Engineers, Norfolk District
6c. ADDRESS (City, State, and ZIP Code) Norfolk, Virginia 23508	7b. ADDRESS (City, State, and ZIP Code) Norfolk, Virginia 23510-1096
8a. NAME OF FUNDING/SPONSORING ORGANIZATION U.S. Army Corps of Engineers, Norfolk District 8b. OFFICE SYMBOL (If applicable) NAOPL; NAOEN	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER DACW65-81-C-0051
8c. ADDRESS (City, State, and ZIP Code) Norfolk, Virginia 23510-1096	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
A Method for the Simulation of ~~Multivariate~~ ENVIRONMENTAL Data Sets

12. PERSONAL AUTHOR(S)
Alden, R.W., III

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM _____ TO _____	1984, July	17

16. SUPPLEMENTARY NOTATION

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Multivariate data analysis, data simulation, statistical analysis

19. ABSTRACT (Continue on reverse if necessary and identify by block number)
MDS method has proven to be a rapid, effective simulation technique. Aids in the evaluation of sampling/statistical regimes or in the interpretation of data sets from nature. A further use of MDS is in the evaluation of the effectiveness of environmental monitoring programs.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Craig L. Seltzer	22b. TELEPHONE (Include Area Code) (804) 441-3767/827-3767 22c. OFFICE SYMBOL NAOPL-R

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

Applied Marine Research Laboratory
Old Dominion University
Norfolk, Virginia 23508

A METHOD FOR THE SIMULATION OF ENVIRONMENTAL DATA SETS

By

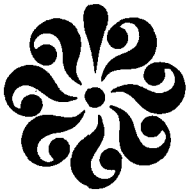
Raymond W. Alden III, Principal Investigator

Supplemental Contract Report

Prepared for the
Department of the Army
Norfolk District, Corps of Engineers
Fort Norfolk, 803 Front St.
Norfolk, Virginia 23510

Under
Contract DACW65-81-C-0051

Submitted by the
Old Dominion University Research Foundation
P.O. Box 6369
Norfolk, Virginia 23508



July 1984

TABLE OF CONTENTS

TITLE ON COVER CORRECT

A METHOD FOR THE SIMULATION OF ENVIRONMENTAL DATA SETS

By

Raymond W. Alden III

Director
Applied Marine Research Laboratory
Old Dominion University
Norfolk, Virginia 23508 U.S.A.

ABSTRACT

A method was developed which allows the simulation of multivariate data sets without requiring a characterization of the distributional "shapes" of each of the variables. The method is based upon the concept that most data sets can be approximately normalized by a family of power transformations. Conversely, a matrix of normal deviates produced by a random number generator can be adjusted to appropriate means and standard deviations and back-transformed to simulate the shape of the observed data. The method was successful in simulating data sets displaying a wide range of theoretical distributions as well as "real" data from an ongoing monitoring program.

Keywords: data simulation; multivariate data analysis

1. INTRODUCTION

Environmental scientists are increasingly being called upon to analyze and interpret large multivariate data sets. Sophisticated statistical computer packages are often employed to test significant patterns in the data. Unfortunately, most of these commonly available statistical techniques are based upon assumptions, such as the concent of multivariate normality which are seldom, if ever, met by data collected from nature. One analytical approach gaining popularity over the use of "cookbook" statistics is the utilization of simulated data sets to test the robustness, power, and sensitivity of various statistical models in the context of natural spatio-temporal variability prior to their application.

Data sets can be simulated through the use of packaged computer programs with random number generation functions (Raeside, 1976; and Green, 1979). Capra and Elster (1971) have demonstrted a method that uses a normal distribution random number generator to simulate data sets with desired means, variances and covariances. Most packaged computer programs today have random number generating functions based upon various families of theoretical distributions (e.g. poisson, binomial, negative binomial, gamma, exponential, etc.). Thus, non-normal variables can be simulated to have a wide range of distributions. Unfortunately, each of the observed variables must be empirically or mathematically evaluated in order to "fit" them with the most appropriate type of distribution. This selection process is often quite time-consuming if a large number of variables are to be simulated, if there is a diversity of distributions among the variables in the data set, or if a number of different data sets are

to be simulated.

The major goal of the present study was to develop a simulation method which could be applied by environmental scientists who may not have a strong background in distributional theory and, moreover, who may not have ready access to a mainframe computer system (e.g. an investigator working on a ship or at a field station). The study has resulted in the development of a method which simplifies the simulation of non-normal multivariate data sets. The method does not involve a preliminary evaluation and fitting of the distributions of the variables to be simulated, nor does it require random number generating functions which produce exotic families of non-normal distributions. As a result, it can be used on most microcomputers, as well as some of the more powerful programmable calculators. The new simulation method is referred to as the "MDS" method for "multivariate data simulation." The term "observed data" is used for the data to be matched with the simulation.

2. METHODS

2.1 General

The development of the MDS method was inspired by a technique presented by Green (1979). In order to simulate a variable with a skewed distribution, Green first employed a random number generator to produce a data set with a standardized normal distribution. Logarithmically transformed values of the desired mean and standard deviation were, respectively, added to and multiplied by the standardized normal deviates. The data set was then untransformed to produce a new variable with a skewed distribution. The concept of using a normal random number generator and the transformation/untransformation process is key to the MDS method.

Box and Cox (1964) introduced a family of power transformations which were designed to normalize data of wide range of distributions. The family of transformations are described by the relationship:

$$y^{(\lambda)} = (y^{\lambda}-1)/\lambda \text{ , if } \lambda \neq 0$$
$$= \log y \text{ , if } \lambda = 0 \qquad (1)$$

where y and $y^{(\lambda)}$ are the raw and transformed variates and λ is a transformation parameter which has been selected to best normalize the data. Box and Cox (1964) presented a maximized log likelihood process by which an optimum λ value can be determined for any given data set. This process is used in the MDS method to select a series of transformations which best normalize each of the variables in the "observed" data set prior to its simulation.

Each variable to be simulated is normalized by the selection transformation where the mean and standard deviation are calculated

for the transformed data. The mean value is added to each of a set of normal standard deviates produced by a random number generator, while the standard deviation value is multiplied by the deviates. The data are then untransformed to produce a distribution of the same type exhibited by the original variable.

2.2. The MDS Method

The MDS method has been incorporated into a computer package programmed in APL (Gilman and Rose, 1976) on a DEC System-10 computer. It can be readily adapted to other languages or computer systems. The data to be simulated are entered as a rxc matrix, where r = the number of cases and c = the number of variables. The process proceeds one variable at a time until the entire data set has been simulated. The basic steps in the procedure can be described as follows:

1. Transformation of the variable to normalize: In order to find the appropriate λ for the optimum power transformation (1), a modification of the maximized log likelihood method is employed. The log likelihood parameter $L_{max}(\lambda)$ is defined by:

$$L_{max}(\lambda) = -1/2\ n \log (S(\lambda;\underline{z})/n), \qquad (2)$$

where n = the number of replicates, and $S(\lambda;\underline{z})$ is the residual sum of squares of $\underline{z}(\lambda)$. The standardized variate $\underline{z}(\lambda)$ is defined by:

$$\underline{z}(\lambda) = (y^{\lambda} - 1)/\lambda \dot{y}^{\lambda-1} \qquad (3)$$

where $\dot{y}$ is the geometric mean of the original variable. The $S(\lambda;\underline{z})$ is calculated by:

$$S(\lambda;\underline{z}) = \Sigma\ (\underline{z}(\lambda) - \underline{\bar{z}}(\lambda))^2. \qquad (4)$$

An initial level of λ is chosen and the corresponding L_{max} value is calculated. Initial λ values of -10 have been shown empirically to be appropriate for most situations. The λ values are then increased incrementally and the L_{max} values are calculated until a maximum value is found. The current MDS computer program iteratively focuses on the L_{max} value until an optimum λ value is defined to two decimal places.

2. Statistical characterization of normalized observed data set: Once the optimum value for λ has been defined, the observed data is tranformed by (1). The mean ($\bar{y}(\lambda)$) and standard deviation ($S_{y(\lambda)}$) are calculated for the transformed data set.

3. Creation of data set of normal deviates: A random number generator is used to create a data set of appropriate size with a normal standard deviate distribution.

4. Adjustment of mean and standard deviation of simulated data: The $\bar{y}(\lambda)$ value is added to each of the values of the normal data set, while the $S_{y(\lambda)}$ value is multiplied by each of the standard deviates.

5. Back transformation of simulated data to the observed distributions: The new data set is then "back transformed," employing the relationship:

$$\begin{aligned} y &= \{(X\lambda)+1\}^{1/\lambda}, \text{ if } \lambda \neq 0 \\ y &= 10^{y}, \text{ if } \lambda \neq 0 \end{aligned} \qquad (5)$$

where X is the data set prior to back transformation and y represents the data set that simulates the distribution of the observed variable.

The program continues with cycles of steps 1-5 until all variables have been simulated. Recently, an option has been included

in the MDS computer program which allows the introduction of the observed autocorrelation/correlation patterns into the simulated data set. The multivariate structure is reproduced by using the APL "indexing" function to sort the values of each of the simulated variables into the same relative numeric order as is exhibited by the observed date. A second option that is available in the program allows the researcher to introduce "impacts" into the simulated data by multiplying the final y values by various factors (e.g. the values are multiplied by 0.5 to decrease them by half or by 2.0 to increase them by 100% etc.).

2.3 Tests of the MDS Method:

The effectiveness of the MDS method has been tested for a variety of theoretical distributions. An APL random number generating computer package was employed to produce data sets containing variables with various poisson, binomial, negative binomial, and gamma distributions. Parameters were varied in each of the families of distribution to provide a wide range of distributional shapes (e.g. from skewed, to normal, to uniform). These data sets were used as the "observed" data to be simulated by the MDS method. Each of the observed data matrices were created to have 200 cases and up to 9 variables of diverse distributions.

The poisson density is defined by the relationship:

$$p(X;\mu) = \mu X(e^{-\mu})/X! \qquad (6)$$

where $p(X;\mu)$ is the probability of X occurrences and μ is the "mean" parameter defining distributional shape. A data matrix consisting of a series poisson variables was generated using equation (6) & μ

values of 0.25, 0.50, 0.75, 1.0, 1.50, 2.0, 4.0, 5.0 and 10.0 to create the observed variables.

The binomial density is defined by the relationship:

$$p(X;N,P) = N!/\{X!(N-X)!\}(P)X(Q)^{N-X} \qquad (7)$$

where P is the "shape" parameters defining the probability of success, Q = 1-P, and N is the sample size, set at a constant value of 10 for these calculations. The values of P used to create the observed data matrix were 0.10, 0.25, 0.50, 0.75, and 0.90.

The negative binomial density is defined by the relationship:

$$P(X,M,R) = \Gamma(M+X)\ (R)^{-X}/\{X!\Gamma(M)(S)^{M+X}\}\ . \qquad (8)$$

One interpretation of the relationship X is the number of trials until M failures, where R = (1-P)/P, P being the probability of success, S = 1+R, Γ is the gamma function and M was set arbitrarily at 10. The values of R used were 0.10, 0.25, 0.50, 0.75, and 0.90.

The gamma density is defined by:

$$F(X) = (e^{-X}X^{\alpha-1})/\Gamma(\alpha), \qquad (9)$$

where is the "shape" parameters, which is also equal to the mean and is the gamma function. The values employed in the generation of the nine variables in the observed data matrix were 0.25, 0.50, 0.75, 1.0, 1.5, 2.0, 4.0, 5.0, and 10.0.

The poisson, binomial, negative binomial and gamma data sets were each introduced into the MDS program three times to test the effectiveness of the simulations for each of the distributional series. The degree of fit of each of the simulated to observed variables was tested with a Kolmogorov-Smirnov two sample test

(Siegel, 1956).

The MDS method was also subjected to a series of empirical tests by simulating water quality data from a monitoring program. The observed data was taken from six bimonthly cruises to a potential dredged material disposal site in the coastal waters off the mouth of the Chesapeake Bay. Each of the six data sets consisted of 16 variables measured on 18 samples. The simulated data was compared to the raw data by Kolmogorov-Smirnov two sample tests.

3. RESULTS

The results of the tests of the MDS simulations of theoretical distributions are presented in Table I. None of the comparisons indicated that the simulations were significantly different from the "observed" data at the $\alpha = 0.05$ level. Graphical comparisons from the four families of distributions were made to emphasize the closeness of fit of the simulations for a wide range of poisson (Figure 1), binomial (Figure 2), negative binomial (Figure 3) and gamma (Figure 4) densities. The simulations appeared to fit equally well for highly skewed data (e.g. Figure 1a-c; Figure 2a,d; Figure 3a; and Figure 4a,b), to more normal densities (e.g. Figure 1e; Figure 2c; Figure 3c; and Figure 4e), to nearly uni- form densities (e.g. Figures 1f, 3d, 4c) and various intermediate patterns (e.g. Figures 1d, 2b, 3b, 4d). The results of tests of MDS simulations of field data are presented in Table II. Despite the fact that the variables displayed a diversity of density patterns, only 3 of the 85 simulations were shown to be significantly different from the observed data. This number of deviations between the distributional patterns of the raw data and simulations would be expected to be due to chance alone.

4. DISCUSSION

The MDS method simulates multivariate data sets containing variables with a wide variety of distributions. It has been evaluated not only with the diverse test set of artifically created distributions, but with numerous data sets collected from nature as well. The method has consistently proven to be a rapid, effective simulation technique.

Techniques for the simulation of multivariate data sets such as the MDS method provide the environmental scientist with numerous techniques to aid in the evaluation of sampling/statistical regimes or in the interpretation of data sets from nature. Green (1979) reports that numerous investigators have evaluated statistical methods in the face of violations of assumptions by simulating and testing data which have the undesirable properties of the data from nature, but which also have been designed to satisfy either the null hypothesis (H_0) or alternate (H_A) hypothesis models. Thus, the actual levels of α and β errors can be compared to nominal values and the effectiveness of the statistical models may be assessed prior to their use. Green further suggests that in situations where the data violate the assumptions of the method quite severely, simulation can be used to test hypotheses directly. A series of data sets can be simulated to have the desirable properties (i.e. non-normality), but to also satisfy the H_0 model. These data sets are then tested by conventional statistical methods along with the observed data. Rather than resorting to statistical tables of critical test values for various levels, probability levels are defined by the percentage of simulated test statistic values exceeded by the

observed data value(s). In other words, H_0 can be rejected at an $\alpha = 0.05$ if at least 95% of the simulated test statistics are exceeded by the observed value.

A further use of simulated multivariate data sets is in the evaluation of the effectiveness of environmental monitoring programs. Data sets can be simulated to follow baseline distributions but with various levels of change in the means of the variables (i.e. true HA models). The simulated data sets can be considered to represent data taken following an environmental impact. The data sets with increasing levels of simulated "impacts" are sequentially tested with appropriate statistical methods until the differences are large enough that they can be detected routinely (at the predetermined level) in the context of the natural spatio-temporal variability. Thus, "minimum detectable impacts" can be defined for each parameters and the effectivenss of the monitoring program can be evaluated in terms of the ecological changes potentially detectable for the level of sampling effort. The MDS method of simulation has been successfully used for each of these techniques.

ACKNOWLEDGEMENT

This research was supported in part by Contract Number DACW65-81-C-0051 from the Norfolk District of the U.S. Army Corps of Engineers.

REFERENCES

Box, G.E.P. and D.R. Cox. 1964. An analysis of transformations. Journal of the Royal Statistical Society B, 26, 211-252.

Capra, J.R. and R.S. Elster. 1971. Notes on generating multivarate data with desired means, variances, and covariances. Educational and Psychological Measurement, 31, 749-752.

Gilman, L. and A.J. Rose. 1976. APL: An Interactive Approach. John Wiley & Sons, New York.

Green, R.H. 1979. Sampling Design and Statistical Methods for Environmental Biologists. John Wiley & Sons, New York.

Raeside, D.E. 1976. Monte carlo principles and applications. Physics in Medicine and Biology, 21, 181-197.

Siegel, S. 1956. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill, New York.

APPENDIX

TABLE I

Kolmogorov-Smirnov D_{max} values for comparisons of distributions of "observed" data matrices with simulated data.

Distribution Type	Variable	Run 1	Run 2	Run 3
Poisson:	1. μ = 0.25	0.16	0.18	0.17
	2. μ = 0.50	0.06	0.12	0.15
	3. μ = 0.75	0.10	0.04	0.12
	4. μ = 1.0	0.06	0.07	0.11
	5. μ = 1.5	0.04	0.05	0.10
	6. μ = 2.0	0.11	0.09	0.08
	7. μ = 4.0	0.08	0.04	0.04
	8. μ = 5.0	0.11	0.11	0.09
	9. μ =10.0	0.11	0.06	0.08
Binomial:	1. P = 0.10	0.05	0.07	0.09
	2. P = 0.25	0.05	0.08	0.11
	3. P = 0.50	0.08	0.09	0.05
	4. P = 0.75	0.08	0.09	0.12
	5. P = 0.90	0.03	0.11	0.05
Negative Binomial:	1. R = 0.10	0.06	0.03	0.08
	2. R = 0.25	0.10	0.07	0.05
	3. R = 0.50	0.08	0.08	0.11
	4. R = 0.75	0.10	0.06	0.08
	5. R = 0.90	0.10	0.07	0.10
Gamma:	1. α = 0.25	0.16	0.14	0.17
	2. α = 0.50	0.15	0.13	0.13
	3. α = 0.75	0.13	0.08	0.09
	4. α = 1.0	0.10	0.07	0.07
	5. α = 1.5	0.13	0.10	0.09
	6. α = 2.0	0.08	0.09	0.13
	7. α = 4.0	0.11	0.10	0.10
	8. α = 5.0	0.07	0.11	0.10
	9. α =10.0	0.06	0.08	0.04

Notes: D_{max} (0.05) (for $n_1=n_2=100$) = 0.192

D_{max} (0.01) (for $n_1=n_2=100$) = 0.231

TABLE II

Kolmogorov-Smirnov D_{max} values for comparisons of distributions of empirical water quality data with simulated data. Raw data were collected from bi-monthly cruises in the coastal waters off the mouth of the Chesapeake Bay.

Variable	Mar.	Apr.	Jun.	Aug.	Oct.	Jan.
Dissolved Oxygen	0.16	0.27	0.22	0.22	0.39	0.22
pH	0.25	0.16	0.33	0.28	0.33	0.27
Turbidity	0.25	0.13	0.11	0.28	0.22	0.20
Nitrite	0.05	0.25	0.22	0.39	--	0.22
Nitrate	--	0.50*	0.05	--	--	0.25
Orthophosphate	--	--	--	0.05	--	0.05
Total Phosphorous	0.08	--	--	0.11	--	0.10
TKN	0.20	0.19	0.21	0.16	0.11	0.17
Ammonia	0.33	0.44*	0.11	0.28	0.22	0.34
Suspended Solids	0.16	0.16	0.44*	0.21	0.11	0.15
Volatile Residue	0.16	0.22	0.22	0.16	0.16	0.18
Chlorophyll *a*	0.08	0.05	0.38	0.16	0.11	0.16
Chlorophyll *b*	0.25	0.05	0.22	0.16	0.39	0.21
Chlorophyll *c*	0.16	0.16	0.16	0.11	0.16	0.15
Phaeophytin	0.08	0.11	0.16	0.16	0.05	0.11

Notes: D_{max} (0.05) (for $n_1=n_2=18$) = 0.44

D_{max} (0.01) (for $n_1=n_2=18$) = 0.55

* Significant at $\alpha = 0.05$ level

-- Most or all samples below detection levels

FIGURE LEGENDS

Figure 1. MDS simulations of representative variables from the Poisson data set. Downswept crosshatched bars represent the density of selected Poisson variables created by a random number generator. Upswept crosshatched bars represent the mean density patterns of three simulations and the vertical lines represent the 95% confidence limits.

Figure 2. MDS simulations of representative variables from the Binomial data set. Downswept crosshatched bars represent the density of selected Binomial variables created by a random number generator. Upswept crosshatched bars represent the mean density pattern of three simulations and the vertical lines represent the 95% confidence limits.

Figure 3. MDS simulations of representative variables from the Negative Binomial data set. Downswept crosshatached bars represent the density of selected Negative Binomial variables created by a random number generator. Upswept crosshatched bars represent the mean density pattern of three simulations and the vertical lines represent the 95% confidence limits.

Figure 4. MDS simulations of representative continuous variables from the Gamma data set. Closed circles represent the density of selected Gamma variables created by a random number generator. Open circles represent the mean density pattern of three simulations and the vertical lines represent the 95% confidence limits.

Figure 1.

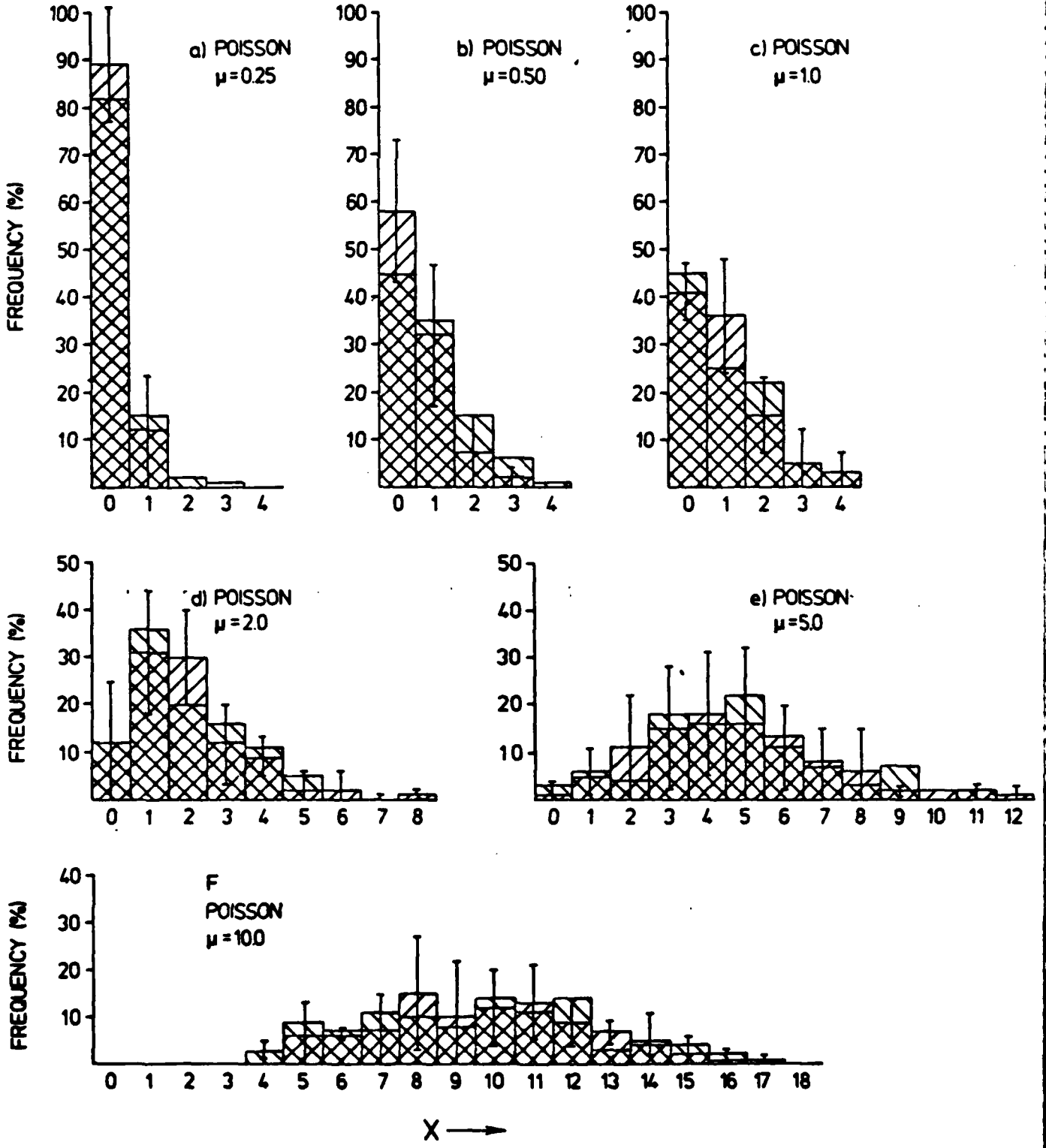
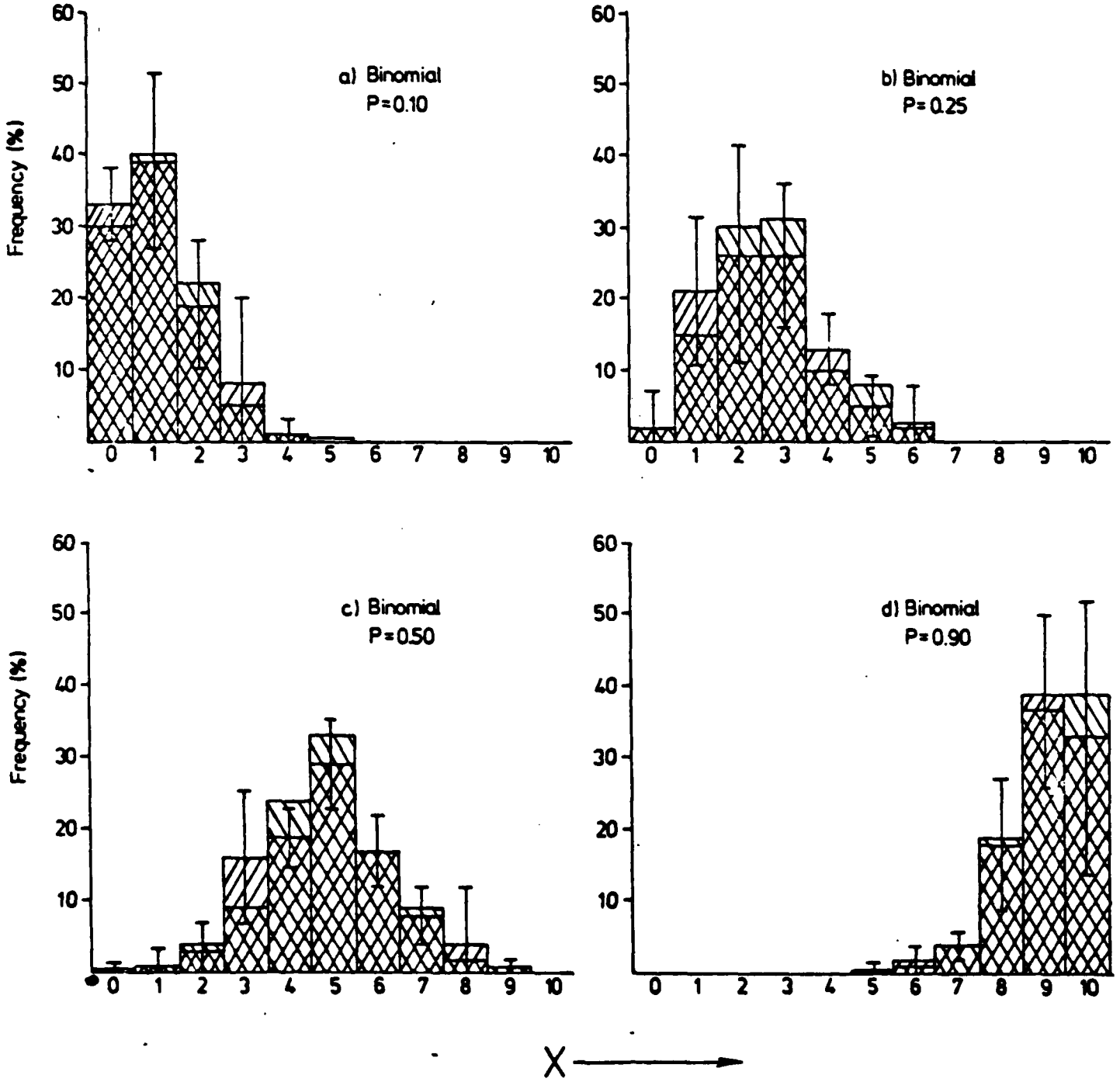

a) POISSON
μ=0.25
b) POISSON
μ=0.50
c) POISSON
μ=1.0
d) POISSON
μ=2.0
e) POISSON
μ=5.0
F
POISSON
μ=10.0
FREQUENCY (%)
X

FIGURE 2.


a) Binomial
P=0.10
b) Binomial
P=0.25
c) Binomial
P=0.50
d) Binomial
P=0.90
Frequency (%)
Frequency (%)
60
50
40
30
20
10
0 1 2 3 4 5 6 7 8 9 10
X

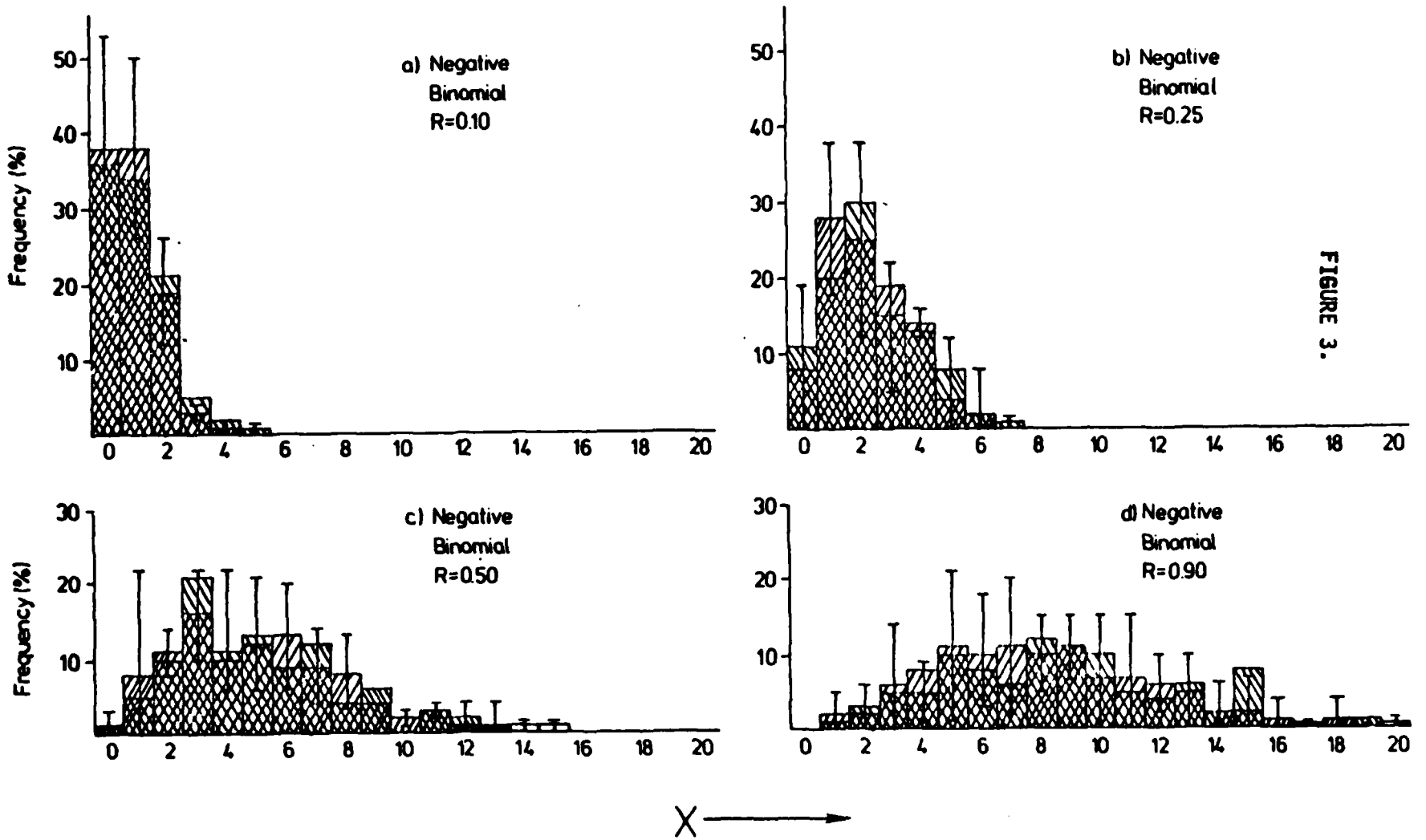
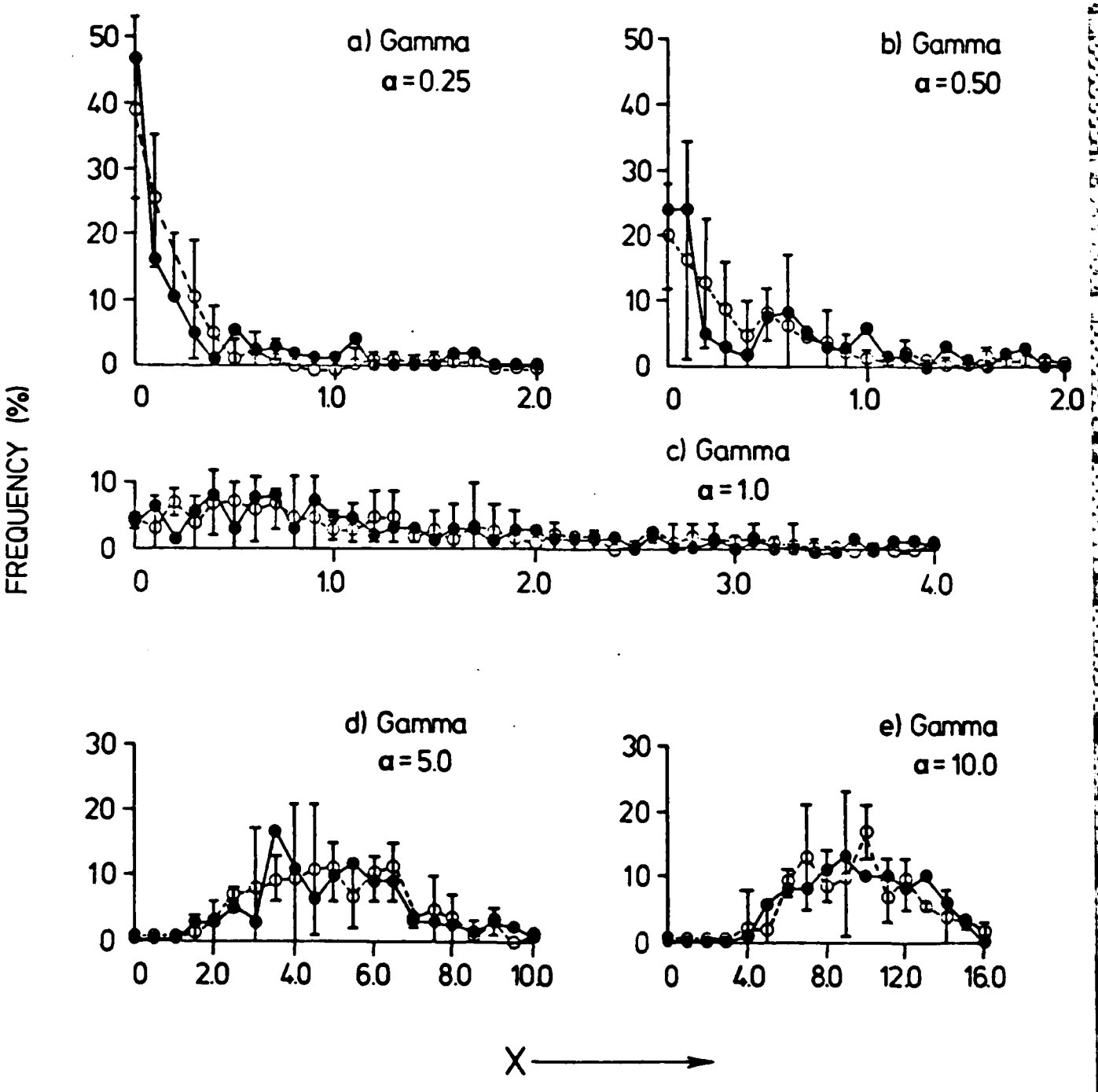

a) Negative Binomial R=0.10
b) Negative Binomial R=0.25
c) Negative Binomial R=0.50
d) Negative Binomial R=0.90
Frequency (%)
X

FIGURE 3.

FIGURE 4.


a) Gamma
α = 0.25
b) Gamma
α = 0.50
c) Gamma
α = 1.0
d) Gamma
α = 5.0
e) Gamma
α = 10.0
FREQUENCY (%)
X

END

FILMED

DTIC