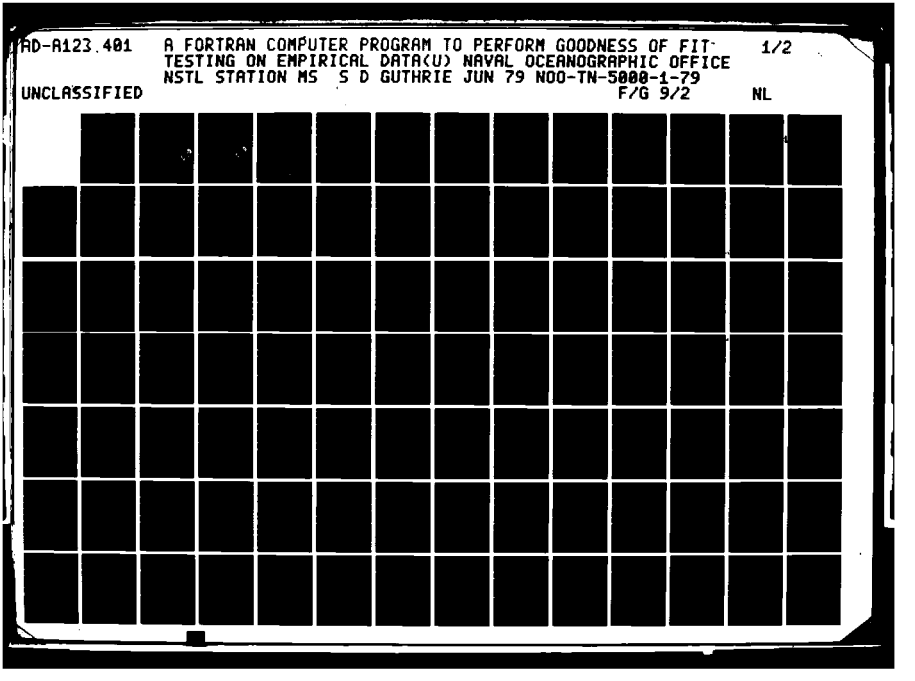
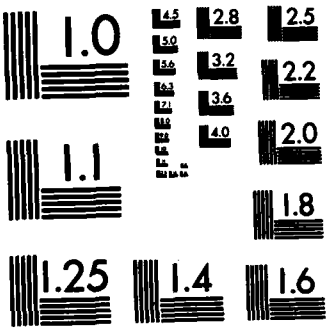
AD-A123 401 A FORTRAN COMPUTER PROGRAM TO PERFORM GOODNESS OF FIT- 1/2
TESTING ON EMPIRICAL DATA(U) NAVAL OCEANOGRAPHIC OFFICE
NSTL STATION MS S D GUTHRIE JUN 79 NOO-TN-5000-1-79
UNCLASSIFIED F/G 9/2 NL





TN 5000-1-79

TECHNICAL NOTE

A FORTRAN COMPUTER PROGRAM TO PERFORM GOODNESS OF FIT TESTING ON EMPIRICAL DATA

SUE D. GUTHRIE

JUNE 1979

ABSTRACT

This document describes a FORTRAN computer program which performs statistical goodness of fit tests on empirical data. Four goodness of fit methods are available: (1) chi-square, (2) Kolmogorov-Smirnov, (3) Cramer-Von Mises, and (4) the moments test for normality. Data may be tested against any one of ten theoretical probability distributions: (1) Poisson, (2) exponential, (3) normal, (4) log-normal, (5) gamma, (6) Erlang-k, (7) chi-square, (8) triangular, (9) uniform, or (10) Weibull.

NAVAL OCEANOGRAPHIC OFFICE
NSTL STATION, BAY ST. LOUIS, MISSISSIPPI 39522

Technical Notes are informal reports prepared primarily for internal use and have not received the rigorous editing and technical review or command approval given formal publications of the Naval Oceanographic Office.

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

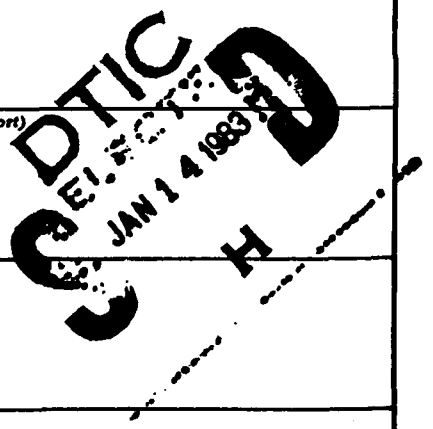
READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: TN 5000-1-79	2. GOVT ACCESSION NO.: AD-A123 401
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE *(and Subtitle)*: A FORTRAN COMPUTER PROGRAM TO PERFORM GOODNESS OF FIT TESTING ON EMPIRICAL DATA	5. TYPE OF REPORT & PERIOD COVERED: TECHNICAL NOTE
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): SUE D. GUTHRIE	8. CONTRACT OR GRANT NUMBER(s):
9. PERFORMING ORGANIZATION NAME AND ADDRESS:	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS: U.S. NAVAL OCEANOGRAPHIC OFFICE, N.S.T.L. Station, Bay St. Louis, MS 39522	12. REPORT DATE: JUNE 1979
	13. NUMBER OF PAGES: 105
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*:	15. SECURITY CLASS. *(of this report)*: UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*



18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

FORTRAN; FITTING FUNCTIONS (MATHEMATICS);STATISTICAL DISTRIBUTIONS; TESTS

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

This document describes a FORTRAN computer program which performs statistical goodness of fit tests on empirical data. Four goodness of fit methods are available: (1) chi-square, (2) Kolmogorov-Smirnov, (3) Cramer-Von Mises, and (4) the mements test for normality. Data may be tested against any one of ten theoretical probability distributions: (1) Poisson, (2) exponential, (3) normal, (4) log-normal, (5) gamma, (6) Erlang-k, (7) chi-square, (8) triangular, (9) uniform, or (10) Weibull.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

TECHNICAL NOTE

A FORTRAN COMPUTER PROGRAM TO PERFORM GOODNESS OF FIT TESTING ON EMPIRICAL DATA

SUE D. GUTHRIE

JUNE 1979

ABSTRACT

This document describes a FORTRAN computer program which performs statistical goodness of fit tests on empirical data. Four goodness of fit methods are available: (1) chi-square, (2) Kolmogorov-Smirnov, (3) Cramer-Vos Mises, and (4) the moments test for normality. Data may be tested against any one of ten theoretical probability distributions: (1) Poisson, (2) exponential, (3) normal, (4) log-normal, (5) gamma, (6) Erlang-k, (7) chi-square, (8) triangular, (9) uniform, or (10) Weibull.

NAVAL OCEANOGRAPHIC OFFICE
NSTL STATION, BAY ST. LOUIS, MISSISSIPPI 39522

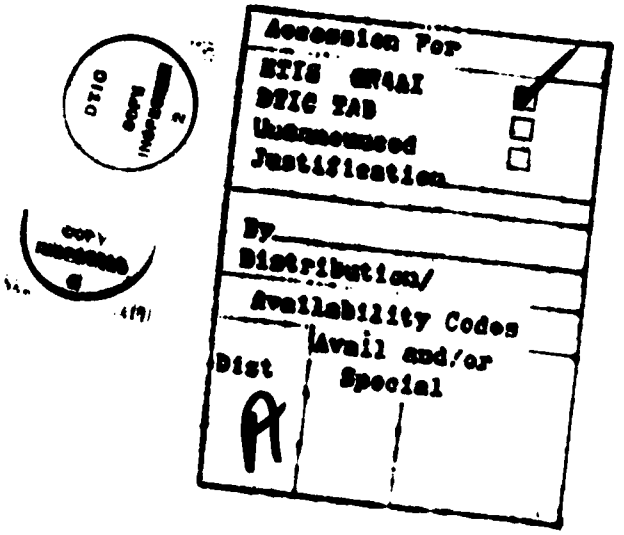
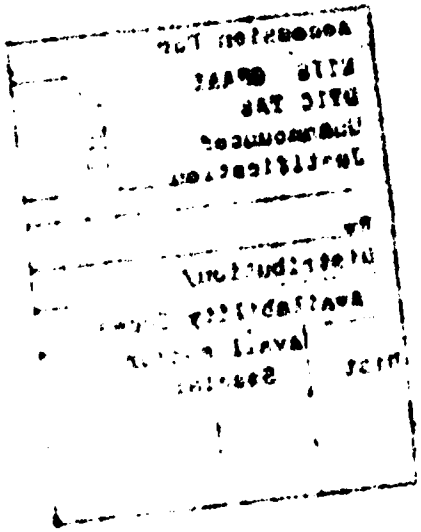


TABLE OF CONTENTS



LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

A. General

Data gathering in an attempt to gain knowledge about a particular phenomenon is a key element in scientific work. It is a preliminary step in the mathematical modeling or simulation of any real world system under study. Historically, much of the data gathered from existing processes is characterized as random in nature and quite often can be identified as belonging to a known theoretical probability distribution. Identification of an underlying probability distribution is one step often used by researchers in their attempts to describe correctly the system they are investigating.

The selection of a possible probability distribution representative of a set of sample data can be attempted from at least four different approaches. Often, historical material, or experience of the researcher with certain phenomena, leads to a decision that a particular sample of data is known to be from a normal distribution, an exponential distribution, or some other distinctive distribution. Another technique used to identify an underlying probability distribution is visual examination--the data is plotted (often as histograms) and the plot is compared with shapes of theoretical distributions. A third approach in identifying a possible distribution for a set of empirical data is through the comparison of descriptive statistics calculated from the sample data with known parameter values of the hypothesized distribution. A fourth method available is the mathematical comparison of the actual number and characteristics of sample observations with the expected number and characteristics of theoretical observations.

Mathematically comparing a theoretical probability distribution with an empirical probability distribution is a "goodness of fit" test. The researcher, in performing a goodness of fit test, makes two preliminary decisions: (1) the theoretical probability distribution against which he wishes to test his data and (2) the level of significance at which he is willing to accept the results of the test. Therefore, goodness of fit testing is simply a mathematical tool to assist the researcher in deciding whether his sample observations fit a theoretical probability distribution well enough for him to describe his system in terms of that distribution.

B. Purpose of Work

In order for a researcher to perform a goodness of fit test on empirical data, he must complete the following six steps:

1. statement of the hypothesis to be tested,
2. selection of a goodness of fit method,
3. specification of the significance level,
4. execution of the mathematical computations for the selected goodness of fit test,
5. formulation of the decision rule for the test, and
6. acceptance or rejection of the null hypothesis.

These six steps require time and knowledge. If a computer is used for sorting, grouping, or calculating, some time may be saved. However, if the researcher must write programs to perform these tasks, additional time and skills are required. If various programs already exist and are available to the researcher, time and effort are saved. Of course, the researcher must have the knowledge to run each program. The computer program described in this report provides the researcher with a single package by which to implement all six steps in goodness of fit testing--saving time and requiring less knowledge than the previously discussed alternatives.

C. Functions of the Program

The goodness of fit (GOF) program offers the following four goodness of fit tests:

1. chi-square test,
2. Kolmogorov-Smirnov test,
3. Cramer-Von Mises test, and
4. moments test.

These four tests can be applied to all of the following ten theoretical probability distributions:

1. Poisson,
2. exponential,
3. normal,
4. log-normal,
5. gamma,
6. Erlang-k,
7. chi-square,
8. triangular,
9. uniform, and
10. Weibull.

In addition to performing the goodness of fit calculations, the GOF program supplies descriptive sample statistics and a printed histogram of the empirical data. Critical value tables for each goodness of fit technique are stored as part of the software allowing the GOF program to evaluate the results of each test at the .01 and .05 levels of significance.

The GOF program operates in either batch or on-line modes. Program options and capabilities are selected by the user through simple-to-use English language statements. These statements are described in Section IV of this document. Sample output from the on-line version of GOF program is included in Appendix A of this document.

II. DESCRIPTION OF GOODNESS OF FIT METHODS

A. General

A goodness of fit test is a statistical technique for comparing a distribution of observed data with a theoretical probability distribution. The theoretical distribution to be tested is stated in the null hypothesis. The purpose of the comparison is to determine how closely the observed values agree with the theoretical values. If they agree sufficiently, the differences between the two sets of values can be attributed to sampling error. If they do not agree sufficiently, the null hypothesis that the observed data comes from the same probability distribution as the theoretical distribution cannot be accepted. A basic assumption of goodness of fit testing is that the observed data must be the result of random sampling.

Several goodness of fit tests exist. The chi-square (also referred to as χ^2) test is the best known technique because it is the oldest and most widely publicized. Almost every text on statistical analysis mentions the chi-square goodness of fit test. The Kolmogorov-Smirnov (also referred to as D_n) and Cramer-Von Mises (also referred to as $n\omega^2$) tests are less frequently documented than the chi-square test, but offer strengths not available from the chi-square test. The moments test is important because it is used as a test for the normal distribution which is the most frequently used distribution of all defined theoretical probability distributions.

Each goodness of fit test has its particular advantages and limitations which are explained later in this report under the descriptions of each individual test. The first three tests, the chi-square, the Kolmogorov-Smirnov, and the Cramer-Von Mises, are broad spectrum tests. They can be used to test null hypotheses for a wide variety of distributions. The fourth test, the moments test, is included only to test null hypotheses for the normal distribution. Cox (5:147-149) gives a synopsis of some of the investigation done into the power of the three broad spectrum tests. A discussion of power of the various tests is beyond the scope of this work.

The flowchart in Figure 1 offers a quick and general path for selecting the goodness of fit tests most applicable to a given set of observations. Decisions not reflected by this flowchart can be used, but are not recommended by this author.

B. Chi-square Test

The test statistic (χ^2) for the chi-square test is a cumulative value computed by

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} ,$$

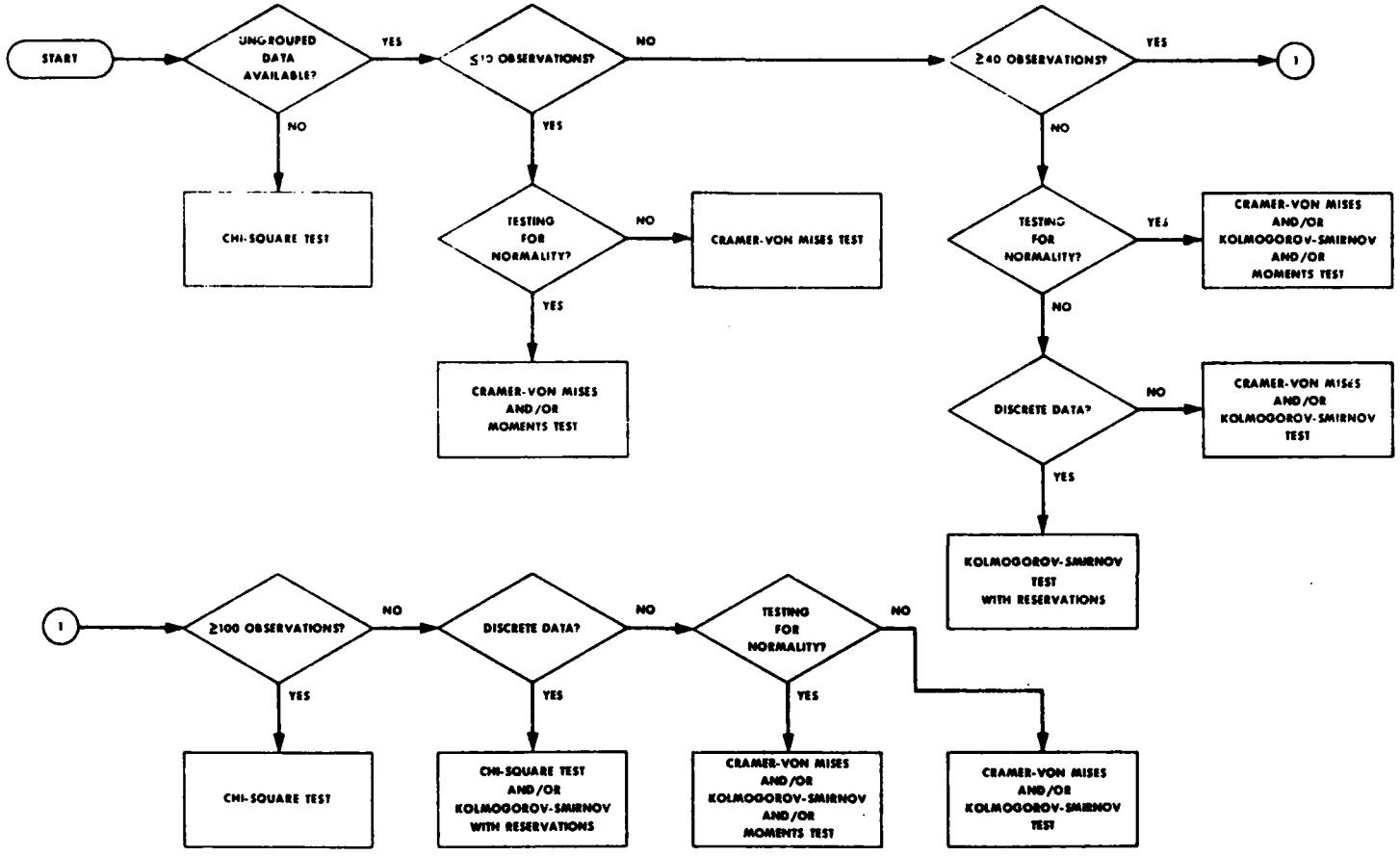


Figure 1

Goodness of Fit Decision Path

where:

n = the number of cells into which the data is grouped,
O_i = the observed absolute frequency of the ith cell, and
E_i = the expected absolute frequency of the ith cell.

Before the chi-square statistic can be computed, the observed data values must be grouped into cells or classes. The number of entries in each cell is counted and each count represents the observed absolute frequency of that cell. The expected frequency values are computed from the relative frequency values derived from the hypothesized theoretical distribution for each cell. The difference between these two frequency values is squared and divided by the expected (theoretical) frequency value of that cell. Each of the resulting values is summed to produce the χ^2 statistic.

The χ^2 statistic is always nonnegative because of the squaring of the differences. Once the χ^2 statistic is computed, its value must be compared with the appropriate value in a table of critical χ^2 values. A χ^2 value of zero would mean a perfect fit of the observed data to the hypothesized theoretical distribution. The larger the difference of $\chi^2_{computed} - \chi^2_{tabular}$, the worse the fit.

A χ^2 critical value table is organized by levels of significance and degrees of freedom. The level of significance is $1-\alpha$ (α is the probability of a Type I error).* The level of significance of a test is selected by the user based on his knowledge of what is acceptable for his situation. The degrees of freedom value reflects the number of restrictions being imposed on the theoretical distribution. Another way of describing the number of degrees of freedom is that it is the "number of ways in which two sets of data that are being compared are free to vary" (Caulcott, 2:113). Usually, the number of degrees of freedom for the chi-square test is expressed as $n-p-1$, where n is the number of cells used to compute the χ^2 statistic and p is the number of distribution (population) parameters which had to be estimated by sample parameters. The constant "1" is always subtracted because the number of cells used for the observed and the theoretical values is restricted to being identical. Therefore, one degree of freedom, representing the restriction that the number of cells be identical, is always lost. The value p varies with the probability distribution named in the null hypothesis. This variation occurs because different distributions have different numbers of characteristic parameters. Table 1 gives the appropriate value of p for each distribution named. If the population parameters are known and, therefore, not estimated by sample parameters, p is not subtracted.

---

*A Type I error occurs when the null hypothesis is falsely rejected.

Table 1

Number of Parameters Per Distribution

p	Distribution
1	Poisson, exponential, chi-square
2	normal, gamma, Weibull, log-normal, uniform, Erlang-k
3	triangular

Prior to performing a chi-square test, the researcher must decide on the number of cells and the cell size he wishes to use. If his data comes already grouped, he simply counts the number of cells given. However, with ungrouped observations, he must select the number of cells to be used in computing the χ^2 statistic. There are no firmly established rules for selecting the number of cells. Cochran (4:332) writes, "I believe that the common practice is to have a moderate number of classes, say between 10 and 25, and to make the class intervals equal." Cochran himself and other authors he cites prefer using unequal cell intervals under certain conditions (4:332-334). Because of the additional complexity introduced by unequal cell intervals and because of the lack of firm convictions about this matter, the choice for the GOF program is equal cell intervals. The choice of the number of equal-interval cells becomes a game of trade-offs under certain data conditions. Most authors recommend that each cell have a frequency count of at least 5 and preferably not more than 50 before performing the chi-square test.

The chi-square test is the only technique implemented in the GOF program which is recommended for use with grouped data. The grouping of data required by the chi-square test is usually referred to as a drawback of the test. However, if only grouped data is available, that characteristic of the chi-square test becomes a valuable feature.

How good is the chi-square test? If the null hypotheses is rejected on the basis of the chi-square statistic, one can feel fairly confident of the test results. However, if the null hypothesis is not rejected, one is on unsafe ground. "For instance, for moderate sample sizes, . . . , the test accepts as normal almost every distribution with a symmetric, single maximum, density" (Breiman, 1:202). To strengthen one's position when the chi-square test indicates acceptance, there are several supplementary tests which can be used in certain situations: a test based on runs in the individual deviations calculated while computing the chi-square statistic and a test on the low moments of the distribution (4:339-340).

Several limitations are direct results of the chi-square test definition.

1. Only actual, not relative, frequency values are acceptable for the test.
2. The frequency count for each class should be at least 5 (17).

3. When the frequency count for a class is too large (n>50), there is a related loss of power* in the test (4).

4. The number of observed values should be at least 40 before the chi-square test is used.

5. When data is difficult, impossible, or expensive to acquire, the requirement of at least 40 observations becomes a severe limitation.

Other more subtle limitations of the chi-square test are described by Cochran (4).

C. Kolmogorov-Smirnov Test

The test statistic for the two-sided Kolmogorov-Smirnov goodness of fit test is a single value computed by

$$D_n = \underset{\text{all } i}{\text{Max}} \left| F(O_i) - F(E_i) \right| ,$$

where:

$F(O_i)$ is the value of the cumulative distribution function for the observed data at point i,

$F(E_i)$ is the value of the cumulative distribution function for the hypothesized (expected) value at point i, and

n is the number of observations in the test.

The preceding formula defines the test statistic (D_n) as the maximum absolute difference between two cumulative distribution functions computed at each sample point. Figure 2 displays graphically the method used by the Kolmogorov-Smirnov test. The maximum D_n is the computed test statistic which is compared to the appropriate critical value from a table of critical values for the Kolmogorov-Smirnov test. The appropriate tabular value is located by sample size (n) for the sample being tested. If the computed D_n value is larger than the tabular critical value, the null hypothesis is rejected.

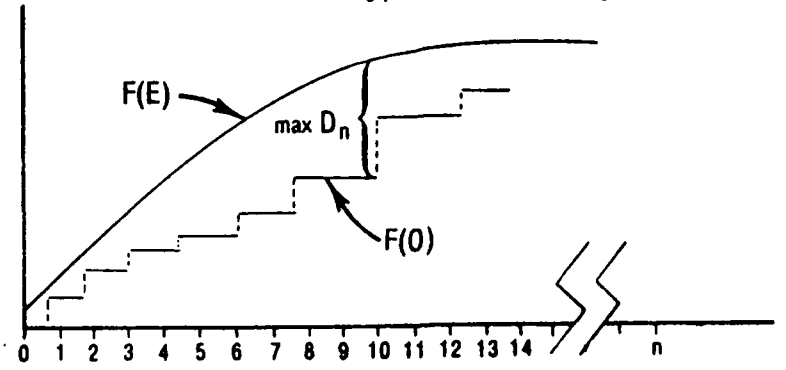


Figure 2
Kolmogorov-Smirnov Test

*Power is the probability of rejecting the null hypothesis when the alternative hypothesis is true.

One of the strengths of the Kolmogorov-Smirnov test lies in its use of each individual observation; whereas the chi-square test requires that the data be grouped. The Kolmogorov-Smirnov test can also be applied to grouped data (12), but is not recommended by this author because the test is less powerful for grouped data.

The Kolmogorov-Smirnov test has two advantages over the chi-square test. Sample observations do not need to be grouped as they do for the chi-square test. Therefore, information is not lost through grouping, which results in the Kolmogorov-Smirnov test being more powerful. Secondly, the number of observations required by the two tests can be less for the Kolmogorov-Smirnov test (15). A range of between 10 and 100 observations is the recommended range for the Kolmogorov-Smirnov test.

Kolmogorov and Smirnov developed their test to be used when the population mean and variance are known. However, it can be employed when the population parameters are estimated by the sample mean and variance and still performs better than the chi-square test (12,16). Although Kolmogorov and Smirnov developed their test to be used mainly in testing continuous distributions, it can be used conservatively with distributions which are not continuous (6).

One final attribute is the simplicity of the critical value table for the Kolmogorov-Smirnov test. One only needs to know the sample size (n) to look up tabulated critical values. There is no worry with degrees of freedom or whether the population parameters are estimated.

D. Cramer-Von Mises Test

The Cramer-Von Mises test statistic (referred to as the ω^2 or $n\omega^2$ statistic) is defined as

$$\omega^2 = \int_{-\infty}^{\infty} (F_o(x) - F_e(x))^2 dF(x),$$

where:

$F_o(x)$ is the cumulative distribution function of the observed samples, and

$F_e(x)$ is the cumulative theoretical distribution function of the null hypothesis.

The integral is computed by the approximation

$$\omega^2 = \frac{1}{12n^2} + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{2i-1}{2n} - F(x_i)\right]^2,$$

where n is the size of the observed sample. Multiplying through by n yields

$$n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[\frac{2i-1}{2n} - F(x_i)\right]^2.$$

The Cramer-Von Mises and the Kolmogorov-Smirnov tests are similar in that they both compare the cumulative theoretical distribution function with the distribution function of the observed sample. However, the Kolmogorov-Smirnov test statistic (D_n) is a single value statistic while the Cramer-Von Mises test statistic ($n\omega^2$) is a cumulative value. The computed $n\omega^2$ test statistic can be compared with the critical values in a Cramer-Von Mises table. The tabular value is selected by the appropriate α value, where α is the probability of a Type I error. If the computed $n\omega^2$ value exceeds the tabular value, the null hypothesis cannot be accepted.

The Cramer-Von Mises test, like the Kolmogorov-Smirnov test, considers each observed sample separately. It does not require that the data be grouped as does the chi-square test. The major attribute of the Cramer-Von Mises test lies in its ability to reliably compute a test statistic for small samples "and has been applied with as few as eight or ten observations" (Phillips, 16:5). The Cramer-Von Mises test, like the chi-square and the Kolmogorov-Smirnov tests, is a distribution-free (non-parametric) test; it does not care which theoretical distribution is being tested in the null hypothesis.

E. Moments Test for Normality

"A moment is the average deviation of a set of data about a point" (Harnett, 8:100). Moments are important because they describe certain characteristics of distributions. The two most often used moments are the first moment, the mean, and the second moment, the variance. They describe the location and the variation, respectively, of a distribution. The moments used in the moments test are the third and fourth moments. The third moment is the measure of skewness of a distribution function. The fourth moment, the measure of kurtosis, describes the peakedness of a distribution function. Figure 3 illustrates the meaning of the first moment, the measure of location. The shape of the two distributions is identical, but their first moments differ. Figure 4 demonstrates the effect of differences in the second moment, the moment of variation (or dispersion). The effects of the moment of skewness (or lack of symmetry) are shown in Figure 5. Figure 6 displays the role of the fourth moment, the moment of kurtosis (or peakedness).

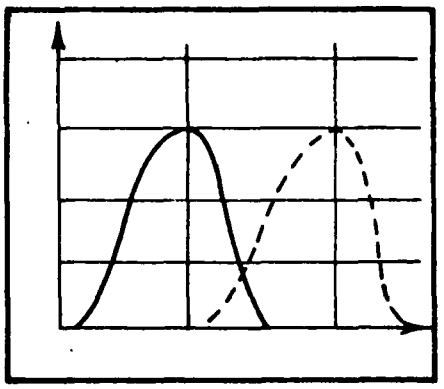


Figure 3
Measure of Location

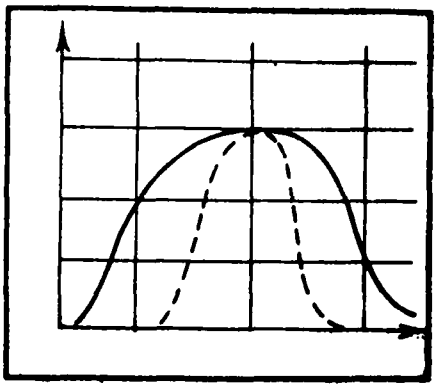


Figure 4
Measure of Variation

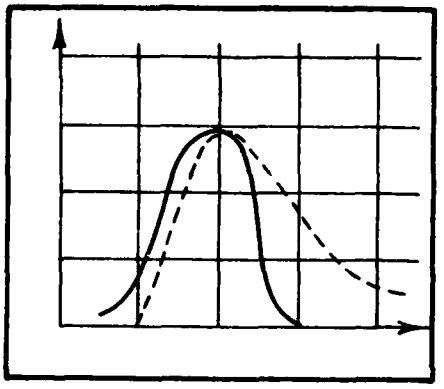


Figure 5
Measure of Skewness

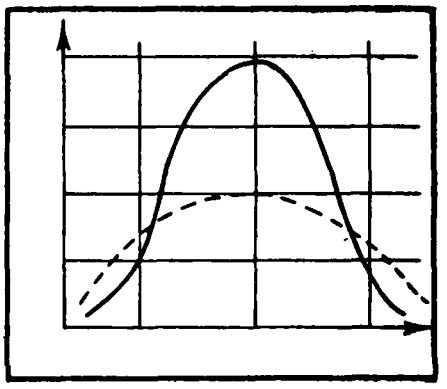


Figure 6
Measure of Kurtosis

When the point about which a moment is calculated is the mean, the moment is called a central moment. For a continuous distribution, the nth central moment is defined by

$$\mu_n = \int_{\text{all } x} (x - \mu)^n f(x)\,dx,$$

where μ is the mean of the distribution. The third central moment describes the skewness. With $n = 3$, all values used in the computation of the third moment retain their positive or negative signs. Therefore, the third moment of a symmetrical distribution is close to zero. The third moment of a truly normal distribution is zero. For a distribution which is skewed to the right, the value of the third moment is positive and when skewed to the left, it is negative. To eliminate the effect

of the size of the units of measurement of the sample observations, a dimensionless quantity is the most common method of describing the third moment. It is

$$S_k = \frac{\mu_3}{\sigma^3},$$

where:

μ_3 is the third moment, and
σ^3 is the cube of the standard deviation.

If the absolute value of S_k is greater than 1, then the distribution being tested is very skewed (8). Figures 7 and 8 demonstrate left (negative value of S_k) and right (positive value of S_k) skewness, respectively.

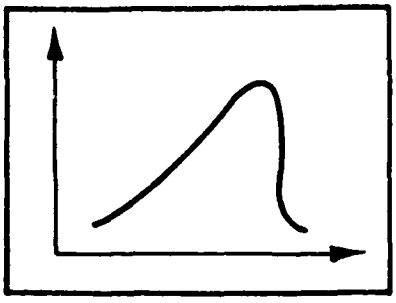


Figure 7
Negative Skewness

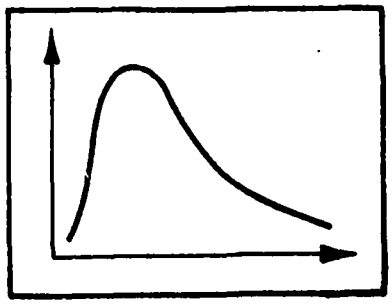


Figure 8
Positive Skewness

The fourth moment about the mean measures kurtosis. Kurtosis is a description of flatness or peakedness of the distribution curve. To create a dimensionless value, independent of the scale of the recorded observations, the following formula is used:

$$K_t = \frac{\mu_4}{\sigma^4},$$

where:

μ_4 is the fourth moment, and
σ^4 is the standard deviation raised to the fourth power.

Because n = 4 for the fourth moment, all values are raised to the fourth power and are, therefore, always positive. The smallest value K_t can be is 1. A normal curve has a kurtosis value of 3 and is called a mesokurtic curve. If a computed K_t is less than 3, the distribution is platykurtic and if K_t is greater than 3, the distribution is described as leptokurtic. Figures 9, 10, and 11 illustrate mesokurtic, platykurtic, and leptokurtic shapes (8).

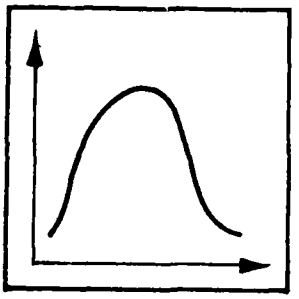


Figure 9
Mesokurtic Shape

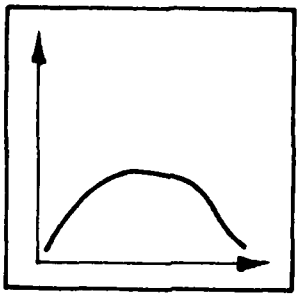


Figure 10
Platykurtic Shape

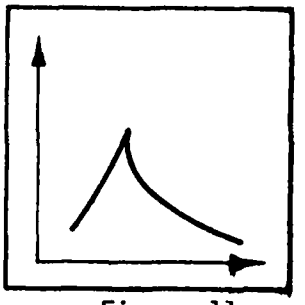


Figure 11
Leptokurtic Shape

The population parameters for skewness and kurtosis are designated by the dimensionless values ν_1, and ν_2, respectively, and are defined as

$$\nu_1 = \frac{\delta_3^2}{\delta_2^3} \quad \text{and} \quad \nu_2 = \frac{\delta_4}{\delta_2^2},$$

where δ_n = the nth moment of the population. The first four moments of the normal distribution have the values μ, σ^2, 0, and $3\sigma^4$. Substituting the values of the moments for the normal distribution into the above equations gives $\nu_1 = 0$ and $\nu_2 = 3$.

For convenience, we define

$$\tilde{\nu}_1 = \sqrt{\nu_1} = \frac{\delta_3}{\delta_2^{3/2}}, \text{ and}$$

$$\tilde{\nu}_2 = (\nu_2 - 3) = \frac{\delta_4 - 3\delta_2^2}{\delta_2^2}.$$

Hence, for a normal distribution $\tilde{\nu}_1 = 0$ and $\tilde{\nu}_2 = 0$ (Phillips, 16:13).

A positive $\tilde{\nu}_1$ denotes a skew to the right; a negative $\tilde{\nu}_1$, a skew to the left. A positive $\tilde{\nu}_2$ denotes a platykurtic shape; a negative $\tilde{\nu}_2$, a leptokurtic shape. Values of zero for $\tilde{\nu}_1$, and $\tilde{\nu}_2$ describe a symmetrical, mesokurtic shape.

The kth moment about the mean is defined as

$$M_k = \int_{-\infty}^{\infty} (x - \mu)^k f(x)\,dx$$

for continuous distributions and

$$M_k = \sum_{-\infty}^{\infty}(x - \mu)^k f(x)$$

for discrete distributions, where μ is the expected value $E(x)$ of the random variable x. "Given a sample of size n, the kth moment about the origin can be estimated by

$$\phi_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \text{"} \quad \text{(Phillips, 16:14).}$$

To compute the moments about the mean for the sample observations using the above estimating formula, the following equations are used:

$$m_1 = 0, \qquad m_3 = \phi_3 - 3\phi_2\phi_1 + 2\phi_1^3, \text{ and}$$

$$m_2 = \phi_2 - \phi_1^2, \qquad m_4 = \phi_4 - 4\phi_3\phi_1 + 6\phi_2\phi_1^2 - 3\phi_1^4.$$

Substituting the sample terms into the equations for ν_1 and ν_2 gives the sample estimators

$$\beta_1 = \frac{m_3}{m_2^{3/2}} \quad \text{and} \quad \beta_2 = \frac{m_4 - 3m_2^2}{m_2^2}.$$

β_1 and β_2 are biased sample estimators of the population parameters ν_1 and ν_2. Unbiased estimators of the population skewness and kurtosis measures can be calculated by using $\tilde{\alpha}_1 = r_3/r_2^{3/2}$ and $\tilde{\alpha}_2 = r_4/r_2^2$. An estimator is considered to be unbiased when its average value (mean), taken over all possible random samples of the same size, is equal to the population parameter for which it is an estimate. To compute the r-values used by the unbiased estimators, Phillips (16:14-15) gives the following equations:

$$r_1 = \frac{S_1}{n},$$

$$r_2 = \frac{nS_2 - S_1^2}{n(n-1)},$$

$$r_3 = \frac{n^2S_3 - 3nS_2S_1 + 2S_1^3}{n(n-1)(n-2)}, \text{ and}$$

$$r_4 = \frac{(n^3+n^2)S_4 - 4(n^2+n)S_3S_1 - 3(n^2-n)S_2^2 + 12S_2S_1^2 - 6S_1^4}{n(n-1)(n-2)(n-3)}$$

where

$$S_k = \sum_{j=1}^{n} x_j^k, \text{ and}$$

n = number of observations.

Once the biased estimators, β_1 and β_2, or the unbiased estimators, $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$, are computed, they must be checked against the corresponding tabular values stored in the GOF program code. The critical values for the moments test are stored for sample sizes $10 \leq n \leq 125$. For sample sizes larger than 125, the distribution of the biased estimator of skewness β_1 approaches the normal distribution with a mean of zero and a variance of 6/n. The biased estimator of kurtosis β_2 also approaches the normal distribution with a mean of 3 and a variance of 24/n. β_1 and β_2 can be transformed to standard normal deviates by the following two equations:

$$z_1 = \frac{\beta_1}{\sqrt{6/n}} \quad \text{and} \quad z_2 = \frac{\beta_2 - 3}{\sqrt{24/n}} .$$

Standard normal tables may be used to determine the critical values, at a selected α value, against which z_1 and z_2 are compared.

The unbiased estimators may be treated in a similar fashion as they both approach normal distributions. The means for both distributions are zero. The variance for the $\tilde{\alpha}_1$ distribution is 6/n and the variance of the $\tilde{\alpha}_2$ distribution is 24/n. Therefore, the standard normal deviates are calculated by

$$z_1 = \frac{\beta_1}{\sqrt{6/n}} \quad \text{and} \quad z_2 = \frac{\beta_2 - 3}{\sqrt{24/n}} \qquad (16).$$

III. THEORETICAL PROBABILITY DISTRIBUTIONS

A. General

In simple terms, a theoretical probability distribution is a model--a means of describing some random, rather than deterministic, phenomena. There are two basic types of probability distributions: the discrete distribution and the continuous distribution. A discrete distribution describes the nature of a random variable that can assume only a finite or countable number of values. A continuous distribution describes the nature of a random variable that can take an infinite set of values. Discrete random variables are often described as those values that are countable; continuous random variables may be described as those values that are measurable. Therefore, the number of eggs per chicken per year would be a discrete variable while the weight of the eggs per chicken per year would be a continuous variable.

Ten theoretical probability distributions are described in this section. The only discrete distribution included is the Poisson distribution. The other nine distributions, which describe continuous phenomena are the: exponential, normal, log-normal, gamma, Erlang-k, chi-square, triangular, uniform, and Weibull.

Before stating the null hypothesis for a goodness of fit test, one must first elect the theoretical distribution with which he wishes to compare his observed data. The selection of a theoretical distribution can be aided by answering two simple questions.

1. Does the observed data come from an environment similar to those historically associated with a particular theoretical distribution?
2. Does the histogram of the observed data resemble the characteristic shape of the graph of a theoretical distribution?

An affirmative answer to the first question is the more important criterion for selection. Affirmative answers to both questions are only a beginning. They provide a means of selecting a distribution with which to compare one's observed data.

The flowchart in Figure 12 provides a verbal inquiry path to help evaluate the correspondence between the environment of one's observed data and the appropriateness of each theoretical distribution to that environment. It is a related series of questions similar to question 1 above. Figure 13 supplies a pictorial pursuit with which to compare a histogram of one's observed data.

In addition to Figures 12 and 13, the author has tried to provide three other aids for pinpointing "likely" theoretical distributions to test. Those three aids are the coefficients of variation, skewness, and kurtosis for each distribution (where possible). Section II of this report contains an explanation of the latter two measures. While the moments of a random variable do not uniquely identify its theoretical distribution, they do provide salient information about its nature. The coefficient of variation is the standard deviation divided by the mean of the distribution. For sample data, it is computed by

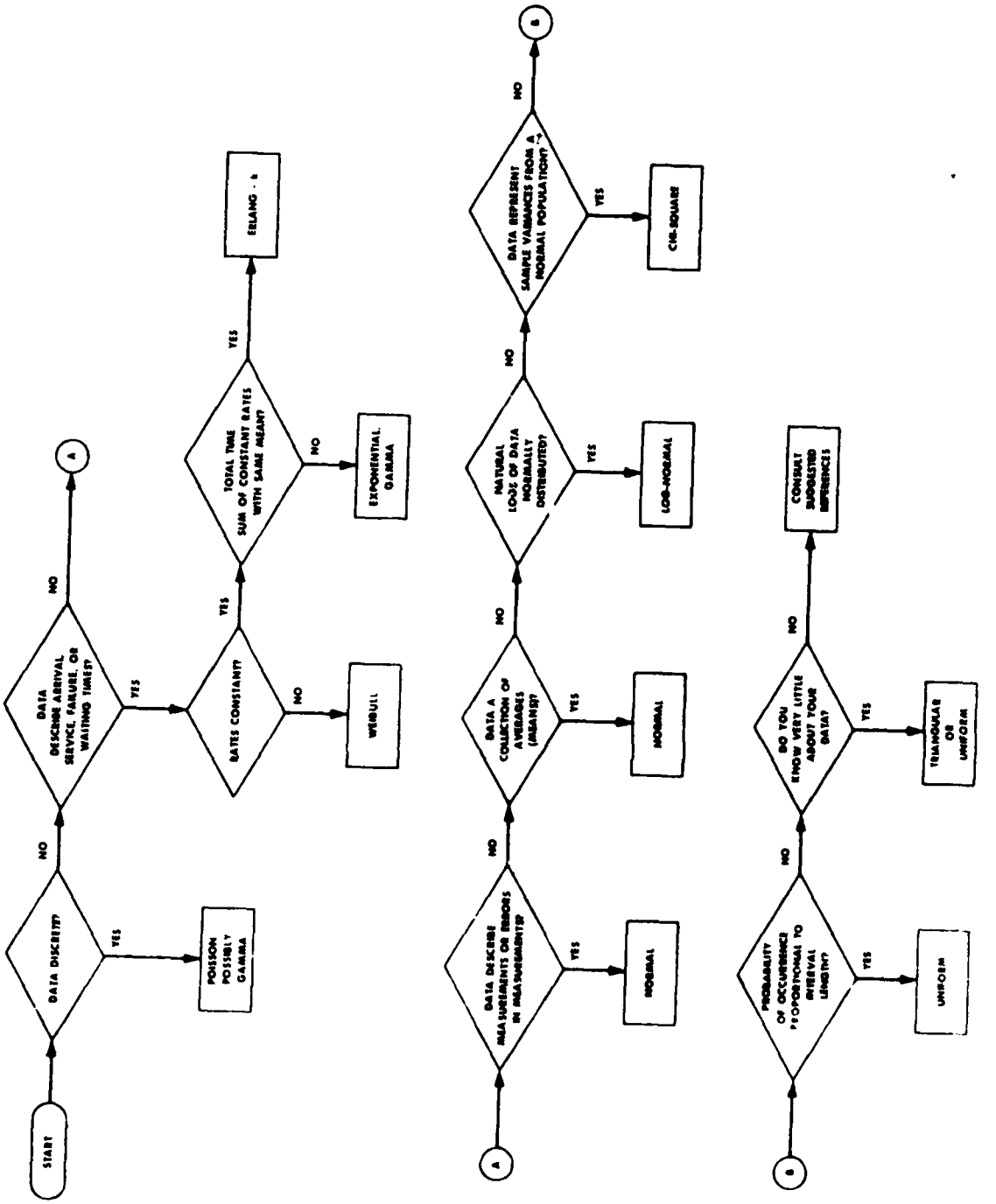


Figure 12

Distribution Selection Flowchart

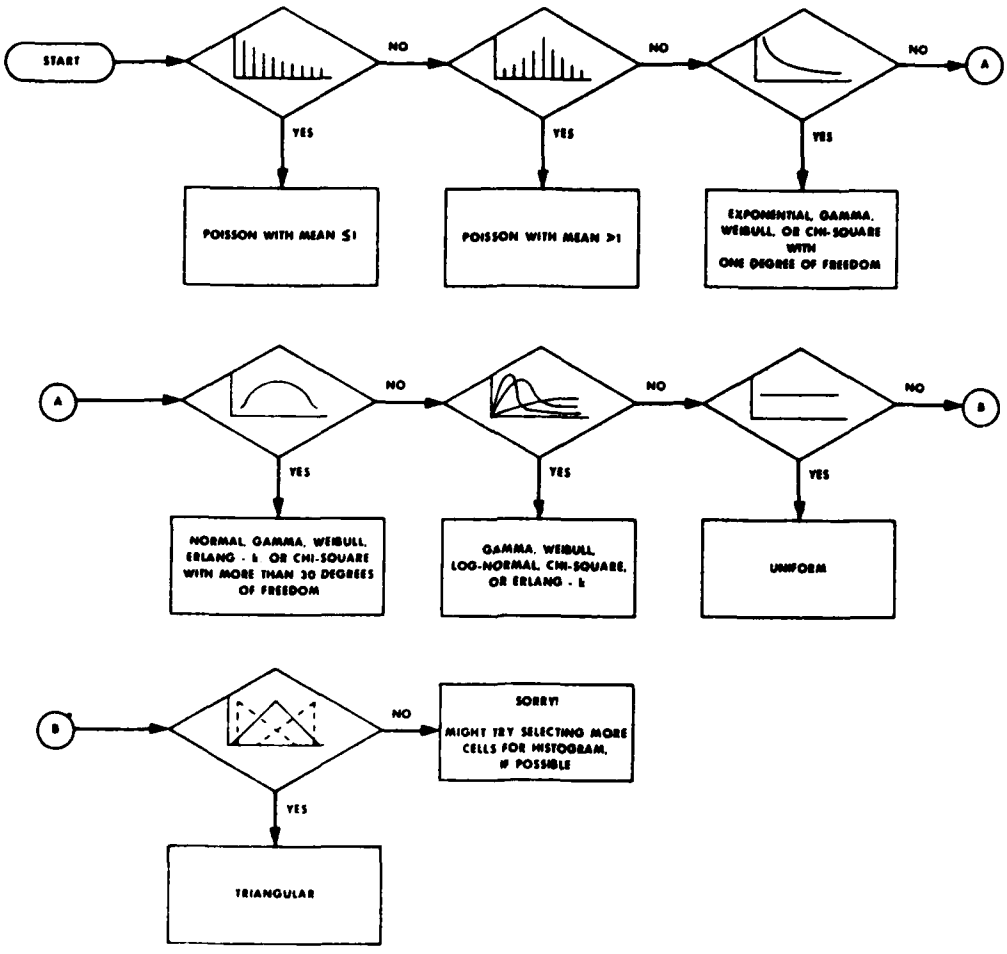


Figure 13

Pictorial Flowchart for Distribution Selection

$$V = s/\bar{x},$$

where s is the sample standard deviation and $\bar{x}$ is the sample mean. It provides a dimensionless measure of dispersion and is sometimes useful in comparing populations with unlike units of measure. For example, comparing the variations of weights of blueberries with variations of pumpkin weights is difficult intuitively to understand unless the units of measure are removed. Some theoretical distributions have constant coefficients of variation, skewness, and kurtosis. Therefore, these values are sometimes useful in selecting plausible distributions against which to apply goodness of fit tests.

B. Poisson Distribution

The Poisson distribution is a discrete distribution. It is used to describe events which are countable, rather than measurable. The Poisson distribution is the distribution of rare events. Mathematically, it is defined as:

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \text{ where } x = 0,1,2,\ldots,\infty, \lambda>0, \text{ and}$$

$$f(x;\lambda) = 0, \text{ otherwise.}$$

The Poisson distribution is characterized by only a single parameter λ which is the mean of the distribution. It represents the average number of times an event occurs in a given space or time. For example, $\lambda = 2.8$ might represent the average number of customers arriving at a stadium box office every minute or it might possibly represent the average number of typographical errors on a typed page. A Poisson variate x can only assume integer values, but the Poisson parameter λ is not restricted to being an integer value. However, λ is restricted to being a value greater than zero. This restriction is logical because λ always represents the number of occurrences--a positive quantity.

One interesting feature of the Poisson distribution is that its variance (σ^2) also equals its mean, λ. In the example above where λ=2.8, the variance of the arrivals (or the errors) would also be 2.8. The standard deviation would be $\lambda^{\frac{1}{2}}$ or $\sqrt{2.8} = 1.67$. Harnett (8:125-126) gives the proof that both the mean and the variance of the Poisson distribution equal λ.

The coefficient of skewness is equal to $\lambda^{-\frac{1}{2}}$. Because λ is always greater than zero, the Poisson distribution always has a positive coefficient of skewness denoting a skew to the right. However, as λ increases the skewness coefficient decreases and for $\lambda>9$, the normal distribution with mean and variance equal to λ may be used to approximate the Poisson distribution (Hastings, 9:112).

The coefficient of kurtosis is calculated by $3+(1/\lambda)$. As λ increases, the kurtosis value approaches 3, the standard of the normal distribution. The coefficient of variation is simply $\lambda^{-\frac{1}{2}}$, the same as the coefficient of skewness.

The shape of the graph for the Poisson distribution is controlled by the value of λ. For small values of λ, where $\lambda \leq 1$, the graph is very skewed to the right. As the value of λ increases, the graph begins to appear more symmetrical. Three graphs are given in Figure 14 to show the effect of an increase in λ on the shape of the distribution.

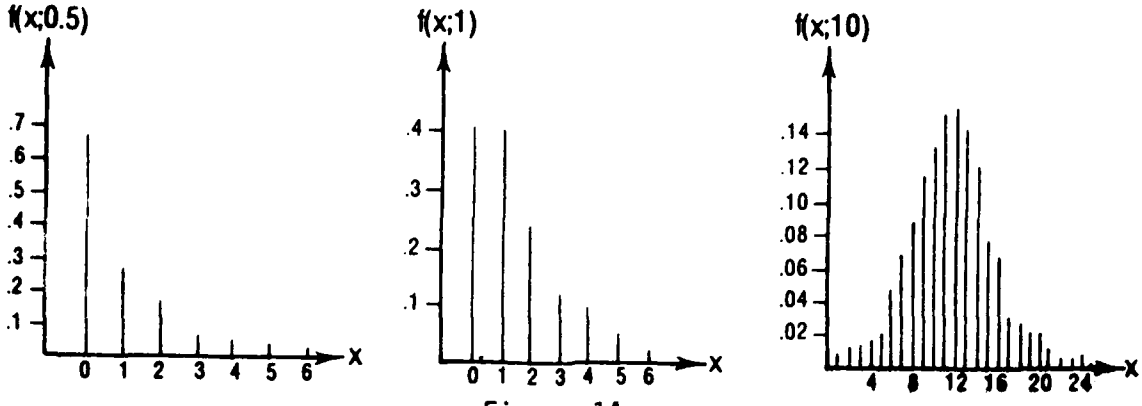


Figure 14
Poisson Distribution (Notice Scale Changes)

The uses for the Poisson distribution are many. Generally, it describes events in time (number of customers arriving at a bank teller's window) and events in space (number of flaws on a wood surface). There are four basic assumptions which must be met before the Poisson distribution can be assumed to represent empirical data.

> First, it must be possible to subdivide the time interval being used into a large number of small subintervals in such a manner that the probability of an occurrence in each of these subintervals is very small. Second, the probability of an occurrence in each of the subintervals must remain constant throughout the time period being considered. Third, the probability of two or more occurrences in each subinterval must be small enough to be ignored. And, finally, an occurrence (or nonoccurrence) in one interval must not affect the occurrence (or nonoccurrence) in any other subinterval-i.e., the occurrences must be independent (Harnett, 8:120).

C. Exponential Distribution

The exponential distribution is a continuous distribution. Its density function is

$$f(x;\beta) = \frac{e^{-x/\beta}}{\beta},$$

where:

β is the reciprocal of the average number of occurrences (arrivals, successes) during a time interval and $\beta > 0$.

The exponential distribution, as the Poisson distribution, has a single parameter. The Poisson distribution's single parameter λ represents the mean number of arrivals during a selected time interval. The exponential distribution's single parameter β represents the mean arrival rate (or interarrival rate) per the same time period. For example, if customers arrive at a checkout counter according to a Poisson distribution with $\lambda = 2.8$ customers per minute, then the exponential mean arrival rate $\beta = 1/\lambda$ or $1/2.8$ which equals about 0.3571 minutes (or one customer about every 21 seconds).

The exponential distribution has the interesting property that its mean equals its standard deviation. The coefficients of skewness and kurtosis for the exponential distribution are both independent of its parameter β. The measure of skewness always equals 2 which shows the exponential distribution as skewed to the right. The measure of kurtosis is always 9, showing that the exponential distribution is considerably more peaked than the normal distribution whose kurtosis coefficient is 3. The coefficient of variation is the constant value 1, which is logical since the mean equals the standard deviation for the exponential distribution.

The graph of the exponential distribution follows the same basic shape for all values of β with its y-intercept equal to $1/\beta$. The x-axis usually represents time while the y-axis represents the number of events. A representative exponential curve, shown by Figure 15, has a mean interarrival time of $\beta = 0.05$.

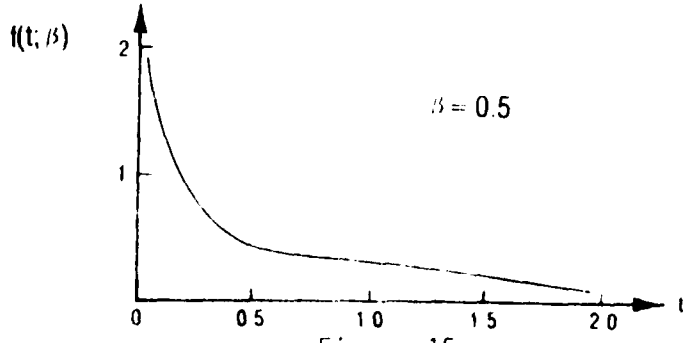


Figure 15
Exponential Distribution

The exponential distribution is considered "one of the best known, most useful, and most thoroughly explored failure distributions . . . applicable to many types of component failures, especially in electrical systems" (Tsokos, 19:186). In addition to describing failure rates, it is often used to describe service times and arrival rates. In using the exponential distribution to describe service times, one major assumption should be adherred to: the interarrival rate or the service rate should be relatively short (Harnett, 8:139). For example, the length of telephone conversations is usually considered to be exponentially distributed and satisfies the "relatively-short" assumption. Martin (11:67) gives seven specific examples of occurrences which have been shown to follow exponential distributions.

1. Lives of electron tubes.
2. Time intervals between successive breakdowns of electronic systems.
3. Life testing in many life distributions.
4. Purely random failure patterns.
5. Pure death processes (fiber failure).
6. Failure of complex mechanisms.
7. Target noise and receiver noise after square law rectification.

D. Normal Distribution

Many statistical superlatives are used to describe the normal distribution: most valuable, most popular, best known, most widely used, most important, and well researched. The normal distribution is a continuous distribution whose equation for its probability density function is

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

where

$-\infty<x<\infty$,
μ is the mean of the distribution,
σ^2 is the variance of the distribution, and
σ is the standard deviation of the distribution.

The characteristic bell shape of the normal curve is a function of its two parameters, μ and σ. The mean μ gives the location on the x-axis of the center of the curve. The standard deviation σ describes how dispersed or varied the random variates are. A large standard deviation means a short, wide bell-shaped curve while a small standard deviation describes a tall, narrow bell-shaped curve.

The coefficients of skewness and kurtosis are mentioned in detail in Section II of this report under the description of the moments test for normality. Both measures are independent of the values of the parameters of the normal distribution. The coefficient of skewness for a normal curve is 0 denoting a symmetrical curve. The coefficient of kurtosis is 3 which describes a mesokurtic shape. The coefficient of variation, by definition, is the standard deviation σ divided by the mean μ.

The shape of the normal curve is always symmetric and always bell-shaped. However, a change in the location parameter μ can change the position the curve is centered about on the x-axis. A change in the scale parameter σ effects the amount of spread the sides of the curve exhibit in relation to the curve's center point. The changes in the shape which are a result of changes in μ (σ held constant) are shown in Figure 16. Figure 17 shows the effects of changing σ while μ is held constant.

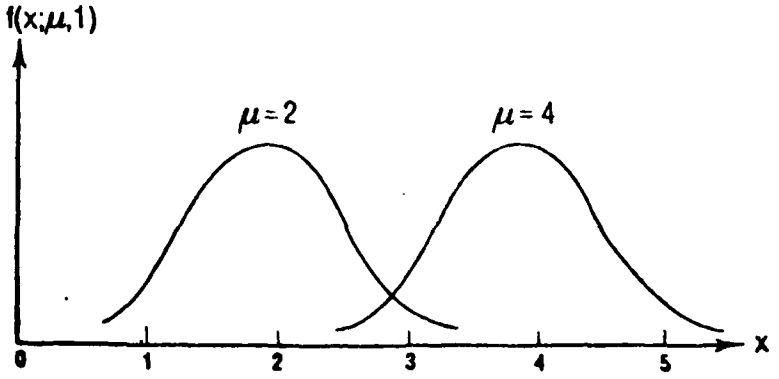


Figure 16
Normal Distribution (σ = 1)

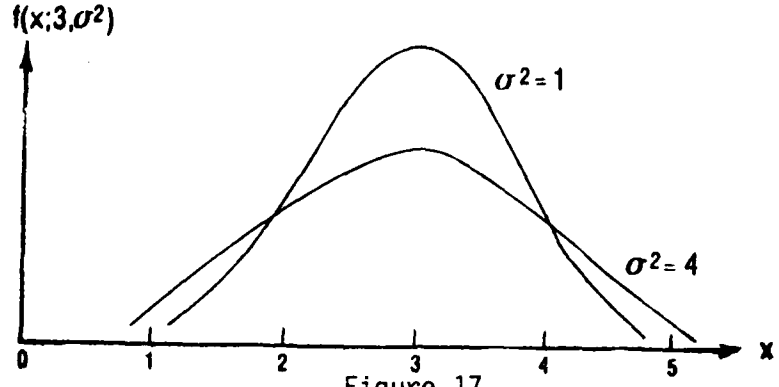


Figure 17
Normal Distribution (μ = 3)

There are five primary reasons why the normal distribution is the most popular theoretical distribution in the study of statistics. First, the normal distribution is easy to use because its probability tables have been extensively and accurately developed. Secondly, many random variables do approach normal distributions. For example, "heights of men, lengths of ears of corn, and, more generally, many linear dimensions . . ." (Snedecor, 18:35). A third reason is that a simple transformation of random variates (whose distributions are not normal) may approximate a normal distribution. The logarithm of a random variate is an example of such a transformation. The Central Limit Theorem is the fourth and probably the most important use of the normal distribution. The Central Limit Theorem states that the distribution of sample means is approximately normal as N>30. The original form of the distribution from which the samples were taken does not matter; the distribution of the sample means would still approach a normal distribution. Therefore, the importance of this theorem is supported by the evidence that many researchers are interested in averages for their data--the average income for a computer systems analyst or the average life of a particular variety of light bulbs. The fifth reason for the extensive use of the normal distribution is that it serves adequately well as an approximation for some non-normal populations (Snedecor, 18:35). Several of the distributions which are sometimes approximated by the normal distribution "are the binomial, the hypergeometric, the Poisson, and the gamma" (Tsokos, 19:160).

Several specific uses of the normal distribution are included to aid in the identification of characteristic environments of sample observations. "Such diverse characteristics as sleeve lengths for adult males, intelligence test scores for school children, errors made by rats learning a maze, and achievement in statistics courses seem to follow the 'normal' distribution" (Klugh, 10:49).

Because the normal distribution is so well known and relatively easy to use, it is often misused to describe phenomena which are not normally distributed. For this reason, the moments test for normality is included and described in Section II of this report.

E. Log-normal Distribution

If Y is a random variate that is normally distributed with a mean μ_y and a variance σ_y^2 and if $Y = \ln X$, then X follows a log-normal distribution. The log-normal distribution is a continuous distribution. If $X = e^y$, where $Y \sim N(\mu_y, \sigma_y^2)$, then the density function of the log-normal distribution is

$$f(x;\mu_y,\sigma_y^2) = \frac{1}{x\sqrt{2\pi\sigma_y^2}} e^{\left[-\frac{(\ln x - \mu_y)^2}{2}\right]},$$

where $x>0$, $-\infty<y<\infty$, and $\sigma_y^2>0$. The density function for Y, the normal variate is

$$f(y;\mu_y,\sigma_y^2) = \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{(y-\mu_y)}{2\sigma_y^2}},$$

where $-\infty<y<\infty$, $-\infty<\mu_y<\infty$, and $\sigma_y^2>0$.

The two parameters used to describe the log-normal distribution are actually the mean and variance of its related normal distribution. The log-normal distribution may be generalized to include a location parameter, thus becoming a three-parameter distribution. It is treated as a two-parameter distribution in this work. For the log-normal distribution, μ_y and σ_y^2 are called the scale and shape parameters, respectively, while they are the location and scale parameters of the normal distribution. The relationship between the parameters of the normal distribution (Y) and the mean and standard deviation of the log-normal distribution (X) is

$$\mu_y = \ln\left(\sqrt{\frac{\mu_x^4}{\mu_x^2+\sigma_x^2}}\right) \text{ and } \sigma_y^2 = \ln\left(\sqrt{\frac{\mu_x^2+\sigma_x^2}{\mu_x^2}}\right) \quad (19:174).$$

The coefficient of skewness for the log-normal distribution can be calculated by

$$C_s = (e^{\sigma_y^2} + 2)(e^{\sigma_y^2} - 1)^{\frac{1}{2}}.$$

The coefficient of kurtosis can be calculated by the formula

$$C_k = (e^{\sigma_y^2})^4 + 2(e^{\sigma_y^2})^3 + 3(e^{\sigma_y^2})^2 - 3.$$

The coefficient of variation is $(e^{\sigma_y^2} - 1)^{\frac{1}{2}}$.

The log-normal distribution is represented by a family of curves, all of which are skewed to the right. The greater the value of the parameter σ_y^2, the more pronounced the skewness of the graph of the density function. Figure 18 illustrates the effects of changes in σ_y^2 while μ_y is held constant.

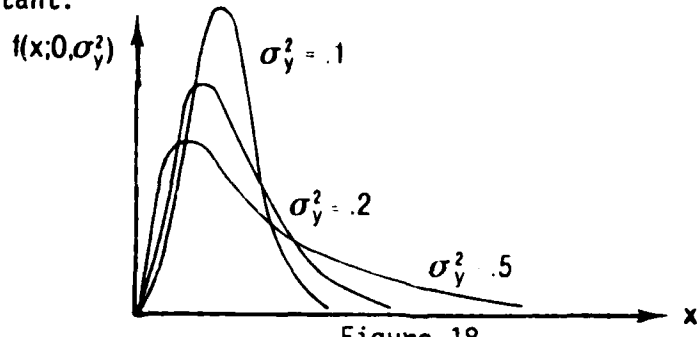


Figure 18
Log-normal Distribution ($\mu_y = 0$)

Figure 19 demonstrates the effects of changing μ_y, the location parameter for Y, while holding σ_y^2 constant.

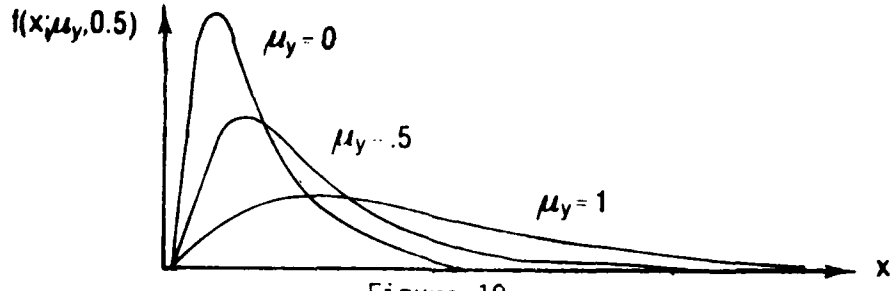


Figure 19
Log-normal Distribution ($\sigma_y^2 = 0.5$)

Random variates suspected to follow a log-normal distribution must always satisfy one requirement--they must be positive values only. This distribution is not defined for negative values.

F. Gamma Distribution

The gamma distribution is a family of continuous distributions with two parameters: α, the shape parameter, and β, the scale parameter. The probability density function is

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}},$$

where $x \geq 0$, $\alpha > 0$, and $\beta > 0$. $\Gamma(\alpha)$ is the gamma function $\Gamma(\alpha) = \int_{-\infty}^{\infty} x^{\alpha-1} e^{-x} dx$, $\Gamma(\alpha) = (\alpha - 1)!$, where α is a positive integer. "The gamma distribution is the appropriate model for the time required for a total of exactly . . . α independent events to take place if events occur at a constant rate . . . β " (Hahn, 7:83).

The mean of the gamma distribution is $\alpha\beta$ and the variance is $\alpha\beta^2$. The measures of skewness and kurtosis for the gamma distribution rely on the α parameter. The coefficient of skewness is $2\alpha^{-\frac{1}{2}}$. The kurtosis coefficient is computed by $3 + 6/\alpha$. The gamma distribution is always skewed to the right, but the skewness decreases as α increases. The coefficient of variation is $\alpha^{-\frac{1}{2}}$.

When the shape parameter equals 1, the exponential member of the gamma distribution drops out with density function

$$f(x;\beta) = \frac{e^{-x/\beta}}{\beta},$$

where $x \geq 0$ and $\beta > 0$. When α represents only positive integers, the Erlang-k distribution is described and the α parameter is referred to as k. The third distinctive member of the gamma family is the chi-square distribution and it occurs when $\beta = 2$ and $\alpha = r/2$, where r is a positive integer. A further description of the chi-square case is available in Hahn and Shapiro's book (7) and under the appropriate section of this report.

The graph of the gamma distribution takes many shapes, which is the reason for its versatility. Its curves are often described as unimodal and reverse J-shaped. A unimodal curve has a single rounded peak. A J-shaped curve rises to a cusp at one end and falls off rapidly at the other end. These adjectives become apparent when one studies the various shapes of the gamma distribution.

The shape parameter α is varied in Figure 20 while β, the scale parameter, is held constant at 1. As α approaches 0, $f(x;\alpha,\beta)$ also approaches 0 for $\alpha,\beta > 0$. In fact, for $0 < \alpha < 1$, $\Gamma(\alpha) = (\alpha - 1)!$ is negative and $f(x;\alpha,\beta)$ becomes infinite as α approaches 0. For cases of $\alpha \geq 1$, the gamma distribution starts at the origin, has its maximum value at $x = \alpha\beta$, and then falls off. "In all cases, the curve approaches the x-axis asymptotically for large values of x" (Wadsworth, 20:110). The shape of the gamma distribution can even resemble that of the normal when $\beta = 1$, and α becomes very large (Shannon, 17:366).

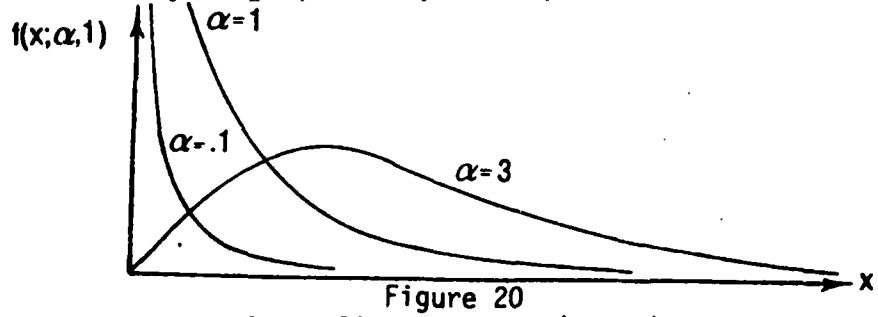


Figure 20
Gamma Distributions ($\beta = 1$)

Figure 21 illustrates the changes which occur in the shape of the gamma distribution when α is held constant and β is varied.

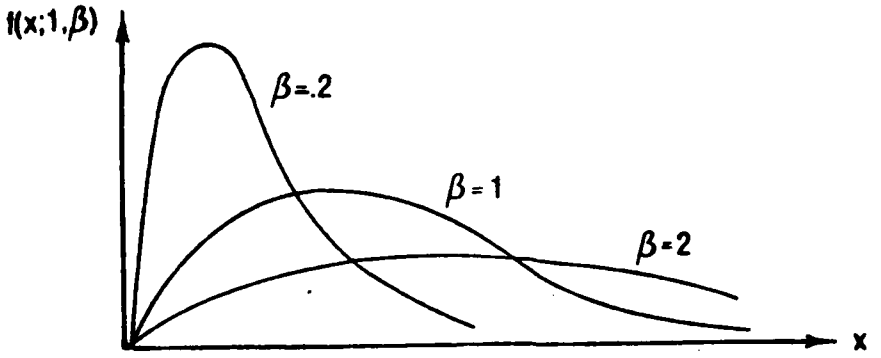


Figure 21
Gamma Distributions (α = 1)

The gamma distribution is used in applications for "determining the probability that t or less time will be required to obtain a success" (Clark, 3:253). In other words, it is often used as a model for waiting times. An example in life testing problems might be the waiting time until the failure of some component. Shannon goes so far as to say, "If the variables from some random phenomenon cannot assume negative values and generally follow a unimodal distribution, then the chances are excellent that a member of the gamma family can adequately simulate the phenomenon" (17:363-364).

G. Erlang-k Distribution

The Erlang-k distribution is a continuous distribution with the same two parameters as the gamma distribution: α, the shape parameter, and β, the scale parameter. The shape parameter, however, must be a positive integer for the Erlang-k distribution. The probability density function is the same as that of the gamma distribution:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}},$$

where α and $\beta>0$ and $x\geq 0$. The gamma function is evaluated for α by $\Gamma(\alpha) = (\alpha - 1)!$, if α is a positive integer.

The mean, variance, coefficients of skewness, kurtosis, and variation are the same as those of the gamma distribution and can be obtained from the portion of this section on the gamma distribution. The only difference between the two is the requirement that the shape parameter of the Erlang-k distribution be a positive integer. When the shape parameter equals 1, the Erlang-k distribution reduces to the exponential distribution.

The coefficient of variation for the exponential, Erlang-k, and gamma distributions is always $1/\sqrt{\alpha}$, where α is the shape parameter. For the special case of the exponential when α is 1, the coefficient of

variation is 1. As the shape parameter increases, the coefficient of variation decreases which implies that Erlang-k data tends to cluster more closely toward the mean than does exponential data. In other words, Erlang-k observations would be less likely to have as many high and low values as would exponential data.

The characteristic shapes of the Erlang-k distribution are identical to those diagrammed previously in this section for the gamma distribution.

H. Chi-square Distribution

The chi-square distribution is a continuous, single-parameter distribution defined only for values of the random variable greater than or equal to zero. Its single parameter ν is the number of degrees of freedom. The degrees of freedom parameter is usually defined as the number of sample observations that are independent to vary after certain limitations on the data have been taken into account. An example of a limitation is using the sample mean $\bar{x}$ as an estimate for the unknown population mean. The $\bar{x}$ limitation would remove one degree of freedom.

The chi-square distribution is really a special member of the gamma family of distribution. The chi-square distribution occurs when the gamma parameter $\beta = \frac{1}{2}$ and $\alpha = k/2$, where k is a positive integer. (The parameter k is the degrees of freedom parameter.) The density function for the chi-square distribution is

$$f(x;\nu) = \frac{x^{(\nu/2-1)} e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)},$$

where $x>0$, $\Gamma(\nu/2) = (\nu/2 - 1)!$, and $\nu/2$ is an integer.

The chi-square distribution has numerous important properties and relationships to variates of other distributions. Hastings and Peacock describe the chi-square variate relationships to the gamma, F, Student's t, Poisson, and normal variates (9:46-50). Two of the more interesting properties of the chi-square distribution are:

1. if X is a standard normal variate ($X \sim N(0,1)$), then X^2 is a chi-square variate with one degree of freedom ($X^2 \sim \chi^2(1)$), and
2. if X_1 is $\chi^2(n)$ and X_2 is $\chi^2(m)$ and X_1 and X_2 are independent, then $X_1 + X_2$ is $\chi^2(n+m)$.

The chi-square distribution is actually defined as the sum of the squares of n independent standard normal variates. Therefore, the first property stated above is simply the definition of the chi-square distribution.

The mean and variance of the chi-square distribution are very simple functions of the parameter ν. The mean μ is equal to ν and the variance σ^2 is equal to 2ν. Therefore, the coefficient of variation is computed simply by $(2/\nu)^{\frac{1}{2}}$ which gives it a range between 0 and $\sqrt{2}$. The coefficient of skewness is computed by $2^{3/2}\nu^{-\frac{1}{2}}$ and the kurtosis coefficient is $3 + 12/\nu$. Since degrees of freedom ν is always greater than

or equal to one, the skewness coefficient is always positive, denoting skew to the right. The kurtosis coefficient is always greater than three, denoting leptokurtic shape approaching mesokurtic as $\nu \to \infty$. The chi-square distribution approaches the normal distribution for $\nu > 30$. For larger values of ν, it is easy to see that the coefficients of skewness and kurtosis approach the values which are characteristic of the normal distribution.

Figure 22 illustrates the various shapes of the chi-square distribution caused by a change in its single parameter ν. The chi-square curve is always skewed to the right, but approaches the symmetry of the normal curve as ν becomes large.

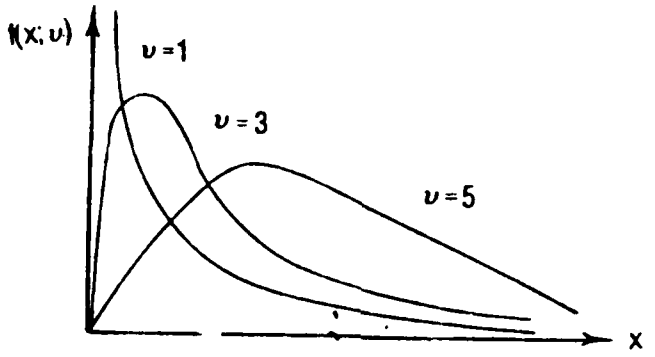


Figure 22
Chi-square Distribution

The chi-square distribution is most valuable in testing hypotheses on the variances of normal populations. It is also used as the limiting distribution in the chi-square goodness of fit test and it is important in conducting tests of independence.

1. Triangular Distribution

The triangular distribution is a three-parameter continuous distribution. The three parameters are the minimum value (X_{min}) the distribution can attain, the maximum value (X_{max}) the distribution reaches, and the most probable value (X_{most}) which occurs over the interval between X_{min} and X_{max}. The probability density function for the triangular distribution has two parts, defined with respect to the X_{most} position:

$$f(x;X_{min},X_{max},X_{most}) = \begin{cases} \dfrac{2(x-X_{min})}{(X_{most}-X_{min})(X_{max}-X_{min})} & \text{for } x \leq X_{most} \\ \dfrac{2(X_{max}-x)}{(X_{max}-X_{most})(X_{max}-X_{min})} & \text{for } x > X_{most} \end{cases}$$

The triangular distribution looks like its name indicates it does--a triangle. Its general form is shown below.

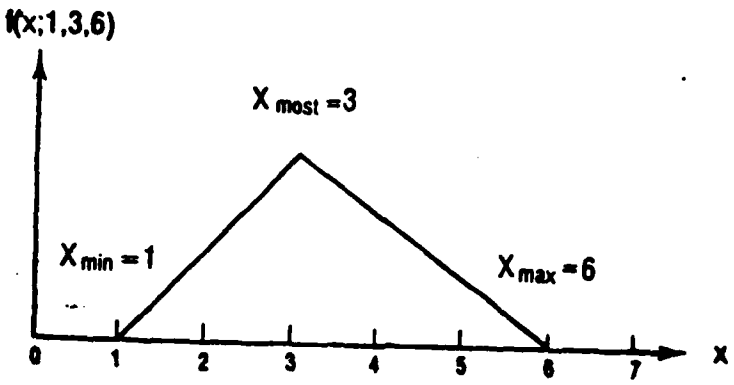


Figure 23
Triangular Distribution

X_{min} is the minimum allowable value for x; X_{max}, the maximum allowable value for x; and, X_{most} is the most probable value which can occur over the defined range.

Three special cases of the triangular distribution exist: a right triangle is defined when $X_{most} = X_{max}$; a left triangle is defined when $X_{most} = X_{min}$; and a pyramidal triangle is defined when $X_{most} = 1/2(X_{max} - X_{min})$.

The triangular distribution is used primarily by the researcher who knows the minimum value of his distribution, the maximum value of his distribution, and the most probable value of his distribution.

J. Uniform Distribution

The uniform distribution has two parameters, usually referred to as a (or A) and b (or B), or sometimes as α and β. The parameters define the endpoints of the interval over which the distribution is defined and a is always less than b. The probability density function for the uniform distribution is

$$f(x;a,b) = \frac{1}{b - a},$$

where $a \leq x \leq b$ and the function is 0 elsewhere. The mean is easily calculated by

$$\mu = (a + b)/2$$

and the variance is almost as simple with

$$\sigma^2 = (b - a)^2/12.$$

The measure of skewness is 0 which is obvious from a graph of the uniform distribution. The measure of kurtosis is constant at 1.8 which implies flatness--again, obvious from a graph. The coefficient of variation can be computed by $(b - a)/\sqrt{3}(a + b)$.

The graph of the uniform distribution is simply a straight line between the endpoints a and b of the interval over which it is defined. The y-coordinate of the line is $1/(b - a)$ as can be seen in Figure 24.

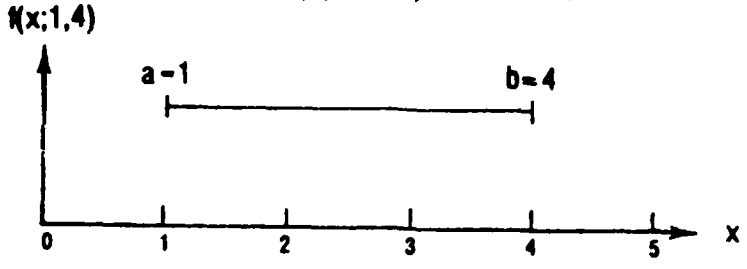


Figure 24
Uniform Distribution

The uniform distribution is used when there is an equal probability of each defined value being selected and that probability is directly proportional to the length of the interval. In nature, variates are not usually uniformly distributed. However, whenever completely random choices can be made from among a set of alternatives, the uniform distribution is often used to describe this situation.

K. Weibull Distribution

The Weibull distribution is a continuous distribution. It is sometimes described as a two-parameter distribution and sometimes as a three-parameter distribution. The third parameter, when used, is a location parameter and permits the introduction of an arbitrary origin. The two-parameter distribution is described here (location parameter is equal to zero). In applications language, the location parameter represents the initial period of time before any failure takes place (7:110-111). The word "failure" suggests that the Weibull distribution is another model used extensively in reliability studies.

The two-parameter probability density function for the Weibull distribution is

$$f(x;\alpha,\beta) = \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}}$$

where $x>0$, $\alpha>0$, and $\beta>0$. For the Weibull distribution, α is the scale parameter and β is the shape parameter.

The mean of the Weibull distribution can be calculated by evaluating the integral

$$\mu = \int_0^{\infty} x\alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}} dx.$$

Miller (14:466) simplifies that integral until the equation for the mean reduces to

$$\mu = \alpha^{-1/\beta}\Gamma(1 + 1/\beta).$$

The variance of the Weibull distribution is defined as

$$\sigma^2 = \alpha^{-2/\beta}\{\Gamma(1 + 2/\beta) - [\Gamma(1 + 1/\beta)]^2\}.$$

The coefficients of skewness and kurtosis are not readily available from the referenced literature.

The Weibull distribution takes on many shapes as the values of its shape parameter β change. The Weibull curves are always skewed to the right. For values of $\beta<1$, the Weibull curves are asymptotic to both axes. When $\beta = 1$, the exponential curve appears. For $\beta>1$, a unimodal family of curves is generated, but still the curves are skewed to the right. Figure 25 illustrates shapes of the Weibull curves for the conditions of β just described.

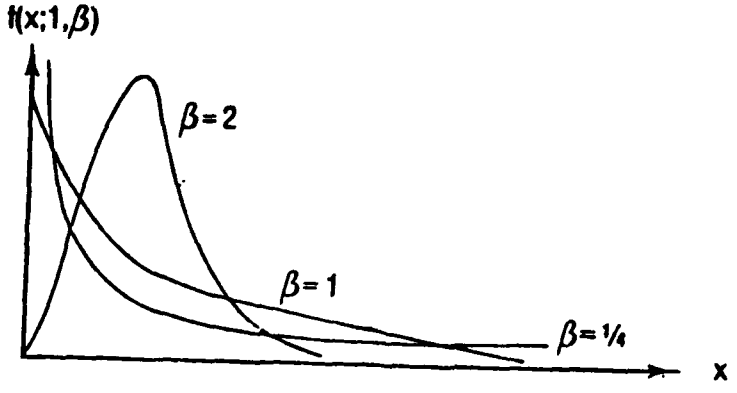


Figure 25
Weibull Distribution (α = 1)

The Weibull distribution has its primary use in reliability studies when the failure rate of the component being tested is not constant for the time frame under consideration. The exponential distribution is often used when the failure rate of a component is a constant. By contrast, Weibull probabilities describe the time to "wear-out failure rather than chance failure" (Miller, 14:465). The Weibull distribution has been used successfully to represent failure models for "electron tubes, relays, and ball bearings . . . . ." (Hahn, 7:109).

IV. SOFTWARE DESCRIPTION

A. General

The goodness of fit (GOF) program, described in this section, is designed to be simple enough to be understandable by the non-statistician and non-programmer. However, it is also designed to be comprehensive enough to give the analyst a tool that provides a first step in the characterization of his research data. The methodology of the GOF program closely follows the general steps of statistical hypothesis testing. Table 2 provides a comparison of the steps in hypothesis testing with the procedures the user must follow when running the GOF program. Once the user is familiar with the techniques of hypothesis testing, the flow of the GOF program is logical and easy to follow.

B. Capabilities

The GOF program operates in two modes: batch and on-line. It is written entirely in FORTRAN IV. Portability is an important design objective for this program. Machine-dependent characteristics and software functions are avoided in the coding of the GOF program. The program was developed on a UNIVAC 1108 and was also successfully implemented on a Xerox Sigma 9 computer.

The GOF program performs four major functions:

1. Interprets easy-to-use input instructions,
2. Prints a histogram (optional) and descriptive statistics calculated from the user's empirical data,
3. Performs selected goodness of fit test(s) to a hypothesized probability distribution, and
4. Notifies the user as to the acceptance or rejection of his test hypothesis.

The first function is performed by the GOF language translator which accepts GOF problem oriented commands and translates them into internal code for use by the computational section of the program. The GOF language is easy to master and its statements are in free form format making them simple to input. The second, third, and fourth major functions are performed by the computational section of the GOF program. That section is responsible for reading and sorting the input data, calculating observed and theoretical distribution frequencies, printing a histogram of the observed data, performing the goodness of fit mathematics, checking the critical value tables for the appropriate tests being run, and printing the results of the run. The output of each goodness of fit test is given in a blocked format with a sentence-structure synopsis of the results of the test. The output is, therefore, easy to identify and understand.

C. Options

Delayed Distribution Selection. The user of the GOF program is provided several options allowing him to tailor the operation of the program to his individual needs. If he is unsure of the distribution

Table 2

Methodology of GOF Program

Steps in Hypothesis Testing	GOF Program Procedures
1. The null and alternative hypotheses are stated.	1. The user selects the distribution to be tested. The GOF program constructs the hypotheses as: H_0: The observed data comes from the selected distribution, and H_1: The observed data does not come from the selected distribution.
2. The test statistic is selected.	2. Selecting the goodness of fit test to be run determines the test statistic to be calculated.
3. The level of significance at which to evaluate the results of the test is specified.	3. The user specifies an ALPHA value to the GOF program. That value is the probability of rejecting the null hypothesis when it is correct.
4. The experiment is performed to obtain the sample observations.	4. The sample values are input to the GOF program.
5. The test statistic is computed and the results are evaluated.	5. The goodness of fit test is performed, the results compared with the appropriate critical value, and the acceptance or rejection message is printed.

against which he wishes to test his data, he can delay the selection of a distribution until after the program has printed the histogram of his observed data. The GOF program also calculates and prints mean, variance, standard deviation, and coefficients of variation, skewness, and kurtosis for the sample data prior to the specification of the hypothesized distribution. The GOF program then returns control to the language translator and waits for the user to select the distribution to be tested. If the user is running in the batch mode, to take advantage of the delayed-distribution-selection option, he would have to execute the GOF program twice. The first run prints the histogram and descriptive sample statistics. The user studies the results of the first run and then executes the GOF program to perform the goodness of fit tests against the chosen distribution.

If the user knows the distribution against which he wishes to test the sample data, he may specify that distribution during the initial input session with the program. In this mode, the program continues uninterrupted through its computational section calculating sample statistics, testing the null hypothesis, printing test results, and printing the histogram of the sample observations.

Population Parameters. Another optional feature of the GOF program concerns the specification of theoretical population parameters. The user may input the parameters of the hypothesized distribution if they are known. If they are not known, the GOF program estimates them from the sample data. The null and alternative hypotheses are printed with the distribution parameters labeled either "theoretical" or "estimated" according to the method selected by the user.

Various Input Formats. Many program options allow versatility for the input of sample observations. The GOF program accepts either individual sample observations or grouped data. Grouped data is input by supplying the class lower and upper boundaries and the absolute class frequencies. Empirical data values, whether grouped or ungrouped, may be input under one of three available methods: (1) free field format of integer, floating, or scientific notation (E-formatted) values; (2) user supplied format; or (3) program default format of 8F10.5. Empirical data may be read from the same source (logical unit) as the language commands or it may be read from an alternate device. The number of input data values may be specified or the GOF program can be instructed to count the number of data items read before a terminal value is encountered. The terminal value may be user specified or a program default value of 999. The terminal value, whether specified or default, must be followed by a system end-of-file designation. Appendix B of this report illustrates run streams for the UNIVAC 1108. The job control statements necessary to implement the alternate device option are included.

Miscellaneous Options. Other program options control the types of calculations to be performed and the quantity of output. The printing of the histogram, which may be time consuming on slow speed terminals, can be omitted at the user's discretion. The user may also specify the number of cells he wishes the sample data to be grouped into initially, or he may let the program automatically group it into 15 cells. The user may decide to have the GOF program calculate either biased

or unbiased coefficients of skewness and kurtosis. Section II of this report gives the equations used in both methods. During the operation of the GOF program in the batch mode, numerous intermediate calculation values are printed. These values are omitted during on-line execution because of the time and extra width output fields required to print them. These options and others are explained in more detail under the part of this section which describes the individual commands.

The GOF program gives the user the results of each goodness of fit test it is requested to run. These results are evaluated internally by the program at the 95% and 99% levels of significance for all but one of the tests. The results of the moments test can only be evaluated by the GOF program at the 0.05 level of significance. Prior to running a test, the user selects an alpha value of 0.01 (for 99%) or 0.05 (for 95%). If neither value is specified, the program defaults to an alpha value of 0.05.* After each selected goodness of fit test is run, the program checks internally stored critical value tables for the appropriate value of alpha and prints the results of the test. Critical values, computed value of the test statistic, value of alpha, and the number of degrees of freedom (where appropriate) are all printed for both on-line and batch versions. This internal evaluation feature eliminates the necessity of the user's bringing extra material to a GOF session. It also spares the occasional user the additional burden of remembering how to use the various critical value tables. For values of alpha other than 0.01 and 0.05 and for tests whose number of observations or degrees of freedom exceed those available from the GOF internal tables, the user must refer to other critical value sources to determine the results of his test.

Because the GOF program runs in both batch and on-line modes, the user should specify through the command language if he is running in the on-line mode. The primary differences between the two modes of operation are the quantity and spacing of the output from the GOF program. Output for the on-line mode is restricted to 64 horizontal print poisitions while batch output extends through 132 print positions. Because of the unpredictable nature of data related error messages and optional feature messages, no control is maintained within the program on vertical output spacing.

D. Macro Flowchart

Figure 26 is the macro flowchart of the GOF program. The narrative which follows is related to the flowchart by the block numbers which appear to the upper left of each major block in the flowchart. This description is primarily for the on-line version of the GOF program. Any significant differences of the batch version are also mentioned.

---

*The user of the GOF program may always evaluate the results of a goodness of fit test at any level of significance he wishes by referring to appropriate critical value tables. The value of the test statistic and the number of degrees of freedom or sample size are printed for every test executed. Therefore, the user is not restricted to the 0.01 and 0.05 levels of significance automatically available from the GOF program.

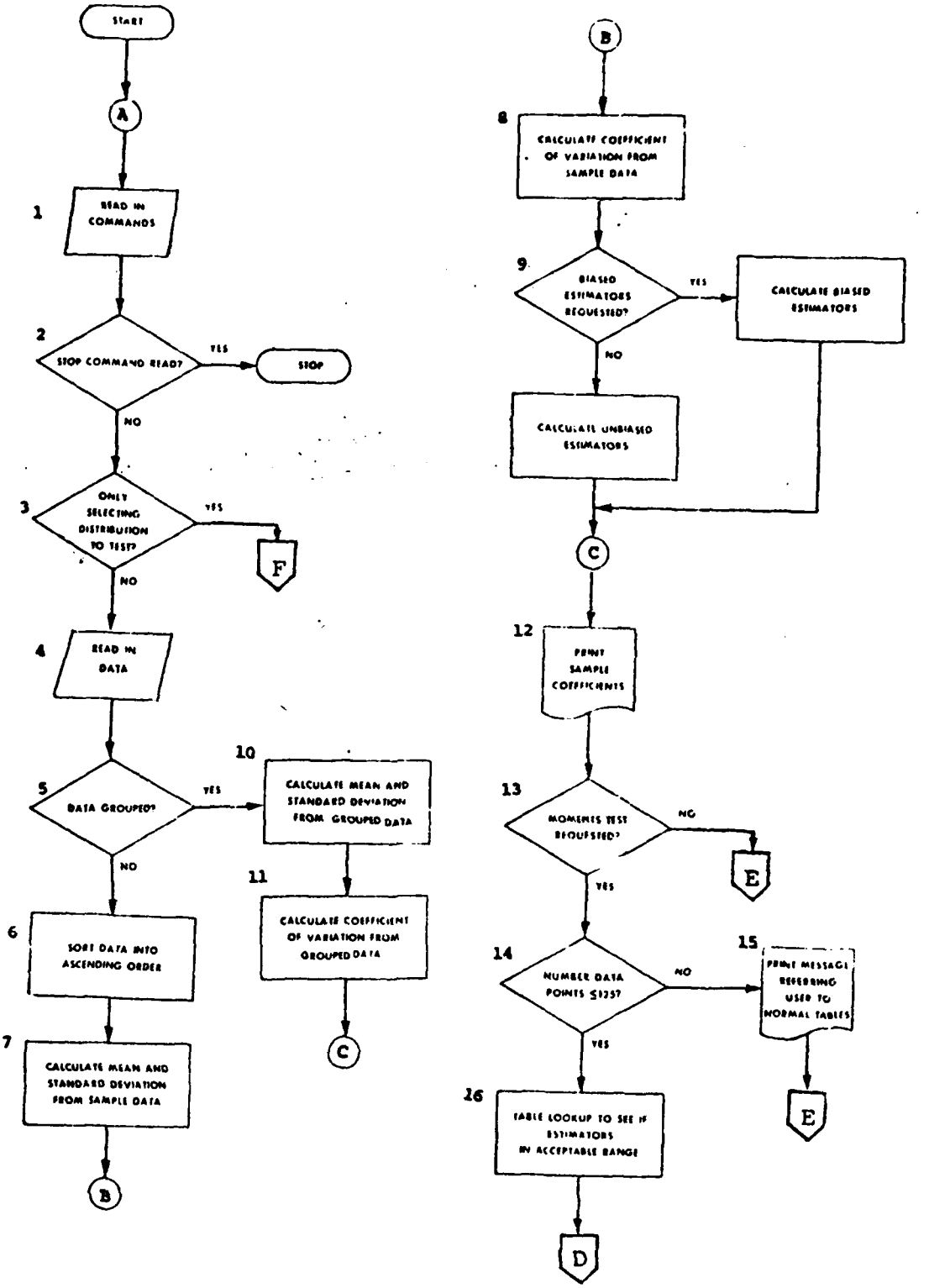


Figure 26

GOF Program Macro Flowchart

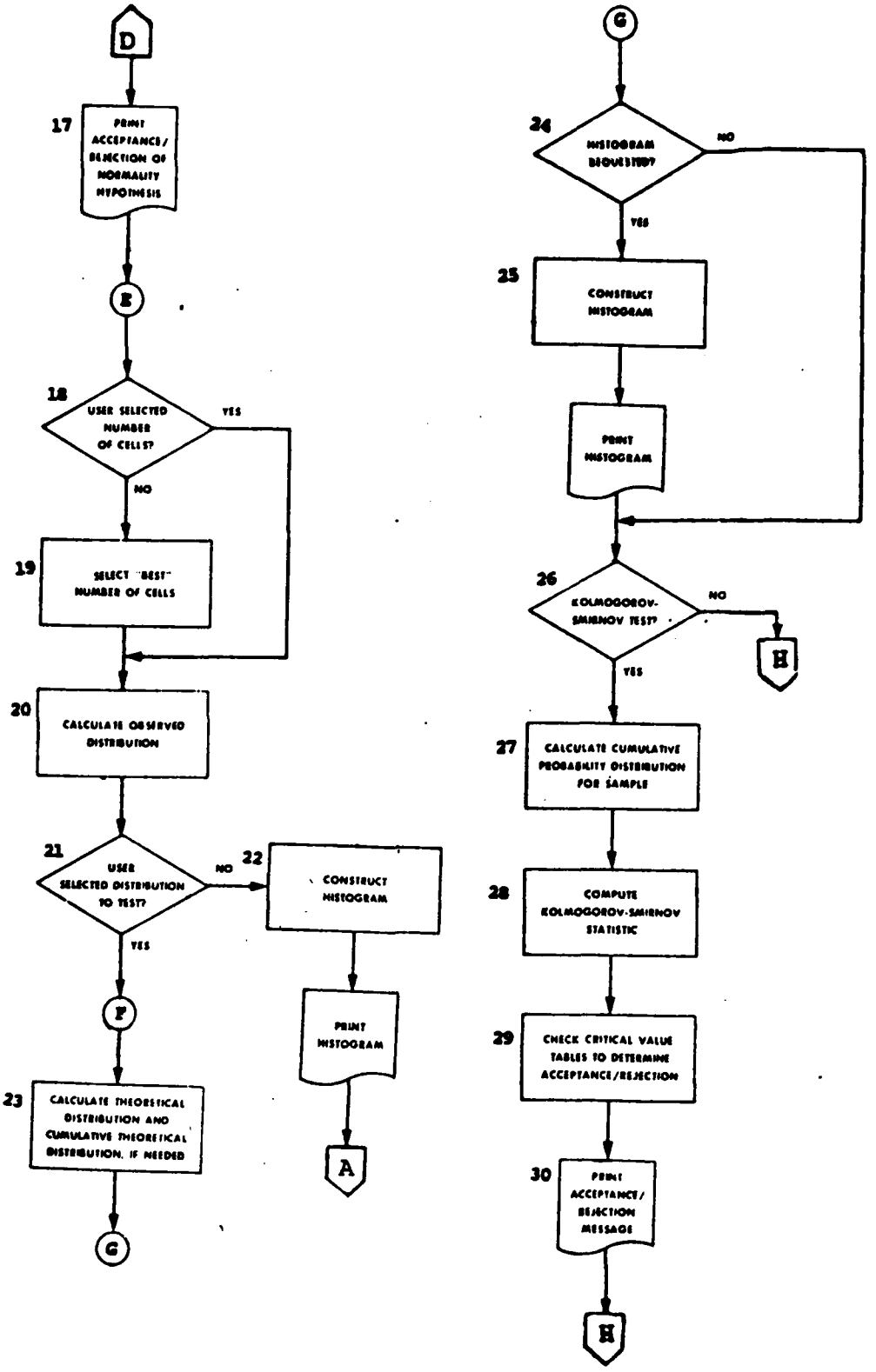


Figure 26 (continued)

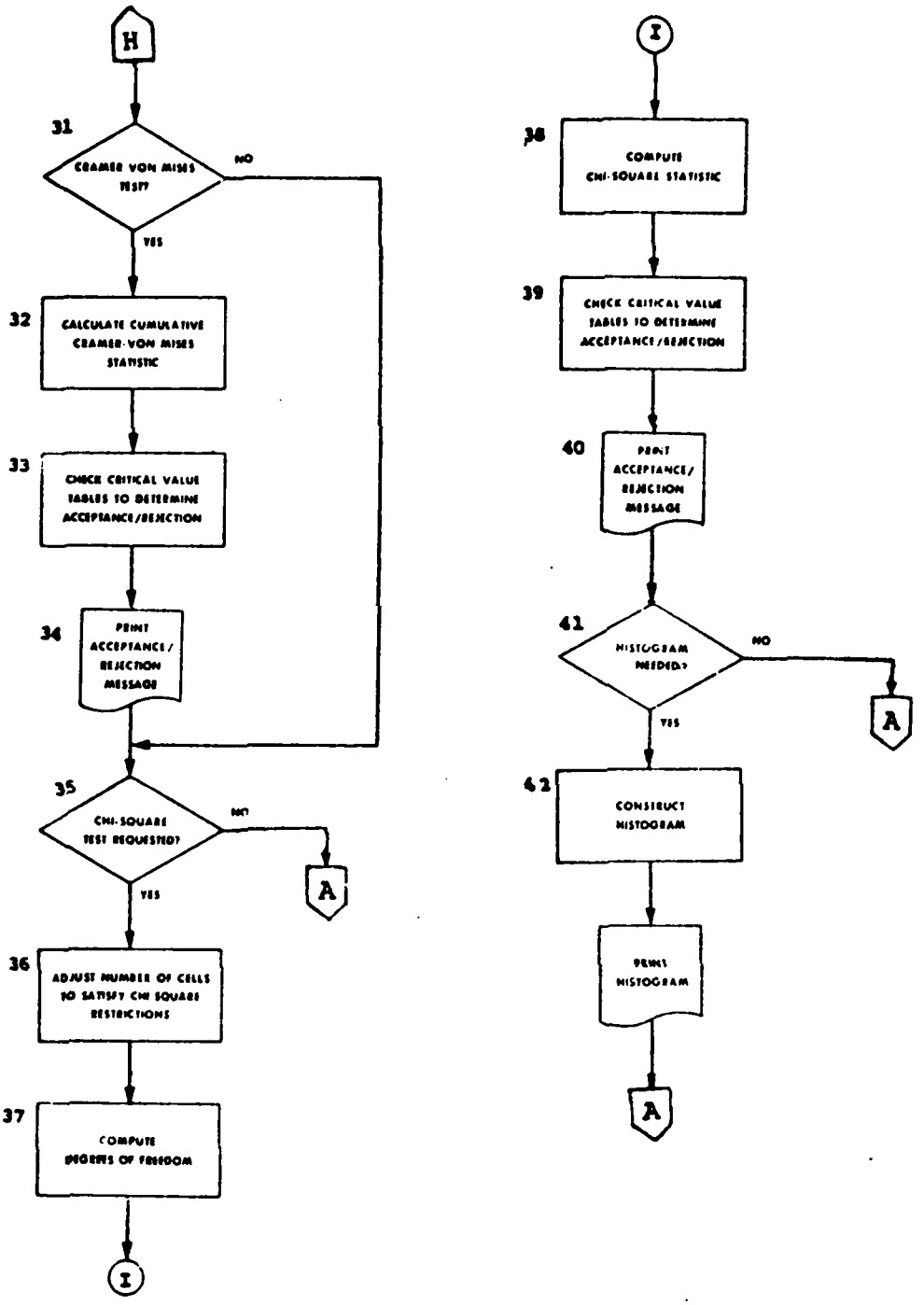


Figure 26 (continued)

Block 1. The GOF language translator is ready to accept the user's input commands immediately following the printing of the GOF program's title line. The system prompt character notifies the user to begin inputting his testing and operational statements. These statements are explained in the next part of this section.

Block 2. At any time the user may input a STOP statement and terminate execution of the GOF program.

Block 3. Often a researcher constructs a histogram of the sample observations from his experiment. From the shape of the histogram, he may obtain a clue as to a possible distribution against which to test his data. Therefore, the GOF program permits the user to see a histogram before he must select the hypothesized distribution. At this point in the program, if the user has seen the histogram and has specified the distribution to test, the GOF program jumps to Block 23 and begins executing the selected goodness of fit tests.

Block 4. The sample observations are read under one of three selectable formats: (1) free-field format, (2) user supplied format, or (3) program default format. These three options are described in detail in the next part of this section. After the values are read, they are echo-printed for the user's verification.

Block 5. At this point, the GOF program handles grouped and ungrouped data differently. The user has the option of inputting either form of data to the GOF program.

Block 6. If the input data is ungrouped, the values are arranged in numerically ascending order using an efficient tree sort algorithm.

Block 7. The mean, adjusted variance, and standard deviation are calculated from the sample observations. This step is performed automatically to provide the user with several descriptive sample statistics.

Block 8. The program also calculates the coefficient of variation for the user. This sample statistic is sometimes helpful in the selection of a distribution to test. Section III of this report contains the formulas (or constant values, in some cases) for the theoretical quantities being estimated for most of the ten distributions included in the GOF program.

Block 9. Either biased or unbiased coefficients of skewness and kurtosis may be calculated. The program defaults to the unbiased calculations, but the user may override that default by the BIASED command. (This command is explained in the next part of this section.) Section II of this report contains the equations used for both methods of computing these two descriptive sample statistics. For ungrouped data, the GOF program transfers to Block 12.

Block 10. For grouped data, the GOF program calculates the mean, adjusted variance, and standard deviation by using each class mark (midpoint) times its absolute class frequency in the equations for these three statistics.

Block 11. Once the mean and standard deviation for the grouped data are calculated, the coefficient of variation is computed. The coefficients of skewness and kurtosis are not computed for grouped data.

Block 12. The descriptive sample statistics are printed. Examples of these printouts are included in Appendix A of this report.

Block 13. The GOF program checks to see if the user requested that a moments test for normality be run. If the user does not want the results of a moments test, the GOF program transfers control to Block 18.

Block 14. The critical value tables for the moments test, which are stored internally by the GOF program, contain entries for values of N up through 125. If N is less than or equal to 125, control goes to Block 16.

Block 15. For values of N greater than 125, the user should refer to a set of tables for the normal distribution to determine the appropriate critical values against which to check the test statistic. This procedure is explained in Section II under the discussion of the moments test. For values of N greater than 125, the GOF program prints a message referring the user to tables for the normal distribution. Control then goes to Block 18.

Block 16. For values of N less than or equal to 125, the GOF program does a table lookup to the nearest value of N (greater than or equal to N) for the 95% level of significance and extracts the test critical values. These values are compared with the calculated values of the skewness and kurtosis coefficients and the results are printed for the user. The tables contain entries only for the 0.05 level of significance.

Block 17. The GOF program prints the results of the moments test. Examples of this type of output are included in Appendix A.

Block 18. If the user inputs ungrouped data, the GOF program must arrange it into cells (or classes) at this point. The GOF program checks to see whether the user specified the number of cells into which he wants his data grouped. If he did, the program transfers to Block 20.

Block 19. If the user does not specify the number of cells for grouping the sample observations, the GOF program, in most cases, defaults to 15. The Poisson distribution is handled differently from the other distributions by the GOF program.

> ...the minimum and maximum values in the data set are determined. If the minimum value is not equal to zero, a change of variables is made to translate all observed data points to a $(0-\infty)$ range. This will not effect the goodness of fit test, but will reduce the possibilities of an overflow (underflow) in the calculations of $e^{-\lambda}$. Once the data have been scaled to zero, (if appropriate) a cell is established for each possible value of the random value X on the range of $x=0,1,2,\ldots(xmax)$. The logic of this approach is that by creating a cell for each possible (discrete) value, the maximum degrees of freedom can be obtained for the Chi-square test (Phillips, 16:17).

The Cramer-Von Mises test uses each observation individually. The Kolmogorov-Smirnov test runs with the number of cells less than or equal to the number of data points. The chi-square test runs with grouped values and requires that each cell contain at least five expected observations. Therefore, the chi-square test could be run with less cells than the user selects (or the program calculates). At this point in the GOF program, the cell values are only calculated. Prior to performing the chi-square test the cell values are checked and may be adjusted to meet the five-observation criterion. The user can find a description of each goodness of fit test and its relationship to grouped data in Section II of this report.

Block 20. The sample observation distribution is calculated for ungrouped input data immediately after the cell boundaries are constructed. If the user is running the batch version of the GOF program, he receives a printout of the cell boundaries and the absolute cell frequencies at this point. The on-line user does not receive this output, but sees it graphically represented in the histogram.

Block 21. The GOF program now checks to see whether the user has specified a hypothesized distribution. If the distribution has been selected, the GOF program continues at Block 23.

Block 22. Because the user has not selected a hypothesized distribution, the GOF program can only print a histogram of the sample observations at this point. After the histogram is printed, the user receives a prompting message asking him to now select a hypothesized distribution. The GOF program transfers control to Block 1 to accept the user's command for the selection of a distribution.

Block 23. The values for the theoretical and cumulative probability distributions are calculated for the hypothesized distribution. If the user inputs the theoretical parameters for the hypothesized distribution, they are used. If the user does not specify the theoretical parameters, sample estimates of these parameters are calculated. Table 3 gives the equations used for each distribution to calculate its parameters from sample estimates. The sample mean is denoted by $\bar{x}$; the sample variance by s^2; and the sample standard deviation by s.

The GOF program calculates the values of the theoretical and cumulative probability distributions for the chi-square and Erlang-k cases by using the same equations it uses for the gamma distribution values. This method is used because the chi-square and Erlang-k distributions are special cases of the gamma distribution. Therefore, if the user inputs the theoretical number of degrees of freedom for the chi-square distribution, that value is divided by two to get the gamma distribution's shape parameter. The gamma scale parameter is set equal to 2 for the calculation of chi-square values. In the Erlang-k case, the estimated shape parameter is rounded to the nearest integer before the gamma distribution equations are used. If the user inputs the theoretical shape parameter for the Erlang-k distribution, he should input an integer value. Section III of this report contains a description of each distribution and its parameters.

Table 3

Sample Estimators of Distribution Parameters

Distribution	Parameter	Estimator
1. Poisson	λ	$\bar{x}$
2. Exponential	β	$\bar{x}$
3. Normal	μ, σ^2	$\bar{x}, s^2$ (unbiased)
4. Log-normal	μ, σ^2	$\bar{x}, s^2$ (of transformed data)
5. Gamma	β (scale) α (shape)	$s^2/\bar{x}$ $(\bar{x}/s)^2$
6. Erlang-k	β (scale) α (shape)	$s^2/\bar{x}$ $(\bar{x}/s)^2$ (rounded to nearest integer)
7. Chi-square	ν (degrees of freedom)	$[(\bar{x}/s)^2]/2$
8. Triangular	X_{min} X_{most} X_{max}	Minimum sample value (3**x) - X_{max} - X_{min} Maximum sample value
9. Uniform	A (lower limit) B (upper limit)	$\bar{x} - s\sqrt{3}$ $\bar{x} + s\sqrt{3}$
10. Weibull	α (scale) β (shape)	See reference 27, pages 465-469 for explanation of the techniques used to estimate Weibull parameters.

Block 24. If the user has requested that no histogram of his empirical data be printed, the GOF program transfers to Block 26. If the user is running under the delayed distribution selection option, the GOF program transfers to Block 26 because the user has already seen a histogram of his empirical data (Block 22).

Block 25. The GOF program constructs and prints a histogram of the empirical data. The number of cells in the histogram is either:

1. user selected,
2. a program default value of 15, or
3. the number of possible integer values in the random variable range (for Poisson tests only).

Block 26. At this point, the GOF program is ready to perform the required goodness of fit tests. If the Kolmogorov-Smirnov test is not required, control is transferred to Block 31.

Block 27. The Kolmogorov-Smirnov test compares the cumulative distribution function of the sample distribution with the cumulative distribution function of the theoretical distribution. Therefore, at this point, the cumulative distribution values for the sample distribution are calculated.

Block 28. The Kolmogorov-Smirnov test statistic is calculated. Section II of this report contains an explanation of how the Kolmogorov-Smirnov test works.

Block 29. The Kolmogorov-Smirnov calculated test statistic is compared with the appropriate critical values stored internally in the GOF program. These critical values are available for values of ALPHA of 0.01 and 0.05. (ALPHA is the probability of rejecting the null hypothesis when it is true. Its use is explained in the next part of this section.) The internally-stored tables contain entries for values of N from 1 through 20 and for the values 25, 30, and 35. For values of N between 21 and 34 which are not in the tables, the next higher multiple of 5 is used. For example, if N=23, the critical value for N=25 is used. For values of N greater than 35, the critical values are calculated by the following equations:

$$1.36/\sqrt{N} \text{ for ALPHA=0.05, and}$$
$$1.63/\sqrt{N} \text{ for ALPHA=0.01.}$$

Block 30. The GOF program prints the results of the Kolmogorov-Smirnov test. An example of this output is contained in Appendix A of this report.

Block 31. The GOF program, at this point, checks to see if the Cramer-Von Mises test is desired. If it is not, the program continues at Block 35.

Block 32. The cumulative Cramer-Von Mises test statistic is computed.

Block 33. The calculated Cramer-Von Mises test statistic is compared with the appropriate critical value by the GOF program. For a 0.01 level of significance, the Cramer-Von Mises critical value is 0.743. For the 0.05 level of significance, it is 0.461.

Block 34. The results of the Cramer-Von Mises test are printed. An example of this output is contained in Appendix A of this report.

Block 35. The GOF program checks to see if the user wishes to have the chi-square goodness of fit test run. If not, the program goes to Block 1 for the next run or termination of the GOF program.

Block 36. One of the requirements of the chi-square test is that each cell of the theoretical distribution contain at least five observations. At this point, the GOF program checks to be sure this requirement is satisfied. If it is not, adjacent cells are merged until each cell contains at least five entries.

Block 37. The number of degrees of freedom is computed. An explanation of how the number of degrees of freedom is determined for the chi-square test is included in Section II of this report.

Block 38. The chi-square test statistic is computed. Section II contains the equations used for the calculation.

Block 39. The GOF program enters the internally stored chi-square critical value tables with the computed number of degrees of freedom and the user selected ALPHA value. The appropriate critical value is then compared with the computed chi-square test statistic. Internally, the GOF program contains values for the following numbers of degrees of freedom: 4-29, 30, 40, 50, 60, 70, 80, 90, and 100. Critical values are available internally only at the 0.01 and 0.05 levels of significance. For values between 31 and 99 not contained in the tables, a linear interpolation is performed to calculate the required critical value. For values over 100, a message is printed referring the user to the normal distribution tables.

Block 40. The user receives a printout of the results of the chi-square test. An example of this output is included in Appendix A of this report.

Block 41. If the number of classes describing data is changed to meet the requirements of the chi-square goodness of fit test, a second histogram is printed. This second histogram provides the user insight into the arrangement of his data for the chi-square test only. If the user is running with the NO HISTOGRAM option, the GOF program transfers to Block 1. If the number of classes is not changed for the chi-square test, the GOF program transfers to Block 1.

Block 42. The components of the histogram are constructed and the histogram of empirical data as it is grouped for the chi-square test is printed. The GOF program returns to Block 1 for the next run or for notification of termination.

E. Instructions for Use

This section is devoted mainly to a description of the use of the on-line version of the GOF program. The batch version's operation is almost identical and the GOF command language is the same in either mode. Rather than typing in GOF language statements through a terminal, the batch GOF user punches the identical commands on cards. Any differences between the two modes are explained as they occur.

Upon execution, the GOF program identifies itself to the on-line user by printing:

***** GOODNESS OF FIT PROGRAM *****.

Following its initial identifying message, the system prompt character appears on the terminal and the GOF language translator is waiting to accept GOF language commands. All statements have the general format shown below.

<INSTRUCTION>::=<VAR>:<VALUE>;|<VAR>:<VALUE><VAR>;|
<VAR>;|<VAR>[:<VALUE>];

The entire instruction name and key words are acceptable or a three-character shorthand is sufficient. The user, for example, can enter TEST:UNIFORM DISTRIBUTION; or he may type only TES:UNI; and either instruction is satisfactory to tell the GOF program to test for a uniform distribution. Several instructions may be entered on a single line or one instruction may be input across more than one line. Blanks within an instruction or between instructions are ignored. All instructions must end with a semicolon. Each instruction is read and its syntax checked by the GOF language translator. However, instructions are not executed until the START command is read. The START command is the signal to the GOF program to begin execution of its calculation section.

The GOF instructions can be divided into two types: testing instructions and operational instructions. The testing instructions describe which calculations the GOF program is to make while the operational instructions govern the non-mathematical flow of the program. Each instruction is described, its full name is given, its three-character abbreviation is also given in parentheses, and examples of its use are included in this section.

F. Testing Instructions

$... ANY COMMENT STRING ...;

This statement is a comment statement and may be included at anytime in the command session. It is used to identify the particular job being run. It may be of any length, but must begin with a "$" and terminate with a ";". An example is:

$TESTING DISTRIBUTION OF WEIGHTS OF GRAIN-FED CALVES;

DATA POINTS:<VALUE>; (DAT:<VALUE>;)

The VALUE in this command represents the number of data points which are read during the execution of the GOF program. For ungrouped data, VALUE represents the actual number of sample observations. For grouped data, VALUE is the number of input data items which are to be read. For grouped data, the lower class boundary, the upper class boundary, and the absolute class frequency must be input in that respective order for each cell. Therefore, the number of individual data items to be read as input is actually three times the number of cells into which the data is grouped. Example 3 of the sample problems (in Appendix A) illustrates the input of grouped data. The number of data observations which constitute the original sample prior to grouping is calculated later by the GOF program. It simply adds the absolute frequency counts of all cells.

The DATA POINTS instruction is optional. If it is not used, the GOF program must be told to count the number of data values being input. In the case of grouped data, counting must produce a value which is a multiple of three or an error message results and program execution terminates.

Examples of the DATA POINTS statement include:

```
DATA POINTS:100;  (100 data points are to be read.)
DAT:63;           (63 data points are to be read.)
D A T A :  4 6 1 ; (Blanks have not effect.)
```

TEST[:<VALUE>]; (TES[:<VALUE>];)

This statement specifies which theoretical probability distribution is to be tested. There may be one to four entries for the VALUE portion of this statement. The first and only mandatory entry must be the name of the theoretical probability distribution to be tested. The valid distributions are:

1. Poisson,
2. exponential,
3. normal,
4. log-normal*,
5. gamma,
6. Erlang-k,
7. chi-square,
8. triangular,
9. uniform, and
10. Weibull.

---

*The GOF program transforms all hypothesized log-normal data to normal random variables by $y=\ln(x)$. After the transformation, it applies the goodness of fit tests against a hypothesized normal distribution. Because y is not defined at x=0, the GOF program sets y=-10.0 when it encounters an input value of zero. Hypothesized log-normal data should not contain zero values, because the log-normal distribution is only defined for values greater than zero. The program executes with y=-10.0, but the results are biased by this substitution (16:26).

The acceptable three-character shorthand is the first three letters of each distribution name.

If the parameters of the distribution are known, they may be input as part of this instruction or they may be input separately by using SCALE, SHAPE, DEGREES OF FREEDOM, LOWER, UPPER, MEAN, VARIANCE, MAXIMUM, MOST, and MINIMUM instructions. If the user chooses to input the distribution parameters as part of the TEST statement, Table 4 shows the order which must be used for distributions having more than one parameter.

Table 4
Input Order of Theoretical Distribution Parameters

Distribution	Parameter Input Order
1. Normal, log-normal	1. Mean, variance
2. Erlang-k, gamma, Weibull	2. Scale, shape
3. Uniform	3. Lower limit, upper limit
4. Triangular	4. Minimum, most, and maximum

If one parameter of a multi-parameter distribution is input, all other parameters for that distribution must also be input. A mixture of theoretical and estimated parameters is not acceptable to the GOF program. If the user inputs the known values of the parameters for his hypothesized distribution, the GOF program counts the number of parameters being input. If an incorrect number of parameters is input for any hypothesized distribution, the user receives an error message and the GOF program terminates. Appendix C of this report contains all error messages which the user might receive from the GOF program.

If the user does not want to specify his distribution parameters with the TEST statement, he may input these values at another time using the statements which identify the parameters by name. The individual parameter statements are explained at a later point in this section.

If the user chooses to delay the selection of a hypothesized distribution until after the printing of the histogram of empirical data, the program returns a prompting message after the histogram which says:

** SELECTION OF DISTRIBUTION FOR NULL HYPOTHESIS MUST BE MADE NOW **.

After printing the message, control is returned to the GOF language translator which waits for the user to input a TEST statement. The TEST statement is the only statement which can be delayed until after the

histogram is printed. A delayed TEST statement should be followed by a START statement to continue execution.

Examples of the TEST statement follow.

```
TEST:POISSON;
TES:POI;
TES:POI:1.0;  (Test for the Poisson distribution with a
               mean of 1.0.)
TEST:NORMAL:5.0:1.5;  (Test for the normal distribution
                       with a mean of 5.0 and a variance
                       of 1.5.)
TES: W E I;
```

RUN[:<VALUE>]TEST; (RUN[:<VALUE>];)

The VALUE portion of the RUN statement may contain any of five acceptable character strings:

1. CHI-SQUARE or CHI,
2. KAS for the Kolmogorov-Smirnov test,
3. CVM for the Cramer-Von Mises test,
4. MOMENTS or MOM, or
5. ALL if all four goodness of fit tests are desired.

The user may select from one to four tests to run during a single execution of the GOF program. The coefficients of skewness and kurtosis are automatically calculated for all ungrouped sample data. However, these coefficients are not compared with the critical values for the moments test unless the moments test is requested via the RUN statement.

Examples for this statement take the following forms.

```
RUN:CHI:KAS;
RUN:ALL TEST;
RUN:MOM TEST;
```

MEAN OF POPULATION:<VALUE>; (MEA:<VALUE>;)

If the user decides to input the theoretical mean of his distribution and does not wish to do it through the TEST statement, he may use the MEAN command. This command is valid only for the Poisson, exponential, log-normal, and normal distributions because only those four GOF-implemented distributions have means as a characteristic parameter.

Examples for this statement include the following commands.

```
MEAN OF POPULATION:12.5;
MEA:0.305;
```

VARIANCE OF POPULATION:<VALUE>; (VAR:<VALUE>;)

The theoretical variance of a distribution may be input to the GOF program via the VARIANCE statement. Only the log-normal and normal distributions might require the use of the VARIANCE statement. The theoretical variance can also be input as the second parameter value in the TEST statement if desired. Examples can be illustrated by the following statements.

```
VARIANCE:  4;
VAR:  16;
```

```
MINIMUM VALUE:<VALUE>;          (MIN:<VALUE>;)
MOST PROBABLE VALUE:<VALUE>;    (MOS:<VALUE>;)
MAXIMUM VALUE:<VALUE>;          (MAX:<VALUE>;)
```

The MIN, MOS, and MAX statements are used only if the user is testing for the triangular distribution and wishes to specify the theoretical parameters of this distribution. The MIN value represents the smallest value the distribution can attain. The MOS value is the most probable value of the distribution. The MAX value is the largest value possible for the distribution. For certain types of triangular distributions, the MIN value equals the MOS value or the MAX and MOS values are the same. For those cases, all three parameters must be individually supplied even though two may be equal. Section III of this report describes the triangular distribution and its parameters. Theoretical parameter values for the triangular distribution may also be input by the TEST statement. If the theoretical parameters of the triangular distribution are not known, these three commands are omitted and the sample estimates of these three parameters are calculated. Several examples are listed below.

```
MIN:  4;
MAXIMUM VALUE:  165.23;
MOST:  37;
```

```
SCALE PARAMETER:<VALUE>;        (SCA:<VALUE>;)
SHAPE PARAMETER:<VALUE>;        (SHA:<VALUE>;)
```

The SCALE and SHAPE PARAMETER statements supply a method for designating the theoretical scale and shape parameters for the Erlang-k, gamma, and Weibull distributions. The user can refer to Section III of this report or to the Hastings and Peacock book (9) for a description of the relationship of the scale and shape parameters to the mean and variance of these three distributions. The user is cautioned not to input a mean or a variance as a scale or shape parameter. Both of these statements must be used if either is used because the GOF program does not accept a mixture of theoretical and estimated distribution parameters. The SCALE and SHAPE commands can be used as illustrated in these two examples.

```
SCA:  1.3;
SHAPE PARAMETER:  2.3;
```

DEGREES OF FREEDOM:<VALUE>; (DEG:<VALUE>;)

The DEGREES OF FREEDOM statement is uniquely applicable to the chi-square distribution for the GOF program. It is used to input the one theoretical parameter of that distribution, if it is known. The degrees of freedom parameter is actually the shape parameter of the chi-square distribution, but is more commonly known by its degrees of freedom name. It is always an integer value. Its use is illustrated by two examples.

```
DEGREES OF FREEDOM:15;
DEG:26;
```

LOWER LIMIT:<VALUE>; (LOW:<VALUE>;)
UPPER LIMIT:<VALUE>; (UPP:<VALUE>;)

The two parameters of the uniform (or rectangular) probability distribution are usually referred to as the lower and upper limits of the distribution. If known, they may be input to the GOF program through the LOWER and UPPER statements or as part of the TEST statement. Occasionally, an alternative parameter to the upper limit is used to describe the uniform distribution. This alternative parameter may be called the range or scale parameter. The range parameter is defined to be the upper limit minus the lower limit. Hastings and Peacock develop the rectangular distribution using the range parameter (9:116). The user is cautioned not to input a range parameter in lieu of the expected upper limit. The lower limit must be less than the upper limit. These two statements might take the following forms.

```
LOWER LIMIT:  10;
UPP:  25;
```

CELLS:<VALUE>; (CEL:<VALUE>;)

The CELLS statement is optional for ungrouped input data and useless for grouped input data. If one inputs ungrouped data, he may specify the number of cells (or classes) to be used for grouping the data. The integer supplied for VALUE may be less than or equal to the number of data values to be input. If the user does not use the CELLS statement for ungrouped data, the GOF program defaults to 15 cells. If the chi-square goodness of fit test is selected to be run, the number of cells, whether user selected or program default, may have to be adjusted to meet the requirements of that test. (Section II contains a discussion of the chi-square test.)

For grouped data, the number of cells is implicitly provided through the DATA POINTS statement. After the class boundaries and absolute frequencies are read, the GOF program, for grouped data, calculates the number of cells by dividing the number of input items by three. For grouped data, three input values are required to define each cell: (1) the lower class boundary, (2) the upper class boundary, and (3) the absolute class frequency. Therefore, should the user inadvertently use the CELLS instruction for grouped data, the value he inputs is destroyed by the value calculated by the GOF program. Example 3 of the sample runs which are described in Appendix A of this report illustrates the use of grouped input data.

The CELLS statement can be used as the following examples demonstrate.

```
CELLS:  20;
C E L : 1 5 ;
```

```
BIASED ESTIMATION;          (BIA;)
UNBIASED ESTIMATION;        (UNB;)
```

One of the options of the GOF program is the selection of biased or unbiased calculations for the coefficients of skewness and kurtosis. These calculations are described in Section II of this report. The coefficients are calculated automatically from the ungrouped sample observations and the results printed as part of the descriptive sample statistics. If the user does not specify them, the program defaults to calculating the unbiased estimates. These two coefficients are computed regardless of whether the moments test for normality is run.

For grouped input data, neither the coefficient of skewness nor kurtosis is computed. Therefore, these two instructions are meaningless if the user has only grouped data to input.

Examples of these instructions are simple.

```
BIA;
UNBIASED;
```

```
ALPHA:<VALUE>;      (ALP:<VALUE>;)
```

The ALPHA command selects the level of significance at which each goodness of fit test is evaluated. ALPHA is the probability of a Type I error--the error of rejecting the null hypothesis when the null hypothesis is actually true. For the GOF program, the VALUE for ALPHA may be either 0.01 or 0.05. This does not mean that a test cannot be run for another level of significance. It does mean that, for values other than 0.01 and 0.05, the user is responsible for checking critical value tables for the goodness of fit test being run to evaluate the results of the test. The GOF program always prints the value of the test statistic and the number of degrees of freedom or the sample size for each test it runs. To evaluate the results of a test for an ALPHA value other than 0.01 or 0.05, the user compares the printed test statistic value to the critical value for the level of significance he desires. The number of degrees of freedom or the sample size is necessary to locate the appropriate critical value for every test except the Cramer-Von Mises test. If the test statistic value exceeds the critical value, the user has insufficient evidence to accept the null hypothesis. If the test statistic is less than or equal to the critical value, the user has insufficient evidence to reject the null hypothesis. If the user does not specify the value of ALPHA, the GOF program defaults to 0.05. The moments test can only be evaluated at the 0.05 level of significance by the GOF program.

The ALPHA statement is acceptable in the following ways.

```
ALPHA:  0.01;
ALP:  0.05;
```

G. Operational Instructions

COUNT DATA POINTS [:<VALUE>]; (COU[:<VALUE>];)

If the exact number of sample observations is not known, the COUNT statement initiates counting of the data items upon input. The VALUE portion of this command should contain the numeric value of the trailer number. If a trailer number is not input as part of the COUNT statement, the GOF program defaults to 999. In addition to the trailer number, a system end-of-file must be included immediately following the last data record. The trailer value is not considered part of the data set. The COUNT statement works for either grouped or ungrouped data. If the number of sample observations is known, the user should supply that information to the GOF program through the DATA POINTS:<VALUE>; statement. The COUNT command can be used as the following samples illustrate.

```
COUNT:  7777 ;  (The terminal value is 7777.)
COU;            (The terminal value is 999, by default.)
```

FREE FIELD READ; (FRE;)

The FREE FIELD READ statement allows the user to input sample observations using a free field format. This command invokes a subroutine internal to the GOF program and does not rely on system-dependent free field read capabilities. Data values in integer, floating, or scientific notation (E-format) are acceptable. The following requirements must be satisfied for FREE FIELD input.

1. Data items must be separated by commas except for the last data item in each record--it is not followed by a comma.
2. A data item may not be started on one record and continued to the next.
3. Zeroes may be input by multiple concatenated commas.
4. Blanks are allowable in any location, but cannot be used as numeric placeholders (substitutes for zeroes).
5. Data values may be input under a single notation, or a mixture of integer, floating, and scientific notation. The FREE FIELD READ statement examples are followed by examples of data sets which are acceptable to the GOF program during a free field read. The numbers in parentheses represent the converted values which are used by the GOF program.

```
FREE FIELD READ;
FRE;

-0.125,.23,2E-2,-1.0E-1,1,+1.23E-01,1.9678
(-0.125,.23,.02,-.10,1,.123,1.9678)
1.05,-2.61754,3,-3,,,+1.24E+1,-.15,0,100E-2
(1.05,-2.61754,3,-3,0,0,12.4,-.15,0,1.00)
```

START RUN; (STA;)

After the testing parameters are defined, the START RUN statement is entered to begin execution of the defined model. The START RUN command should be the last instruction issued for any particular testing model.

If the selection of a distribution to be tested is delayed until after the histogram has been printed, the GOF program prints the following message:

** SELECTION OF DISTRIBUTION FOR NULL HYPOTHESIS MUST BE MADE NOW **.

After printing that promt message, the GOF language translator waits for a TEST:<VALUE>; statement. The TEST command must be followed by a START RUN; command. In this special mode of operation which delays specification of a hypothesized distribution, two START statements are required--the first triggers the construction and printing of the histogram and the second causes the hypothesized model to be tested.

The START RUN; statement is acceptable in any of the following forms.

```
START RUN;
STA;
```

FORMAT READ:<VALUE>; (FOR:<VALUE>;)

This command allows the user to input his own FORTRAN FORMAT to be used in reading the input data. The content of VALUE is any valid FORTRAN FORMAT statement and must include the opening and closing parentheses. The GOF program defaults to an (8F10.5) format without the use of this statement or the FREE FIELD READ command. Integer (I) and alpha (A) formats are not acceptable. Integer values may be input under an F format. For example, a four digit integer like 7654 may be input using an F4.0 format. Examples of this statement follow.

```
FORMAT READ:(10F6.3);
FOR:  (3F11.4);
```

NEXT PROBLEM; (NEX;)

This statement causes the GOF program to clear all relevant storage locations in preparation for accepting parameters and data for the next problem. It may be used as the examples illustrate.

```
NEXT PROBLEM;
NEX;
```

STOP; (STO;)

The STOP statement causes the GOF program to terminate execution. Examples of this instruction are trivial.

```
STOP;
STO;
```

ALTERNATE DATA DEVICE:<VALUE>; (ALT:<VALUE>;)

This instruction provides the user with the ability to input data to the GOF program from a device other than the one he is using to supply testing and operational commands. The VALUE of this command must contain the integer number which corresponds to the appropriate system linkage to the alternate input device. The alternate device may physically be a tape file, disk file, or drum file. Appendix B contains sample Univac 1108 run streams which illustrate the use of an alternate input device. Two examples of this command follow.

```
ALTERNATE DATA DEVICE:  10;
ALT:  30;
```

INTERACTIVE MODE; (INT;)

The INTERACTIVE MODE command notifies the GOF program that the user is running in an on-line mode. The primary effect of this statement is to limit the intermediate printed output and the width of the print field. It can be used as the following illustrations demonstrate.

```
INTERACTIVE;
INT;
```

NEW DATA; (NEW;)

The NEW DATA instruction allows the user to use the same testing model defined for the previous run, but read new data. All parameters which are associated with data input must be redefined. The following commands are concerned with data input specifications:

1. DATA POINTS:<VALUE>;
2. COUNT DATA POINTS:<VALUE>;
3. FREE FIELD READ;
4. FORMAT READ:<VALUE>;
5. ALTERNATE DATA DEVICE:<VALUE>;, and
6. GROUPED DATA;.

Examples of the NEW DATA statement follow.

```
NEW DATA;
N E W;
```

DUMP INPUT MODEL; (DUM;)

The DUMP statement provides the on-line user with a tabular printout of the input parameters of the model currently being run. The batch user gets this dump automatically. Therefore, the DUMP INPUT MODEL statement is redundant in the batch mode. Figure 27 illustrates the output provided by the DUMP command.

```
*****************************************************************
*      G O O D N E S S   O F   F I T   P R O G R A M            *
*                                                               *
*  REWRITTEN AND EXPANDED BY:  SUE D. GUTHRIE                   *
*                              UNIV SOUTHERN MISS (1979)        *
*  PREVIOUS ENHANCEMENTS BY:   RALPH B. BISLAND,JR.             *
*                              UNIV SOUTHERN MISSISSIPPI        *
*                              LARRY SCHEUERMANN                *
*                              NICHOLS STATE UNIV               *
*  ORIGINALLY WRITTEN BY:      DON T. PHILLIPS                  *
*                              TEXAS A&M UNIV                   *
*---------------------------------------------------------------*
*                                                               *
*      S U M M A R Y   O F   I N P U T   P A R M S              *
*                                                               *
*      DISTRIBUTION TO BE TESTED - - - - - - - NORMAL           *
*      TYPE OF TEST(S) - - - - - - - - - - - - 1 CHI-SQRE       *
*                                              2 K AND S        *
*      TYPE OF ESTIMATION  - - - - - - - - -    UNBIASED        *
*      CLASSIFICATION OF INPUT DATA  - - - -  UNGROUPED         *
*      NUMBER OF CELLS - - - - - - - - - - - UNSPECIFIED        *
*      NUMBER OF DATA POINTS - - - - - - - -         40         *
*      MEAN OF POPULATION  - - - - - - - - -     10.000         *
*      VARIANCE OF POPULATION  - - - - - - -     10.240         *
*      ALTERNATE DEVICE NUMBER - - - - - - -         10         *
*      MODE OF OPERATION - - - - - - - - - -      BATCH         *
*      INPUT FORMAT FOR DATA - - - - - - - -     8F10.5         *
*      HISTOGRAM REQUESTED - - - - - - - - -        YES         *
*      LEVEL OF SIGNIFICANCE - - - - - - - -        .01         *
*                                                               *
*****************************************************************
```

Figure 27
Output Generated by DUMP INPUT MODEL;

Examples of the DUMP command are obvious.

```
DUMP INPUT MODEL;
DUM;
```

NO HISTOGRAM; (NOH;)

The GOF program omits printing the histogram of the sample observations if the user inputs the NO HISTOGRAM; command. When this command is used, the selection of the hypothesized distribution cannot be delayed. Such a run is meaningless and the GOF program terminates with an error message. The examples for this statement follow.

```
NO HISTOGRAM;
NOH;
```

GROUPED DATA;

The GOF program assumes that all input data represents individual sample observations. If the user wishes to input grouped data, he must notify the GOF program by using this GROUPED DATA; command. Grouped data requires three data values be input for each class: (1) lower class boundary, (2) upper class boundary, and (3) absolute class frequency. Grouped data is only acceptable for equal-interval classes. The first and last classes should not be open classes. The calculation of the sample descriptive statistics is incorrect if open-ended classes are input. This statement can take the following forms.

```
GROUPED;
GRO;
```

There are no more commands which are acceptable to the GOF language translator. Table 5 provides a three-character synopsis of all the testing and operational instructions which are described in this section.

Table 5
Synopsis of GOF Language Statements

Testing Statements		Operational Statements
$COMMENT	LOW:<VALUE>;	COU[:<VALUE>];
DAT:<VALUE>;	UPP:<VALUE>;	FRE;
TES[:<VALUE>];	CEL:<VALUE>;	STA;
RUN:<VALUE>;	BIA;	FOR:<VALUE>;
MEA:<VALUE>;	UNB;	NEX;
VAR:<VALUE>;	ALP:<VALUE>;	STO;
MIN:<VALUE>;		ALT:<VALUE>;
MOS:<VALUE>;		INT;
MAX:<VALUE>;		NEW;
SCA:<VALUE>;		DUM;
SHA:<VALUE>;		NOH;
DEG:<VALUE>;		GRO;

There are no additional operating instructions for the GOF program. The GOF language translator monitors input commands and notifies the user of errors or default conditions. Appendix C of this report contains a list of all possible error messages the user might receive from the GOF program. Additional information and suggestions for corrective actions are included for many of the error messages.

The GOF program does not require the use of any external storage mediums (tapes, disks, drums) unless their use is imposed on the program through the ALTERNATE DATA DEVICE command. Internal storage for sample observations is set at 500. This value can easily be changed by altering a DIMENSION statement and the value of the variable MAXDAT in the main GOF program and recompiling the program. No other changes need to be made to any GOF program component to allow for a change in the maximum number of data observations which the program accepts.

The GOF program is not coded to run in an overlay structure. However, Appendix D of this report describes each of the GOF program components and their interrelationships for the user who wants to construct an overlay version of the GOF program.

V. SUMMARY

A computer program to perform four statistical goodness of fit tests is described in this work. Statistical goodness of fit testing is a mathematical procedure for evaluating how closely a set of observed sample values fits a hypothesized theoretical probability distribution. Ten theoretical probability distributions are available in the computer program documented by this report.

The scope of the program includes calculations for four goodness of fit methods:

1. chi-square,
2. Kolmogorov-Smirnov,
3. Cramer-Von Mises, and
4. moments test for normality.

The ten probability distributions available are frequently found in engineering environments and the natural and physical sciences. These distributions are: (1) Poisson, (2) exponential, (3) normal, (4) log-normal, (5) gamma, (6) Erlang-k, (7) chi-square, (8) triangular, (9) uniform, and (10) Weibull. The methodology of the program's operation follows closely the steps employed in testing any statistical hypothesis. One of the major goals in developing the GOF software is to make it truly self-contained. Therefore, the program:

1. applies the goodness of fit tests,
2. constructs and prints a histogram,
3. calculates descriptive sample statistics, and
4. automatically evaluates the results of the tests at either the 0.05 or the 0.01 level of significance.

VI. CONCLUSIONS

The goodness of fit computer program in combination with this report provides the researcher with a valuable tool to use in the analysis or summary of empirical observations. The following four paragraphs outline the primary benefits of the program.

Time-Saving. Section I of this report outlines the steps necessary for performing a goodness of fit test against a theoretical probability distribution. Many of the steps are lengthy and require that the user have a considerable amount of familiarity with the assumptions and equations of the method. Incorporating many of these steps into a single, easy-to-use FORTRAN computer program offers a time-saving product to the scientist who needs to analyze random phenomena. The user of the program only needs to manually perform one goodness of fit test on empirical data to a hypothesized distribution to appreciate the services at his disposal by this program.

Completeness. In addition to the time-saving benefit, the GOF program is self-contained. It requires neither special purpose hardware nor software. Inclusion of the critical value tables in the software eliminates the need for additional books or papers to evaluate the goodness of fit test results. The software is designed to assist the user in all three phases of computing--input, processing, and output.

Methodology of Operation. The goodness of fit program operates in a manner similar to that of a reseracher developing and testing a statistical hypothesis. The program could operate in a less-disciplined manner and attempt to fit empirical data to a variety of distributions during a single run. Operating in this manner is not conducive to promoting careful scrutiny of sample observations by the user. Certain distributions are not applicable to empirical data by reason of the environment from which the data comes. Mathematically, data may "fit" a distribution when logically it is impossible. Therefore, the structure of the goodness of fit program helps the user to examine the data and the ten distributions which are part of the program for a logical relationship before attempting to find a mathematical relationship.

Software Capabilities. The goodness of fit program exists in two versions--batch and on-line. Each version is controlled through easy-to-use English language commands. Modularity is built into the software. Therefore, the addition of new distributions can be accomplished through a set of concise well-defined steps. Comprehensive evaluation of input parameters is included to help circumvent invalid executions. Extensive English language diagnostic messages are output to describe error and default conditions or optional operational procedures.

VII. APPENDIX A - SAMPLE ON-LINE OUTPUT

A. Example 1

The first example uses 100 sample observations taken from a process believed to be normally distributed with a mean of 10.0 and a variance of 10.24. The input statements and on-line output from this sample are reproduced in Figure 28.

Following the identification line which the GOF program prints as its first function, the GOF language statements are shown. The Univac 1108 system prompt character (>) precedes every on-line statement. The GOF program echoes each command. The statements for example 1 are explained as they relate to this particular example. A general explanation of all GOF statements is included in Section IV of this report. All GOF statements, except for the START and STOP commands, may be entered in any order.

DAT:100;. The GOF program is notified to expect 100 sample data values by this command.

ALP:0.01;. The value of ALPHA (the probability of rejecting the null hypothesis when it is true) is set to 0.01 by this instruction. The results of all goodness of fit tests are analyzed at the 0.01 level.

TES:NOR:10.0:10.24;. The hypothesized distribution is the normal distribution with a known mean of 10.0 and a known variance of 10.24. The variance, not the standard deviation, is the second expected parameter for this distribution. The mean and the variance, if known, could be omitted from the TEST command and entered using MEAN and VARIANCE statements.

RUN:KAS:CHI;. Two goodness of fit tests are requested: (1) the Kolmogorov-Smirnov test (KAS); and (2) the chi-square test (CHI). The order of input of the tests to be run is not significant. For example, RUN:KAS:CHI; produces the same results as RUN:CHI:KAS;.

ALT:10;. The 100 sample observations are to be input through an alternate input device known as logical unit 10.

INT;. The GOF program is being run in the on-line (interactive) mode. Therefore, the amount of intermediate output is curtailed and the width of the print field is reduced to 64 characters or less.

DUM;. A summary of all input variables and options is desired. This summary is printed automatically in the batch mode, but must be requested by the on-line version of the GOF program.

STA;. All testing and operational commands are entered and the execution of the computational section of the GOF program begins with the START instruction.

The observed data values are read from device 10 and printed. Because the DUM statement requests a synopsis of input parameters, the boxed information following the observed data is given. The following entries in the input summary information are a result of default conditions within the GOF program for this particular example:

1. TYPE OF ESTIMATION,
2. CLASSIFICATION OF INPUT DATA,
3. NUMBER OF CELLS,
4. INPUT FORMAT FOR DATA, AND
5. HISTOGRAM REQUESTED.

The 100 sample observations are sorted into ascending order and the GOF program prints the ordered data. The descriptive sample statistics automatically follow. The coefficients of skewness and kurtosis are printed, but are not compared to critical values because the moments test is not required in this example.

The hypothesis statement follows and reflects the fact that the theoretical mean and variance are known for this example. Example 2 illustrates hypothesis statements with the mean and variance estimated from the sample observations.

The histogram of the empirical data is printed with a default grouping of 15 cells. Because the distribution is hypothesized to be normal, the first and last cells are modified to be open intervals extending over the defined range of the random variable X for the normal distribution. (In the GOF program, ±9999.999 is a substitute for $\pm \infty$.)

The Kolmogorov-Smirnov test may be run on either grouped or ungrouped data. This example illustrates the execution of this test on grouped data. If the number of cells is specified to be equal to the number of data points, the GOF program executes the Kolmogorov-Smirnov test on ungrouped data. The input of a CELL:100; command for the 100 data points in this problem causes execution of the Kolmogorov-Smirnov test on ungrouped data. Running the Kolmogorov-Smirnov test on ungrouped data is recommended by this author. Section II of this report discusses this particular goodness of fit test.

The results of the Kolmogorov-Smirnov and chi-square goodness of fit tests are printed giving the level of significance at which the tests are evaluated. The number of degrees of freedom (or sample size for the Kolmogorov-Smirnov test), the computed test statistic, and the critical value are all printed. Both tests show that there is insufficient evidence to reject the normality hypothesis at the 0.01 level of significance. In this example, the Kolmogorov-Smirnov critical value is located by the value of ALPHA and the number of groups (not individual samples) used in calculating the test statistic.

When the chi-square test is run, the number of cells into which the data is grouped is subject to modification. The reason for this potential adjustment is that the chi-square test requires that each cell contain at least five expected observations. Therefore, the GOF program

checks each cell of expected values, regroups the expected cells if necessary, and also regroups the observed cells to maintain an equal number of expected and observed cells. When this regrouping does occur, the GOF program prints a second histogram of the empirical data which displays the grouping used by the chi-square test. For example 1, the data is regrouped from 15 to 10 cells before the chi-square test is performed. In this example, the true mean and variance of the population are known. Therefore, the chi-square critical value is located for 9 degrees of freedom. If the two population parameters are estimated from the sample data, the critical value (in this case) is located for 7 degrees of freedom. The number of degrees of freedom for the chi-square test is computed by the GOF program as

$$n-p-1,$$

where

n=the number of cells, and
p=the number of parameters estimated.

After completing the histogram which reflects the results of the chi-square regrouping, the GOF program returns to its language translator and waits for another run to be described or the termination of execution to be requested. Figure 28 shows the entry of the STOP statement which causes the final printout by the GOF program and an end to execution.

Example 1 provides a case in which both goodness of fit tests show acceptance of the normality hypothesis and both histograms suggest the shape of the normal curve.

```
     *****  GOODNESS OF FIT PROGRAM  *****

>DAT:100;
 DAT:100;

>ALP:0.01;
 ALP:0.01;

>TES:NOR:10.0:10.24;
 TES:NOR:10.0:10.24;

>RUN:KAS:CHI;
 RUN:KAS:CHI;

>ALT:10;
 ALT:10;

>INT;
 INT;

>DUM;
 DUM;

>STA;
 STA;
 OBSERVED DATA

   7.56042  11.11718   9.13020  12.76546  10.38620  13.95235
  15.86326   8.37642  10.64314  16.57776  12.47317  11.60722
  12.18313  10.38219   5.78083   5.91987  12.38836  11.00146
   7.90016   8.96447  11.48305  14.86468   8.12733  10.69868
  10.74629   5.86495   4.02501  11.69355  15.09608   7.68292
  13.83797  10.06154  15.18804   7.69657   9.12563   9.31729
  10.36702  14.31983  12.18753   9.79069   4.42008  14.16544
  10.70023   8.64389   5.13957   8.50520  11.81053  12.03744
  10.32967  11.18203   9.13243   9.30852  12.05967  10.51681
  11.25789   7.17938   9.99768  15.64463   6.17127   8.84607
   6.87402   9.13407  13.68619   8.33830  12.18405    .79262
  12.31962  10.78885   8.64017  13.49714  10.26625  12.70168
  10.58027   9.13014   6.90645   8.57410  13.13434  10.21526
  13.35620  12.52760   8.36794  15.90941  12.06191   9.20603
  13.72725  14.04590  14.44035   5.06786  12.20868   6.68642
  11.91668  12.67990  15.48195  14.67340  12.20061   9.51926
   5.03732  11.95755  10.58181   9.45669
```

Figure 28

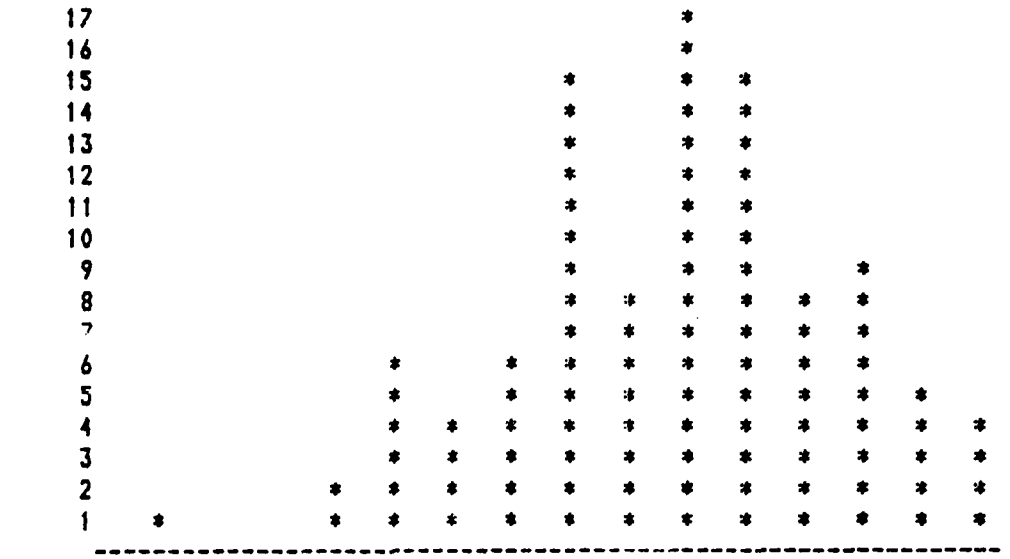
Example 1

```
*****************************************************************
*      G O O D N E S S   O F   F I T   P R O G R A M            *
*                                                               *
*  REWRITTEN AND EXPANDED BY:  SUE D. GUTHRIE                   *
*                              UNIV SOUTHERN MISS (1979)        *
*  PREVIOUS ENHANCEMENTS BY:   RALPH B. BISLAND,JR.             *
*                              UNIV SOUTHERN MISSISSIPPI        *
*                              LARRY SCHEUERMANN                *
*                              NICHOLS STATE UNIV               *
*  ORIGINALLY WRITTEN BY:      DON T. PHILLIPS                  *
*                              TEXAS A&M UNIV                   *
*---------------------------------------------------------------*
*                                                               *
*      S U M M A R Y   O F   I N P U T   P A R M S              *
*                                                               *
*      DISTRIBUTION TO BE TESTED - - - - - - - - NORMAL         *
*      TYPE OF TEST(S) - - - - - - - - - - - - - 1 CHI-SQRE     *
*                                                2 K AND S      *
*      TYPE OF ESTIMATION  - - - - - - - - - -   UNBIASED       *
*      CLASSIFICATION OF INPUT DATA  - - - - -  UNGROUPED       *
*      NUMBER OF CELLS - - - - - - - - - - - - UNSPECIFIED      *
*      NUMBER OF DATA POINTS - - - - - - - - -         100      *
*      MEAN OF POPULATION  - - - - - - - - - -      10.000      *
*      VARIANCE OF POPULATION  - - - - - - - -      10.240      *
*      ALTERNATE DEVICE NUMBER - - - - - - - -          10      *
*      MODE OF OPERATION - - - - - - - - - - - INTERACTIVE      *
*      INPUT FORMAT FOR DATA - - - - - - - - -     8F10.5      *
*      HISTOGRAM REQUESTED - - - - - - - - - -         YES      *
*      LEVEL OF SIGNIFICANCE - - - - - - - - -         .01      *
*                                                               *
*****************************************************************
ORDERED DATA

    .79262    4.02501    4.42008    5.03732    5.06786    5.13957
   5.78083    5.86495    5.91987    6.17127    6.68642    6.87402
   6.90645    7.17938    7.56042    7.68292    7.69657    7.90016
   8.12733    8.33830    8.36794    8.37642    8.50520    8.57410
   8.64017    8.64389    8.84607    8.96447    9.12563    9.13014
   9.13020    9.13243    9.13407    9.20603    9.30852    9.31729
   9.45669    9.51926    9.79069    9.99768   10.06154   10.21526
  10.26625   10.32967   10.36702   10.38219   10.38620   10.51681
  10.58027   10.58181   10.64314   10.69868   10.70023   10.74629
  10.78885   11.00146   11.11718   11.18203   11.25789   11.48305
  11.60722   11.69355   11.81053   11.91668   11.95755   12.03744
  12.05967   12.06191   12.18313   12.18405   12.18753   12.20061
  12.20868   12.31962   12.38836   12.47317   12.52760   12.67990
  12.70168   12.76546   13.13434   13.35620   13.49714   13.68619
  13.72725   13.83797   13.95235   14.04590   14.16544   14.31983
  14.44035   14.67340   14.86468   15.09608   15.18804   15.48195
  15.64463   15.86326   15.90941   16.57776
```

Figure 28 (continued)

```
**********************************************************************

        ----- DESCRIPTIVE SAMPLE STATISTICS -----

MEAN - - - - - - - - - - - - - - - -   10.53070
VARIANCE - - - - - - - - - - - - - -    9.04810
STANDARD DEVIATION - - - - - - - - -    3.00801
COEFFICIENT OF VARIATION - - - - - -     .28564
UNBIASED COEFFICIENT OF SKEWNESS -     -.39361
UNBIASED COEFFICIENT OF KURTOSIS -      .21009


**********************************************************************


**********************************************************************


        ----- HYPOTHESIS STATEMENT -----


H(0): POPULATION IS NORMAL
      WITH THEORETICAL MEAN OF    10.0000
      AND THEORETICAL VARIANCE OF    10.2400

H(1): POPULATION IS NOT NORMAL
      WITH THEORETICAL MEAN OF    10.0000
      AND THEORETICAL VARIANCE OF    10.2400


**********************************************************************
```

Figure 28 (continued)



NUMBER	1.	0.	0.	2.	6.	4.	6.	15.	8.	17.	15.	8.	9.	5.	4.
CLASS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

LOWER LIMIT OF FIRST CLASS = -9999.999
UPPER LIMIT OF LAST CLASS = 9999.999
NUMBER OF OBSERVATIONS = 100
MINIMUM OBSERVED VALUE = .793
MAXIMUM OBSERVED VALUE = 16.578

****

----- RESULTS OF KOLMOGOROV-SMIRNOV TEST -----

AT THE ALPHA = .01 LEVEL OF SIGNIFICANCE
THE KOLMOGOROV-SMIRNOV CRITICAL VALUE = .404
FOR A SAMPLE SIZE OF 15

THE COMPUTED K-S TEST STATISTIC = .113

BECAUSE THE K-S TEST STATISTIC IS LESS THAN
THE K-S CRITICAL VALUE, THERE IS INSUFFICIENT
EVIDENCE TO REJECT THE NULL HYPOTHESIS.

****

Figure 28 (continued)

```
**********************************************************************

      ----- RESULTS OF THE CHI-SQUARE TEST -----

   AT THE ALPHA =  .01 LEVEL OF SIGNIFICANCE
   FOR   9 DEGREES OF FREEDOM, THE CHI-SQUARE
   CRITICAL VALUE =  21.7

   THE COMPUTED CHI-SQUARE STATISTIC =  10.49


   BECAUSE THE TEST STATISTIC IS LESS THAN
   THE CRITICAL VALUE, THERE IS INSUFFICIENT
   EVIDENCE TO REJECT THE NULL HYPOTHESIS.
**********************************************************************
** THE NUMBER OF CELLS HAD TO BE ADJUSTED TO MEET
   THE REQUIREMENTS OF THE CHI-SQUARE TEST. THE HISTOGRAM
   BELOW SHOWS HOW THE EMPIRICAL DATA WAS GROUPED TO MEET
   THE CHI-SQUARE TEST REQUIREMENTS. **
       *****   HISTOGRAM   *****

 17                        *
 16                        *
 15              *         *   *
 14              *         *   *
 13              *         *   *
 12              *         *   *
 11              *         *   *
 10        *     *         *   *
  9        *     *         *   *       *   *
  8        *     *   *     *   *   *   *   *
  7        *     *   *     *   *   *   *   *
  6        *  *  *   *     *   *   *   *   *
  5        *  *  *   *     *   *   *   *   *
  4        *  *  *   *     *   *   *   *   *
  3  *     *  *  *   *     *   *   *   *   *
  2  *     *  *  *   *     *   *   *   *   *
  1  *     *  *  *   *     *   *   *   *   *
    ----------------------------------------
NUMBER   3. 10.  6. 15.  8. 17. 15.  8.  9.  9.

CLASS     1   2   3   4   5   6   7   8   9  10

LOWER LIMIT OF FIRST CLASS = -9999.999
UPPER LIMIT OF LAST CLASS =   9999.999
NUMBER OF OBSERVATIONS =           100
MINIMUM OBSERVED VALUE =          .793
MAXIMUM OBSERVED VALUE =        16.578
>STO;
STO;
    ***** GOODNESS OF FIT PROGRAM FINISHED *****
```

Figure 28 (continued)

B. Example 2

The data values in the second example are the same observations shown in Figure 28 for example 1. However, in example 2, the strategy in executing the GOF program is changed. The hypothesized distribution of the 100 sample observations is not specified during the initial phase of GOF execution. The user wishes to see a histogram of the 100 values before the null hypothesis is formulated. Therefore, the selection of the hypothesized distribution is delayed. Figure 29 contains the GOF language statements and the output for example 2.

Example 2 demonstrates the use of several unabbreviated GOF commands; example 1 in Figure 28 illustrates the use of GOF three-character shorthand commands. The following sentences provide an explanation of the nine GOF commands of example 2.

1. The number of data values is known to be 100.
2. The second GOF statement requests the execution of the Kolmogorov-Smirnov and chi-square goodness of fit tests.
3. The 100 sample observations are to be input from logical unit 10.
4. The test statistics are to be evaluated at the 0.01 level of significance.
5. The GOF program is to run in the interactive mode.
6. A dump of the input parameters and program defaults is desired.
7. Both requested goodness of fit tests are to be executed using 10 cells for data grouping.
8. Biased estimators are to be calculated for the coefficients of skewness and kurtosis. Section II gives the equations used for these calculations.
9. The START command is the signal to begin execution of the calculation portion of the GOF program.

The 100 observations are printed and are followed by a synopsis of the input parameters which are being used for this example. Because the hypothesized distribution is not selected at the onset of example 2, it is listed as "PENDING". A comparison of the SUMMARY OF INPUT PARMS block in Figure 29 with the one in Figure 28 illustrates several variations.

1. The coefficients of skewness and kurtosis are unbiased values in example 1 and biased estimators for example 2.
2. In example 1, the number of cells is left to the GOF program default value of 15 cells. In example 2, the number of cells is specified to be 10.
3. The population parameters are known in example 1; example 2 lists the population parameters as "PENDING". It is illogical to input theoretical population parameters if the hypothesized probability distribution is not selected.

The 100 sample observations are printed in ascending order following the input parameters summary. The sample statistics for example 2 include the biased coefficients for skewness and kurtosis instead of the unbiased ones printed in Figure 28 for example 1.

The histogram of empirical data follows the sample statistics. It is composed of 10 groups as designated by the CELLS:10; command. In the five summary values following the histogram, the GOF program shows the lower limit of the first class as .000 and the upper limit of the last class as 16.578 which is the maximum observed data value. In this example, at this point in execution, the GOF program does not know a hypothesized distribution. Therefore, the upper limit of the last class is not extended over the defined range for the random variable as it is in example 1. The lower limit of the first class is extended, but no farther than zero. The range of all possible hypothesized random variables in the GOF program includes zero.*

At this point in the output, the GOF program notifies the user that the selection of a hypothesized distribution must be made before program execution can continue. The system prompt character signals the user to input, via the TEST statement, the distribution to be tested. Example 2 is a test for the normal distribution. The parameters of the distribution are not known in this example and must be estimated from the sample observations. The START command follows the TEST instruction and the GOF program continues execution.

The GOF program performs the requested goodness of fit tests using sample estimates for the mean and variance of the hypothesized normal distribution. The hypothesis statement is phrased with the mean and variance as estimated parameters. Example 1 in Figure 28 reflects the hypothesis statement with the population parameters as known values.

The Kolmogorov-Smirnov and chi-square test results and conclusions are printed next. A note to the user signals that the number of cells is adjusted for the chi-square test. The number of cells for example 2 is specified at 10, but the data is regrouped into 7 cells before the chi-square test requirements are satisfied. Because of this adjustment, the GOF program prints an adjusted histogram representative of the chi-square grouping.

Example 2 terminates after accepting a STOP command. The delayed distribution selection option which example 2 illustrates provides the user the opportunity to study the descriptive sample statistics and histogram of the empirical data before the establishment of the null hypothesis. Figure 13 in Section III of this report provides characteristic graphs of the ten GOF probability distributions with which to compare the shape of the printed histogram. Section III also contains a discussion of the characteristics of the coefficients of skewness, kurtosis, and variation for most of the ten distributions. Therefore, once the

---

*The log-normal distribution cannot be handled at zero by the GOF program. Any zero value encountered in log-normal data is transformed to -10.0. A discussion of the transformation of log-normal data occurs in Section IV of this report as a footnote to the explanation of the TEST command.

histogram and sample descriptive statistics are reviewed, the user can better select the distribution to be tested. In the interactive mode, most operating systems have a time limit in which the user must respond. Therefore, if the user needs to spend some quantity of time studying the histogram and sample statistics, he should terminate the GOF program with a STOP command and reenter the job after he studies the intermediate results.

```
     *****  GOODNESS OF FIT PROGRAM  *****
>DATA POINTS:100;
 DATA POINTS:100;

>RUN:KAS:CHI;
 RUN:KAS:CHI;

>ALTERNATE DATA DEVICE:10;
 ALTERNATE DATA DEVICE:10;

>ALPHA:0.01;
 ALPHA:0.01;

>INTERACTIVE;
 INTERACTIVE;

>DUMPIT;
 DUMPIT;

>CELLS:10;
 CELLS:10;

>BIASED ESTIMATORS;
 BIASED ESTIMATORS;

>START;
 START;
 OBSERVED DATA

   7.56042  11.11718   9.13020  12.76546  10.38620  13.95235
  15.86326   8.37642  10.64314  16.57776  12.47317  11.60722
  12.18313  10.38219   5.78083   5.91987  12.38836  11.00146
   7.90016   8.96447  11.48305  14.86468   8.12733  10.69868
  10.74629   5.86495   4.02501  11.69355  15.09608   7.68292
  13.83797  10.06154  15.18804   7.69657   9.12563   9.31729
  10.36702  14.31983  12.18753   9.79069   4.42008  14.16544
  10.70023   8.64389   5.13957   8.50520  11.81053  12.03744
  10.32967  11.18203   9.13243   9.30852  12.05967  10.51681
  11.25789   7.17938   9.99768  15.64463   6.17127   8.84607
   6.87402   9.13407  13.68619   8.33830  12.18405   .79262
  12.31962  10.78885   8.64017  13.49714  10.26625  12.70168
  10.58027   9.13014   6.90645   8.57410  13.13434  10.21526
  13.35620  12.52760   8.36794  15.90941  12.06191   9.20603
  13.72725  14.04590  14.44035   5.06786  12.20868   6.68642
  11.91668  12.67990  15.48195  14.67340  12.20061   9.51926
   5.03732  11.95755  10.58181   9.45669
```

Figure 29

Example 2

```
*****************************************************************
*      G O O D N E S S   O F   F I T   P R O G R A M            *
*                                                               *
*  REWRITTEN AND EXPANDED BY:  SUE D. GUTHRIE                   *
*                              UNIV SOUTHERN MISS (1979)        *
*  PREVIOUS ENHANCEMENTS BY:   RALPH B. BISLAND,JR.             *
*                              UNIV SOUTHERN MISSISSIPPI        *
*                              LARRY SCHEUERMANN                *
*                              NICHOLS STATE UNIV               *
*  ORIGINALLY WRITTEN BY:      DON T. PHILLIPS                  *
*                              TEXAS A&M UNIV                   *
*---------------------------------------------------------------*
*                                                               *
*      S U M M A R Y   O F   I N P U T   P A R M S              *
*                                                               *
*      DISTRIBUTION TO BE TESTED - - - - - - - - PENDING        *
*      TYPE OF TEST(S) - - - - - - - - - - - - - 1 CHI-SQRE     *
*                                                2 K AND S      *
*      TYPE OF ESTIMATION - - - - - - - - - - -   BIASED        *
*      CLASSIFICATION OF INPUT DATA - - - - - -  UNGROUPED      *
*      NUMBER OF CELLS - - - - - - - - - - - -         10       *
*      NUMBER OF DATA POINTS - - - - - - - - -        100       *
*      PARAMETERS OF POPULATION - - - - - - -   PENDING         *
*      ALTERNATE DEVICE NUMBER - - - - - - - -         10       *
*      MODE OF OPERATION - - - - - - - - - - - INTERACTIVE      *
*      INPUT FORMAT FOR DATA - - - - - - - - -     8F10.5       *
*      HISTOGRAM REQUESTED - - - - - - - - - -        YES       *
*      LEVEL OF SIGNIFICANCE - - - - - - - - -        .01       *
*                                                               *
*****************************************************************
ORDERED DATA

   .79262    4.02501    4.42008    5.03732    5.06786    5.13957
  5.78083    5.86495    5.91987    6.17127    6.68642    6.87402
  6.90645    7.17938    7.56042    7.68292    7.69657    7.90016
  8.12733    8.33830    8.36794    8.37642    8.50520    8.57410
  8.64017    8.64389    8.84607    8.96447    9.12563    9.13014
  9.13020    9.13243    9.13407    9.20603    9.30852    9.31729
  9.45669    9.51926    9.79069    9.99768   10.06154   10.2[illegible]526
 10.26625   10.32967   10.36702   10.38219   10.38620   10.51681
 10.58027   10.58181   10.64314   10.69868   10.70023   10.74629
 10.78885   11.00146   11.11718   11.18203   11.25789   11.48305
 11.60722   11.69355   11.81053   11.91668   11.95755   12.03744
 12.05967   12.06191   12.18313   12.18405   12.18753   12.20061
 12.20868   12.31962   12.38836   12.47317   12.52760   12.67990
 12.70168   12.76546   13.13434   13.35620   13.49714   13.68619
 13.72725   13.83797   13.95235   14.04590   14.16544   14.31983
 14.44035   14.67340   14.86468   15.09608   15.18804   15.48195
 15.64463   15.86326   15.90941   16.57776
```

Figure 29 (continued)

```
**********************************************************************

        ----- DESCRIPTIVE SAMPLE STATISTICS -----

MEAN - - - - - - - - - - - - - - - -   10.53070
VARIANCE - - - - - - - - - - - - - -    9.04810
STANDARD DEVIATION - - - - - - - - -    3.00801
COEFFICIENT OF VARIATION - - - - - -     .28564
BIASED COEFFICIENT OF SKEWNESS - -     -.38768
BIASED COEFFICIENT OF KURTOSIS - -       .14031


**********************************************************************

         *****   HISTOGRAM   *****

        EACH * REPRESENTS  2 POINTS

    22                          *
    20                          *   *
    18                          *   *
    16                      *   *   *
    14                  *   *   *   *
    12                  *   *   *   *   *
    10                  *   *   *   *   *
     8              *   *   *   *   *   *   *
     6          *   *   *   *   *   *   *   *
     4          *   *   *   *   *   *   *   *
     2  *       *   *   *   *   *   *   *   *
      ----------------------------------------
NUMBER   1.  0.  5.  7. 13. 16. 21. 19. 11.  7.

CLASS     1   2   3   4   5   6   7   8   9  10

LOWER LIMIT OF FIRST CLASS =       .000
UPPER LIMIT OF LAST CLASS =      16.578
NUMBER OF OBSERVATIONS =            100
MINIMUM OBSERVED VALUE =           .793
MAXIMUM OBSERVED VALUE =         16.578


 ** SELECTION OF DISTRIBUTION FOR NULL HYPOTHESIS
    MUST BE MADE NOW **
>TEST:NORMAL;
 TEST:NORMAL;

>START;
 START;
 ** PARAMETERS OF TEST DISTRIBUTION NOT SPECIFIED
    DEFAULT TO ESTIMATES FROM SAMPLE DATA. **
```

Figure 29 (continued)

```
**********************************************************************

        ----- HYPOTHESIS STATEMENT -----

H(0): POPULATION IS NORMAL
      WITH  ESTIMATED  MEAN OF    10.5307
      AND  ESTIMATED  VARIANCE OF     9.0481

H(1): POPULATION IS NOT NORMAL
      WITH  ESTIMATED  MEAN OF    10.5307
      AND  ESTIMATED  VARIANCE OF     9.0481

**********************************************************************

**********************************************************************

        ----- RESULTS OF KOLMOGOROV-SMIRNOV TEST -----

   AT THE ALPHA =  .01 LEVEL OF SIGNIFICANCE
   THE KOLMOGOROV-SMIRNOV CRITICAL VALUE =  .490
   FOR A SAMPLE SIZE OF  10

   THE COMPUTED K-S TEST STATISTIC =    .045


   BECAUSE THE K-S TEST STATISTIC IS LESS THAN
   THE K-S CRITICAL VALUE, THERE IS INSUFFICIENT
   EVIDENCE TO REJECT THE NULL HYPOTHESIS.

**********************************************************************

** NOT ENOUGH DATA FOR CELL SPECIFICATION OF   10
   CHI SQUARE TEST WILL BE RUN WITH    7 CELLS. **
```

Figure 29 (continued)

```
**********************************************************************

        ----- RESULTS OF THE CHI-SQUARE TEST -----

   AT THE ALPHA =  .01 LEVEL OF SIGNIFICANCE
   FOR   4 DEGREES OF FREEDOM, THE CHI-SQUARE
   CRITICAL VALUE =  13.3

   THE COMPUTED CHI-SQUARE STATISTIC =   1.31


   BECAUSE THE TEST STATISTIC IS LESS THAN
   THE CRITICAL VALUE, THERE IS INSUFFICIENT
   EVIDENCE TO REJECT THE NULL HYPOTHESIS.

**********************************************************************


** THE NUMBER OF CELLS HAD TO BE ADJUSTED TO MEET
   THE REQUIREMENTS OF THE CHI-SQUARE TEST. THE HISTOGRAM
   BELOW SHOWS HOW THE EMPIRICAL DATA WAS GROUPED TO MEET
   THE CHI-SQUARE TEST REQUIREMENTS. **
         *****   HISTOGRAM   *****

        EACH * REPRESENTS  2 POINTS

    22                *
    20                *   *
    18                *   *
    16            *   *   *
    14    *   *   *   *   *
    12    *   *   *   *   *   *
    10    *   *   *   *   *   *
     8    *   *   *   *   *   *   *
     6    *   *   *   *   *   *   *
     4    *   *   *   *   *   *   *
     2    *   *   *   *   *   *   *
      ------------------------------
NUMBER  13.  13.  16.  21.  19.  11.   7.

CLASS     1   2   3   4   5   6   7

LOWER LIMIT OF FIRST CLASS = -9999.999
UPPER LIMIT OF LAST CLASS =   9999.999
NUMBER OF OBSERVATIONS =           100
MINIMUM OBSERVED VALUE =          .793
MAXIMUM OBSERVED VALUE =        16.578
>STOP;
STOP;

    ***** GOODNESS OF FIT PROGRAM FINISHED *****
```

Figure 29 (continued)

C. Example 3

The sample observations for example 3 are only available in a grouped format. If data exists in both grouped and ungrouped forms, the use of ungrouped data is preferable with the GOF program for four primary reasons.

1. The values calculated for the descriptive sample statistics are more accurate for ungrouped data.
2. The coefficients of skewness and kurtosis are not computed by the GOF program for grouped data. Therefore, the moments test for normality cannot be run for grouped data.
3. The Cramer-Von Mises goodness of fit test is not valid for grouped data.
4. The Kolmogorov-Smirnov goodness of fit test is more powerful when run on ungrouped data.

Section II of this report contains a description of the characteristics of the four goodness of fit tests available from the GOF program.

In example 3, shown in Figure 30, the GOF program expects to read 75 input values. For grouped data, the number of input values equals three times the number of groups (or cells) into which the data is arranged. Grouped data values are input in a mandatory order of lower class boundary, upper class boundary, and absolute class frequency for the first through the last class. These three values are necessary for each group in the data set. Grouped data may also be input using the COUNT option which is explained in Section IV of this report. Example 3 specifies that 75 values are to be read. Therefore, 25 cells are implied. The number of grouped data values to be read must always be a multiple of 3.

The input values for example 3 are read from logical unit 14. The format of data values is free field. The GOF program expects grouped data to be input because of the GRO statement in Figure 30. All goodness of fit tests are to be run at a 0.01 value of ALPHA. Three goodness of fit tests are to be run: (1) the Kolmogorov-Smirnov test, (2) the chi-square test, and (3) the Cramer-Von Mises test. The data in example 3 are hypothesized to come from a uniform distribution whose lower limit is 10.0 and whose upper limit is 45.0. The parameters of the uniform distribution in this example could also be input using the following two statements.

1. LOWER LIMIT: 10.0;, and
2. UPPER LIMIT: 45.0;.

The effect is the same regardless of the technique used to input known population parameters. For this example, a histogram is not desired. The GOF program is to run in the interactive mode. After the problem definition, the START instruction is input and execution of the computation section begins.

Immediately, the GOF program disallows the Cramer-Von Mises goodness of fit test because it is not valid for grouped data. The program continues to run to perform the Kolmogorov-Smirnov and chi-square tests. The input data is printed in a special grouped format.

Following the preliminaries, the descriptive sample statistics are computed and printed. The coefficient of variation is interesting in this example. Because the hypothesized uniform population parameters are given as a=10.0 (lower limit) and b=45.0 (upper limit), the population coefficient of variation can be calculated by the formula:

$$\text{Coefficient of variation} = (b-a) / 3^{\frac{1}{2}}(a+b).$$

For example 3, this equation yields a theoretical coefficient of variation equal to 0.367405. This value is very close to the coefficient of variation statistic computed from the sample observations.

The null and alternative hypotheses are given for the example being run. After the hypothesis statements, the results of the Kolmogorov-Smirnov test is given and evaluated at a probability of 0.01. A sample size of 25 is appropriate for locating the correct critical value because the data is grouped into 25 cells.

No additional grouping is required to execute the chi-square test. Therefore, the chi-square critical value is provided at the 0.01 level and the appropriate degrees of freedom value is 24. The computed chi-square test statistic is compared with the critical value, and because the computed value is less, the null hypothesis cannot be rejected. There is insufficient evidence to reject the assumption that the input data is from a uniform distribution with a known lower limit of 10.0 and a known upper limit of 45.0. The GOF program is terminated in the usual manner with a STOP command.

```
      *****  GOODNESS OF FIT PROGRAM  *****


>DAT:75;
 DAT:75;

>ALT:14;
 ALT:14;

>FRE;
 FRE;

>GRO;
 GRO;

>ALP:0.01;
 ALP:0.01;

>RUN:KAS:CHI:CVM;
 RUN:KAS:CHI:CVM;

>TES:UNI:10.0:45.0;
 TES:UNI:10.0:45.0;

>NOH;
 NOH;

>INT;
 INT;

>STA;
 STA;

 ** THE CRAMER-VON MISES TEST CANNOT BE RUN
    ON GROUPED DATA - REQUEST DENIED. **
```

Figure 30

Example 3

GROUPED INPUT DATA

FROM	TO	ABSOLUTE FREQUENCY
10.54775	11.92441	11.
11.92441	13.30107	10.
13.30107	14.67774	4.
14.67774	16.05440	5.
16.05440	17.43106	5.
17.43106	18.80772	7.
18.80772	20.18438	7.
20.18438	21.56104	10.
21.56104	22.93771	13.
22.93771	24.31437	6.
24.31437	25.69103	12.
25.69103	27.06769	6.
27.06769	28.44435	7.
28.44435	29.82102	12.
29.82102	31.19768	5.
31.19768	32.57434	7.
32.57434	33.95100	6.
33.95100	35.32766	11.
35.32766	36.70432	6.
36.70432	38.08099	7.
38.08099	39.45765	8.
39.45765	40.83431	9.
40.83431	42.21097	11.
42.21097	43.58763	9.
43.58763	44.96429	6.

****

----- DESCRIPTIVE SAMPLE STATISTICS -----

MEAN - - - - - - - - - - - - - - - -	27.92810
VARIANCE - - - - - - - - - - - - - -	99.00627
STANDARD DEVIATION - - - - - - - -	9.95019
COEFFICIENT OF VARIATION - - - - -	.35628

****

Figure 30 (continued)

```
**********************************************************************
        ----- HYPOTHESIS STATEMENT -----


H(0): POPULATION IS UNIFORM
      WITH THEORETICAL LOWER LIMIT OF    10.0000
      AND THEORETICAL UPPER LIMIT OF     45.0000

H(1): POPULATION IS NOT UNIFORM
      WITH THEORETICAL LOWER LIMIT OF    10.0000
      AND THEORETICAL UPPER LIMIT OF     45.0000


**********************************************************************
**********************************************************************

       ----- RESULTS OF KOLMOGOROV-SMIRNOV TEST -----

   AT THE ALPHA =  .01 LEVEL OF SIGNIFICANCE
   THE KOLMOGOROV-SMIRNOV CRITICAL VALUE =  .320
   FOR A SAMPLE SIZE OF  25

   THE COMPUTED K-S TEST STATISTIC =    .046


   BECAUSE THE K-S TEST STATISTIC IS LESS THAN
   THE K-S CRITICAL VALUE, THERE IS INSUFFICIENT
   EVIDENCE TO REJECT THE NULL HYPOTHESIS.


**********************************************************************
**********************************************************************

       ----- RESULTS OF THE CHI-SQUARE TEST -----

   AT THE ALPHA =  .01 LEVEL OF SIGNIFICANCE
   FOR  24 DEGREES OF FREEDOM, THE CHI-SQUARE
   CRITICAL VALUE =  43.0

   THE COMPUTED CHI-SQUARE STATISTIC =  19.49


   BECAUSE THE TEST STATISTIC IS LESS THAN
   THE CRITICAL VALUE, THERE IS INSUFFICIENT
   EVIDENCE TO REJECT THE NULL HYPOTHESIS.


**********************************************************************
>STO;
 STO;

    ***** GOODNESS OF FIT PROGRAM FINISHED *****
```

Figure 30 (continued)

D. Example 4

This sample run uses all four of the goodness of fit techniques available from the GOF program and thus provides a sample of each type of output as shown in Figure 31. In this example, 55 sample observations are believed to be Poisson distributed with a known mean of 3.4. The data values are to be read from logical unit 11 and the GOF program is to run in the interactive mode.

The probability of rejecting the null hypothesis when it is true (ALPHA) is not specified. The GOF program defaults to an alpha value of 0.05 at which to evaluate the results of the goodness of fit tests.

The sample observations are printed in the order they are read, followed by a second printing of them in sorted order. The sample statistics hint that the mean of the sample data is very close to the known mean of the hypothesized Poisson distribution. The calculation of the coefficient of variation for the hypothesized distribution yields 0.54233, which is close to the 0.62982 value computed from the sample observations. The sample coefficient of skewness is positive, suggesting a skew to the right. A positive coefficient of kurtosis denotes a platykurtic shape. All three sample coefficients seem to support the assumption of a Poisson distribution. Section II of this report contains a discussion of the meaning of the coefficients of skewness and kurtosis. Section III provides equations for calculating these three coefficients, using theoretical population parameters for most of the ten GOF distributions.

The moments test for normality is run in example 4 for illustrative reasons. Rejection of normality occurs for the skewness part of the test, but does not hold for the kurtosis evaluation. These results are inconclusive and are only included to illustrate the output of the moments test. These results also support the comments in Section II of this report that this test is best considered a test of nonnormality. Trying to accept a null hypothesis with conflicting results, similar to the ones in this example, leaves the researcher with a questionable premise.

The hypothesis statement is followed by a histogram of the empirical data in Figure 31. The Poisson distribution is treated differently from the other nine distributions by the GOF program. This difference is explained in Section IV of this report under the description of Block 19 of the macro flowchart. For the Poisson distribution, the GOF program establishes a cell for each possible discrete data value over the range of 0 to the maximum data point and the absolute frequencies are tabulated accordingly. Therefore, there will be as many cells as there are possible unique integer values in the defined range, unless the user specifies (via the CELLS statement) that the program is to run with less cells. In this one case, the number of cells is reduced to meet the user's request. The histogram in Figure 31 illustrates the establishment of 11 cells for this test because there are 11 possible integer values between 0 and 10.0 (the maximum value in this example's data set).

The Kolmogorov-Smirnov test results are printed for an alpha value of 0.05 and a sample size of 11--the number of cells into which the data is grouped. The Kolmogorov-Smirnov statistic is much less than the Kolmogorov-Smirnov critical value and the user must conclude in this case that there is insufficient evidence to reject the null hypothesis.

The Cramer-Von Mises and chi-square tests also lead to the conclusion that the null hypothesis cannot be rejected. Therefore, the results of example 4 suggest that the 55 sample observations do come from a Poisson distribution with a known mean of 3.4. Figure 31 has a second histogram showing the effects of regrouping the data to meet the chi-square test requirements. The summary values following the histogram give the possible range of the random variables as $0<x<\infty$, which is the allowable range for the Poisson distribution.

The GOF program concludes with the acceptance of a STOP instruction. The user could run another test by issuing the appropriate testing and operational commands instead of the STOP statement.

Example 4 is included to illustrate the output from each of the four goodness of fit techniques which are available from the GOF program. The user should not arbitrarily apply goodness of fit tests to any distribution. For example, the sample run in example 4 uses the Cramer-Von Mises test with the Poisson distribution which violates the restriction that the Cramer-Von Mises test be used only with continuous distributions. Caution should be exercised in determining which goodness of fit tests are best suited to individual circumstances.

```
     *****  GOODNESS OF FIT PROGRAM  *****
>DAT:55;
 DAT:55;

>RUN:CHI:KAS:MOM:CVM;
 RUN:CHI:KAS:MOM:CVM;

>TES:POI:3.4;
 TES:POI:3.4;

>ALT:11;
 ALT:11;

>INT;
 INT;

>STA;
 STA;

 ** ALPHA VALUE (PROBABILITY OF TYPE I ERROR) NOT SPECIFIED AS
    EITHER 0.01 OR 0.05 - DEFAULT IS 0.05 **
 OBSERVED DATA

    3.00000  10.00000   2.00000   3.00000   2.00000   3.00000
    1.00000   6.00000   2.00000   3.00000   3.00000   2.00000
    3.00000   3.00000   4.00000    .00000   3.00000   4.00000
    5.00000   2.00000   1.00000   3.00000   6.00000   2.00000
    1.00000   4.00000   1.00000   4.00000   4.00000   3.00000
    9.00000   4.00000   4.00000   4.00000   1.00000   8.00000
    4.00000   1.00000   2.00000   6.00000   2.00000   7.00000
     .00000   3.00000   1.00000   4.00000   4.00000   1.00000
    3.00000   3.00000   4.00000   6.00000   7.00000   2.00000
    3.00000

 ORDERED DATA

     .00000    .00000   1.00000   1.00000   1.00000   1.00000
    1.00000   1.00000   1.00000   1.00000   2.00000   2.00000
    2.00000   2.00000   2.00000   2.00000   2.00000   2.00000
    2.00000   3.00000   3.00000   3.00000   3.00000   3.00000
    3.00000   3.00000   3.00000   3.00000   3.00000   3.00000
    3.00000   3.00000   3.00000   4.00000   4.00000   4.00000
    4.00000   4.00000   4.00000   4.00000   4.00000   4.00000
    4.00000   4.00000   4.00000   5.00000   6.00000   6.00000
    6.00000   6.00000   7.00000   7.00000   8.00000   9.00000
   10.00000
```

Figure 31

Example 4

```
**********************************************************************

        ----- DESCRIPTIVE SAMPLE STATISTICS -----

MEAN - - - - - - - - - - - - - - -    3.38182
VARIANCE - - - - - - - - - - - - -    4.53670
STANDARD DEVIATION - - - - - - - -    2.12995
COEFFICIENT OF VARIATION - - - - -     .62982
UNBIASED COEFFICIENT OF SKEWNESS -    1.05270
UNBIASED COEFFICIENT OF KURTOSIS -    1.31165

**********************************************************************

**********************************************************************

        ----- RESULTS OF MOMENTS TEST -----

   FOR SKEWNESS, IF  -.633 LE   1.0527 LE  .633 ONE CANNOT
   REJECT NORMALITY AT THE 95 PER CENT SIGNIFICANCE LEVEL
   WITH A SAMPLE SIZE OF   55

   FOR KURTOSIS, IF  -.89 LE   1.3117 LE 1.58 ONE CANNOT
   REJECT NORMALITY AT THE 95 PER CENT SIGNIFICANCE LEVEL
   WITH A SAMPLE SIZE OF   55

**********************************************************************

**********************************************************************

        ----- HYPOTHESIS STATEMENT -----

H(0): POPULATION IS POISSON
      WITH THEORETICAL MEAN OF       3.40

H(1): POPULATION IS NOT POISSON
      WITH THEORETICAL MEAN OF       3.40

**********************************************************************
```

Figure 31 (continued)

```
      *****   HISTOGRAM   *****

   14                 *
   13                 *
   12                 *   *
   11                 *   *
   10                 *   *
    9             *   *   *
    8         *   *   *   *
    7         *   *   *   *
    6         *   *   *   *
    5         *   *   *   *
    4         *   *   *   *       *
    3         *   *   *   *       *
    2     *   *   *   *   *       *   *
    1     *   *   *   *   *   *   *   *   *   *   *
      ---------------------------------------------

NUMBER   2.  8.  9. 14. 12.  1.  4.  2.  1.  1.  1.

CLASS     1   2   3   4   5   6   7   8   9  10  11

LOWER LIMIT OF FIRST CLASS =      .000
UPPER LIMIT OF LAST CLASS =   9999.999
NUMBER OF OBSERVATIONS =           55
MINIMUM OBSERVED VALUE =         .000
MAXIMUM OBSERVED VALUE =       10.000


**********************************************************************

        ----- RESULTS OF KOLMOGOROV-SMIRNOV TEST -----

   AT THE ALPHA =  .05 LEVEL OF SIGNIFICANCE
   THE KOLMOGOROV-SMIRNOV CRITICAL VALUE =  .391
   FOR A SAMPLE SIZE OF  11

   THE COMPUTED K-S TEST STATISTIC =    .074


   BECAUSE THE K-S TEST STATISTIC IS LESS THAN
   THE K-S CRITICAL VALUE, THERE IS INSUFFICIENT
   EVIDENCE TO REJECT THE NULL HYPOTHESIS.


**********************************************************************
```

Figure 31 (continued)

```
**********************************************************************

      ----- RESULTS OF CRAMER-VON MISES TEST -----

   AT THE ALPHA = .05 LEVEL OF SIGNIFICANCE
   THE CRAMER-VON MISES CRITICAL VALUE =  .461

   THE COMPUTED CRAMER-VON MISES TEST STATISTIC =   .369


   BECAUSE THE TEST STATISTIC IS LESS THAN
   THE CRITICAL VALUE, THERE IS INSUFFICIENT
   EVIDENCE TO REJECT THE NULL HYPOTHESIS.

**********************************************************************


**********************************************************************

      ----- RESULTS OF THE CHI-SQUARE TEST -----

   AT THE ALPHA =  .05 LEVEL OF SIGNIFICANCE
   FOR   5 DEGREES OF FREEDOM, THE CHI-SQUARE
   CRITICAL VALUE =  11.1

   THE COMPUTED CHI-SQUARE STATISTIC =   6.93


   BECAUSE THE TEST STATISTIC IS LESS THAN
   THE CRITICAL VALUE, THERE IS INSUFFICIENT
   EVIDENCE TO REJECT THE NULL HYPOTHESIS.

**********************************************************************
```

Figure 31 (continued)

```
** THE NUMBER OF CELLS HAD TO BE ADJUSTED TO MEET
   THE REQUIREMENTS OF THE CHI-SQUARE TEST. THE HISTOGRAM
   BELOW SHOWS HOW THE EMPIRICAL DATA WAS GROUPED TO MEET
   THE CHI-SQUARE TEST REQUIREMENTS. **


        *****   HISTOGRAM   *****

    14          *
    13          *
    12          *   *
    11          *   *
    10    *     *   *
     9    *   * *   *       *
     8    *   * *   *       *
     7    *   * *   *       *
     6    *   * *   *       *
     5    *   * *   *       *
     4    *   * *   *       *
     3    *   * *   *       *
     2    *   * *   *       *
     1    *   * *   *   *   *
      ------------------------
 NUMBER  10.  9. 14. 12.  1.  9.

 CLASS    1   2   3   4   5   6

 LOWER LIMIT OF FIRST CLASS =       .000
 UPPER LIMIT OF LAST CLASS =   9999.999
 NUMBER OF OBSERVATIONS =           55
 MINIMUM OBSERVED VALUE =         .000
 MAXIMUM OBSERVED VALUE =       10.000
>STO;
 STO;


***** GOODNESS OF FIT PROGRAM FINISHED *****
```

Figure 31 (continued)

XIII. Appendix B - Univac 1108 Run Streams

A. General

The GOF program described in this report allows the user the option of inputting his empirical data values from an input device different from the one through which he enters his GOF language statements. The option is declared to the GOF language translator through the ALTERNATE DATA DEVICE command and is explained in Section IV of this report. The command requires a numerical value which is the number of the logical unit from which the GOF program reads the empirical data values. The two run streams provided in this appendix illustrate the statements necessary to execute the GOF program with the ALTERNATE DATA DEVICE command. To eliminate the use of that option, the user omits the commands which have an asterisk beside them.

B. Input from a File

The Univac 1108 run streams for the batch and on-line versions are identical with one exception--the GOF input command statements are followed by an EOF card in the batch mode. All Univac 1108 operating system commands begin with a master space symbol (@). File names on the Univac 1108 should be followed by a period (.) as shown in Figures 32 and 33. Logical unit numbers 0, 1, 5, 6, and 30 should not be used as an alternate device number because they are reserved for reread, punch, read, write, and reread units, respectively.

If the empirical data to be input resides in a file or an element of a file, the user supplies the commands shown in Figure 32. Basically, these commands cause the computer to copy the empirical data values into a file whose file name is the same as the logical unit number given in the ALTERNATE DATA DEVICE statement. The asterisks shown in Figure 32 are not part of the command, but signal that the associated statement is omitted if the user does not desire the alternate device option.

```
   @RUN XXXXXX,PPPPPPPPPPP/NNNN,MMMMMM
*  @ASG,T 10.
*  @ED,I 10.
*  @ADD FILE.ELEMENT
*  ↵
*  EXIT
   @XQT FILE.GOF

        GOF input commands

   @EOF (required for batch mode only)
   @FIN
```

Figure 32
Univac 1108 Run Stream for Input from a File

The exact format of the RUN command is unique to the host system. In Figure 32, the parameters of the RUN command have the following meanings:

1. XXXXXX is a six character run identification value,
2. PPPPPPPPPPPP is a twelve character account identification code,
3. NNNN represents a 1-12 digit user identification number, and
4. MMMMMM is a project identification entry which must not exceed twelve characters.

The ASG statement temporarily assigns to the user a file whose name is 10. The file is destroyed when the job terminates. The number 10 is the number which the user inputs with the ALTERNATE DATA DEVICE command. For example, one of the GOF input commands necessary in Figure 32 is ALT:10; which defines the alternate input unit to be 10. Section IV of this report explains the use of this command.

The ED statement in Figure 32 invokes the on-line editor which expects to accept information for file 10. The first command given to the editor is the ADD command which causes the empirical data values in FILE.ELEMENT to be copied into file 10. The user must substitute his file and element names where the letters FILE.ELEMENT appear in the ADD statement. The curved arrow (↵) represents a carriage return and is necessary to change the operating mode of the editor. The EXIT command is an instruction to the editor to close file 10 and return control to the Univac 1108 operating system. At this point, a copy of the empirical data values are in file 10.

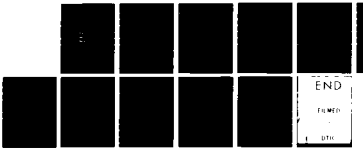
The GOF program is executed following the entry of the XQT statement. Examples of input and output for the program are given in Appendix A of this report. The GOF input commands are followed by an EOF statement only if the program is being run in the batch mode. Upon completion of a GOF session, the user exits from the Univac 1108 system with a FIN command.

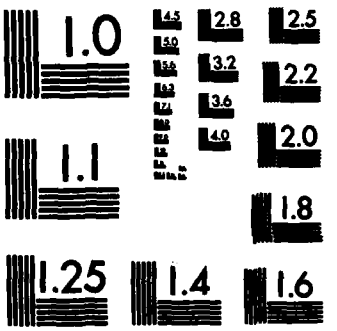
C. Input from Magnetic Tape

If the user wants to enter empirical data values which are recorded on a magnetic tape, he uses the series of Univac 1108 commands illustrated by Figure 33. The RUN, XQT, EOF, and FIN statements have the same parameter requirements and functions as those described in the preceding section of this appendix. Therefore, these four statements are not discussed for this example. Input from magnetic tape must be composed of 80-character records.

AD-A123 401 A FORTRAN COMPUTER PROGRAM TO PERFORM GOODNESS OF FIT TESTING ON EMPIRICAL DATA(U) NAVAL OCEANOGRAPHIC OFFICE NSTL STATION MS S D GUTHRIE JUN 79 NOO-TN-5000-1-79 2/2

UNCLASSIFIED F/G 9/2 NL





```
@RUN XXXXXX,PPPPP/PPPPPPP/NNNN,NNNNNN
* @ASG,T FILE.,DDD,REELNO
* @USE 12.,FILE.
@XQT YOURFL.GOF

     GOF input commands

@EOF (required for batch mode only)
* @FREE FILE.
@FIN
```

Figure 33
Univac 1108 Run Stream for Input from Magnetic Tape

For the ASG statement in Figure 33, the three necessary parameters are explained as follows.

1. FILE. represents any 1-12 character name the user wishes to assign to his magnetic tape. He might select meaningful names like DATA., SAMPLS., INPUT., or EMPVAL.
2. DDD is a 1-3 character designator for the tape drive to be used. For example, U9V specifies the use of a 9-track, very high density unit and U selects a 7-track, 800 BPI unit.
3. REELNO is a representation for the physical reel number of the magnetic tape containing the empirical data.

The USE command associates the user's file name (FILE. in this example) with the logical unit number to be input to the GOF language translator via the ALTERNATE DATA DEVICE statement. The logical unit number in Figure 33 is 12. Therefore, the GOF program is to read from logical unit 12 the empirical data values which are stored on a magnetic tape.

After execution of the GOF program is complete, the user should remember to enter the FREE command which causes the tape to be rewound and the tape drive to be released.

IX. Appendix C - Explanation of Diagnostic Messages

The GOF program verifies the syntax of GOF language commands, checks the appropriateness of many input parameters, validates the combinations of options requested, and monitors certain internal testing procedures. The user is informed of errors or changes in these areas by messages printed by the GOF program. These messages are always enclosed by a double asterisk (**). Some messages are simply informative; some are warning devices; some denote unresolvable (by the GOF program) conditions and cause an error termination of the GOF program.

This Appendix contains a listing of all possible GOF messages. The messages are presented in alphabetical order. Some are followed by additional information or suggestions for corrective actions. Others are considered self-explanatory and are not discussed. The name in parentheses following each message is the name of the GOF subprogram responsible for printing that message. In the case of the ERROR subroutine's printing of a message, a second name is included in the parentheses. The second name is the name of the subprogram which calls ERROR to print the message.

1. ** ALL GOODNESS OF FIT TESTS WILL USE TRANSFORMED DATA - DATA IS ASSUMED POISSON WITH TRANSFORMED MEAN OF XXXXX.XXXX AND ORIGINAL MEAN OF XXXXX.XXXX. ** (GFMAIN)

To prevent possible overflow during Poisson calculations, all hypothesized Poisson data sets are transformed to a minimum base of zero. This transformation is explained in Section IV of this report under the description of Block 19 of the GOF macro flowchart.

2. ** ALPHA VALUE (PROBABILITY OF TYPE I ERROR) NOT SPECIFIED AS EITHER 0.01 or 0.05 - DEFAULT IS 0.05. ** (ERROR,LANG)

Section IV of this report explains the purpose of this message under the description of the ALPHA statement.

3. ** AS SHAPE PARAMETER GETS LARGER, THE GAMMA DISTRIBUTION APPROACHES THE NORMAL DISTRIBUTION. SHAPE PARAMETER = XXXXX.XXXX, TEST IS SWITCHED TO NORMAL. ** (GFMAIN)

When the shape parameter of the gamma distribution exceeds 30.0, the normal distribution provides an excellent approximation to the gamma distribution. If the user did not input the gamma distribution parameters, the normal distribution is tested using estimates of the mean and variance calculated from the sample observations. If the theoretical shape and scale parameters are defined to the GOF program, the mean and variance are calculated using the theoretical gamma parameters. Section III of this report contains the equations for calculating the mean and variance, given the shape and scale parameters.

4. ** CHI SQUARE TEST CANNOT BE RUN - INSUFFICIENT DATA. ** (GFMAIN)

The GOF program does not execute the chi-square goodness of fit test if the number of cells falls below four. The number of sample observations is too small to perform meaningful calculations. The Cramer-Von Mises or Kolmogorov-Smirnov test is recommended. The number of cells could be reduced below four because of the regrouping which occurs in some cases to meet the chi-square test restriction of at least five expected observations per class. Section II of this report contains a discussion of the chi-square test and an explanation of this particular restriction.

5. ** CONTRADICTION: FREE FIELD READ AND FORMAT READ NO DEFAULT. ** (ERROR,LANG)

The user is requesting the GOF program to read in the sample observations using both a free field read and a user-supplied format. The command stream for the current run must be reentered correctly choosing one input mode or the other.

6. ** CONTRADICTION: NO. OF DATA POINTS SPECIFIED AND INSTRUCTION TO COUNT DATA POINTS--DEFAULT TO DATA POINTS SPECIFIED. ** (ERROR,LANG)

The user supplies the number of sample observations for the GOF program to read and also the GOF program to count the number of values it reads. If the default path is not satisfactory, the user should issue a NEXT command (explained in Section IV of this report) and redefine the current run.

7. ** DATA POINT TERMINAL VALUE NOT FOUND--NO DEFAULT. ** (ERROR,INPUT)

The GOF program is reading sample observations under control of the COUNT command, which is explained in Section IV of this report. The program encounters an end-of-file in the data stream, but does not locate the anticipated terminal code. The user must redefine the job either supplying a correct terminal value or the exact number of data points to be read.

8. ** DATA POINTS ARE IN EXCESS OF XXX NO DEFAULT. ** (ERROR, INPUT)

If the GOF program is running under the COUNT option (explained in Section IV of this report) and does not read an end-of-file mark before it reads the maximum number of allowable input values, it terminates with this printout. The maximum allowable number is 500, but may be easily changed. To change the number of allowable data points, the user must change the value of a main program (GOF) variable named MAXDAT.

He must also change the dimensioned values of these arrays: (1) X, (2) XCELS, (3) PXCELS, (4) YCELS, (5) ZCELS, (6) ZZCELS, (7) FROM, (8) FT, (9) Y, (10) XT, and (11) TO. No other changes are necessary.

9. ** DATA WILL NOT SUPPORT XXXXX CELLS. XMAX = XXXXX.XXXX, TEST CONTINUES. ** (GFMAIN)

In testing for the Poisson distribution, the user specifies a desired number of cells into which he wishes his data to be grouped. In grouping the hypothesized Poisson data, the GOF program assigns one cell to each integer value possible over the range of input values (0 to XMAX). Therefore, this message is printed when there are less possible integer values than the number of cells the user requested. The program execution continues; using all available cells unless regrouping occurs for the chi-square test.

10. ** EITHER NUMBER OF DATA POINTS OR COUNT OF DATA POINTS MUST BE SPECIFIED--NO DEFAULT. ** (ERROR,LANG)

The user is trying to excecute the GOF program without using a DATA command or a COUNT command. One statement or the other is mandatory. The user must reenter his GOF command stream with this deficiency corrected.

11. ** FOR HYPOTHESIZED DISTRIBUTION, BOTH THE SHAPE AND THE SCALE PARAMETERS MUST BE GREATER THAN ZERO. ** (ERROR,LANG)

The user is testing for the gamma, the Erlang-k, or the Weibull distribution and has input a scale or shape parameter whose value is less than or equal to zero. Neither parameter is valid for such values. The user must reenter his corrected GOF command stream.

12. ** FOR HYPOTHESIZED DISTRIBUTION, MEAN MUST BE GREATER THAN ZERO - NO DEFAULT. ** (ERROR,LANG)

The hypothesized distribution is either Poisson, exponential, or log-normal; and the user has input a negative or zero mean. The mean of these three distributions is always greater than zero. The user must reenter his corrected GOF command stream.

13. ** FOR HYPOTHESIZED DISTRIBUTION, VARIANCE MUST BE GREATER THAN ZERO - NO DEFAULT. ** (ERROR,LANG)

The user is testing for either a normal or log-normal distribution and is supplying an invalid variance value. The user must reenter his corrected GOF command stream.

14. ** FOR CHI-SQUARE DISTRIBUTION, THE NUMBER OF DEGREES OF FREEDOM MUST BE GREATER THAN OR EQUAL TO 1 - NO DEFAULT. ** (ERROR,LANG)

15. ** FOR TRIANGULAR DISTRIBUTION, MINIMUM VALUE MUST BE LESS THAN MAXIMUM VALUE - NO DEFAULT. ** (ERROR,LANG)

16. ** FOR TRIANGULAR DISTRIBUTION, MOST PROBABLE VALUE MUST LIE IN THE RANGE BETWEEN THE MINIMUM AND MAXIMUM VALUES - NO DEFAULT. ** (ERROR,LANG)

The user is testing for a triangular distribution with known parameters. The value input for the most probable parameter is either less than the minimum value or greater than the maximum value. The user must input a corrected command stream.

17. ** FOR UNIFORM DISTRIBUTION, LOWER LIMIT MUST BE LESS THAN UPPER LIMIT - NO DEFAULT. ** (ERROR,LANG)

18. ** FORMAT EXCEEDS 80 CHARACTERS -- NO DEFAULT. ** (ERROR,LANG)

In selecting the FORMAT option (explained in Section IV of this report), the user is trying to input a FORMAT designation longer than 80 characters. The command stream for the current run must be reentered with this problem corrected.

19. ** 'GAMMA' ERROR 1, X WITHIN 1.E-6 OF A NEGATIVE INTEGER OR ZERO, RETURNED TO EXECUTION, X=XXXXXX.XXXXXXXX. ** (TAMMA)

The user is testing for a gamma distribution. To perform the integrations necessary for solving the gamma density function, numerical approximations are used by the GOF program. Different techniques are used depending on the value of the gamma shape parameter. Execution continues after the user is informed of the evaluation of the shape parameter being used for his run.

20. ** 'GAMMA' ERROR 2, X GT 34.82, RETURNED TO EXECUTION, X=XXXXXX.XXXXXXXX. ** (TAMMA)

Error 19 gives an explanation of the reason for this message to the user.

21. ** 'GAMMA' ERROR 3, X LT - 28.5, RETURNED TO EXECUTION, X=XXXXXX.XXXXXXXX. ** (TAMMA)

Error 19 gives an explanation of the reason for this message to the user.

22. ** GROUPED INPUT DATA NOT ACCEPTABLE WHEN TESTING THE LOG-NORMAL DISTRIBUTION - NO DEFAULT. ** (ERROR,LANG)

The GOF program transforms all hypothesized log-normal data before any goodness of fit tests are performed. All tests are then calculated using the transformed normal data. The GOF program's treatment of log-normal data is discussed in Section IV of this report. The necessary transformations are not appropriate for grouped data. The log-normal distribution is the only distribution restricted in this manner by the GOF program. If the user has ungrouped version of the unacceptable record.

26. ** LANGUAGE INTERPRETER HAS RUN OUT OF DATA BEFORE ENCOUNTERING A START TEST INSTRUCTION. ** (ERROR,SCAN)

The command stream being compiled lacks a START command. The command stream must be reentered with a START statement included.

27. ** LOGNORMAL DISTRIBUTION IS NOT DEFINED AT X=0. HOWEVER, TEST CONTINUES WITH LN(0)=-10.0. ** (LTRSFM)

A zero value is part of the data set hypothesized to be from a log-normal distribution. Because the GOF program transforms all log-normal sample observations before performing goodness of fit tests, it handles this problem by equating the value to -10.0. The program continues to run and performs the requested tests against the transformed normal data.

28. ** NO TERMINAL VALUE SPECIFIED FOR COUNTING DATA POINTS -- DEFAULT IS 999. ** (ERROR,INPUT)

The user is running under the COUNT command, which is explained in Section IV. He is not furnishing the GOF program a terminal value by which to detect the end of the input stream. Therefore, the GOF program defaults to an expected terminal value of 999.

29. ** NOT ENOUGH DATA FOR CELL SPECIFICATION OF NNNNN CHI SQUARE TEST WILL BE RUN WITH NNNNN CELLS. ** (GFMAIN)

Regrouping of the empirical data is necessary to satisfy the requirement of the chi-square goodness of fit test that each cell contain at least five expected observations. This message and a restructured histogram reflecting the chi-square grouping are provided to the user for his information; not for corrective action.

30. ** NUMBER OF CELLS EXCEEDS 15 MAKING HISTOGRAM TOO WIDE TO PRINT ON-LINE. ON-LINE WIDTH RESTRICTED TO 64 PRINT POSITIONS. ** (ERROR,HIST)

The width of the print field is limited to 64 characters for on-line output. Fifteen is the maximum number of cells that can be represented by a histogram in 64 print positions. No histogram is printed. If the chi-square test results in a reduction of the number of cells to 15 or less, a histogram is printed following the chi-square test calculations.

31. ** NUMBER OF CELLS EXCEEDS 32 MAKING HISTOGRAM TOO WIDE TO PRINT IN BATCH MODE. BATCH WIDTH RESTRICTED TO 132 PRINT POSITIONS. ** (ERROR,HIST)

The width of the print field is limited to 132 characters for batch output. Thirty-two is the maximum number of cells that can be represented by a histogram in 132 print positions. No histogram is printed. If the chi-square test results in a reduction of the number of cells to 32 or less, a histogram is printed following the chi-square test calculations.

32. ** NUMERIC FIELD EXCEEDS 10 SIGNIFICANT DIGITS FIELD STARTS WITH AAA. ** (ERROR,CONVRT)

The user is attempting to input a numeric value whose significant digits exceed the number acceptable to the GOF program. The value must be trimmed.

33. ** NUMERIC FIELD EXCEEDS 12 CHARACTERS. NUMBER STARTS WITH A. ** (ERROR,LANG)

The GOF language translator accepts blanks anywhere in its defined statements. When entering numeric parameters, the user should limit the number of numeric values and blanks to twelve characters or less. No value is converted and the user should reenter the unacceptable value in less characters.

34. ** PARAMETERS OF TEST DISTRIBUTION NOT SPECIFIED. DEFAULTS TO ESTIMATES FROM SAMPLE DATA. ** (ERROR,LANG)

This message is for information only and occurs any time the parameters of the hypothesized distribution are not known.

35. ** SEMICOLON OCCURS BEFORE COLON -- INSTRUCTION DELETED. ** (ERROR,LANG)

The acceptable GOF language statement syntax is described in Section IV of this report. The statement immediately preceding this error message violates that syntax and is ignored. The user may continue and correctly reenter the invalid statement. If operating in the batch mode, the GOF program attempts execution without the invalid command.

36. ** TEST TYPE TO BE RUN WAS NOT SPECIFIED. NO DEFAULT - PLEASE SPECIFY NOW. ** (ERROR,LANG)

If the on-line user forgets to specify which goodness of fit tests he wishes to run, he received a second chance from the GOF program. The batch user's job terminates in error. Once the on-line user selects (via the RUN command) which tests are to be run, he must issue another START command to continue processing.

37. ** THE CRAMER-VON MISES TEST CANNOT BE RUN ON GROUPED DATA - REQUEST DENIED. ** (ERROR,LANG)

Section II of this report explains that the Cramer-Von Mises goodness of fit test cannot be run on grouped data. The run continues for any other tests selected.

38. ** THE LOGNORMAL TEST IS PERFORMED ON TRANSFORMED DATA OF THE FORM Y=LN(X). ** (LTRSFM)

This message is for information only. A discussion of the log-normal distribution is part of Section III of this report and Section IV contains an explanation of the GOF program's treatment of hypothesized log-normal data.

39. ** THE MOMENTS TEST CANNOT BE RUN ON GROUPED DATA - REQUEST DENIED. ** (ERROR,LANG)

The GOF program does not compute the coefficients of skewness and kurtosis for grouped data. Therefore, there are no values available for the moments test. The user may run the moments test on the individual sample observations if they are available.

40. ** THE MOMENTS TEST WAS REQUESTED FOR MORE THAN 125 DATA VALUES. ** (ERROR,GFMAIN)

The preceding message is always accompanied by two additional messages which inform the user of techniques suitable for evaluating the results of the moments test. For more than 125 values, the standard normal distribution tables are used. Section II of this work elaborates on the techniques needed to perform this evaluation.

41. ** THE NUMBER OF CELLS HAD TO BE ADJUSTED TO BE ADJUSTED TO MEET THE REQUIREMENTS OF THE CHI-SQUARE TEST. THE HISTOGRAM BELOW SHOWS HOW THE EMPIRICAL DATA WAS GROUPED TO MEET THE CHI-SQUARE TEST REQUIREMENTS. ** (GFMAIN)

Section II of this report describes the chi-square test and its requirement that each cell contain at least five expected observations.

42. ** THE NUMBER OF VALUES INPUT FOR GROUPED DATA WAS NOT THREE TIMES THE NUMBER OF GROUPS - NO DEFAULT. ** (ERROR,GFMAIN)

For grouped data, three values are required to describe each cell: (1) its lower class boundary, (2) its upper class boundary, and (3) its absolute class frequency. Therefore, the GOF program expects to read a number of values which is a multiple of three. The number of cells for the data being input is determined by dividing the number of values to be read by three. The user must reenter his commands with a "DATA" value which is a multiple of three. If the user is entering data under the COUNT option, the GOF program must "count" a multiple of three data values.

43. ** SELECTION OF DISTRIBUTION FOR NULL HYPOTHESIS MUST BE MADE NOW. ** (GFMAIN)

The user is delaying the selection of a hypothesized distribution until the histogram of his empirical data is printed. Immediately after printing the histogram, the GOF program prints the above prompting message and returns to the language translator. At this time the user should enter a TEST command with the distribution specified followed by another START command.

44. ** THEORETICAL DISTRIBUTION TO BE TESTED WAS NOT SPECIFIED. NO DEFAULT - PLEASE SPECIFY NOW. ** (ERROR,LANG)

The user is requesting that no histogram be printed. Therefore, the GOF program must be told which theoretical probability distribution is to be tested. The GOF program waits until the user selects a distribution via a TEST command and follows that TEST command with another START command.

45. ** UNABLE TO EVALUATE RESULTS OF MOMENTS TEST AT 0.01 LEVEL OF SIGNIFICANCE - DEFAULTS TO 0.05 LEVEL FOR THIS TEST ONLY. ** (ERROR,GFMAIN)

The GOF program only contains the critical values for the 0.05 level of significance for the moments test. The 0.01 values are not available from the referenced literature.

46. ** UNKNOWN DISTRIBUTION TO BE TESTED STARTING WITH AAA. ** (ERROR,LANG)

A spelling error probably exists in the TEST command just entered. The GOF program accepts the following shorthand distribution names: (1) POI, (2) GAM, (3) ERL, (4) CHI, (5) NOR, (6) LOG, (7) WEI, (8) EXP, (9) UNI, and (10) TRI. The TEST command should be reentered. This type of error is fatal in the batch mode of operation.

47. ** UNKNOWN INSTRUCTION TYPE STARTING WITH AAA. ** (ERROR,LANG)

The user is attempting to enter an instruction to the GOF language translator which is not in its repertoire. Section IV of this report contains a complete list of all acceptable GOF commands.

48. ** UNKNOWN TEST TYPE STARTING WITH AAA. ** (ERROR,LANG)

The user is asking the GOF program to run a goodness of fit test other than:

1. KAS (Kolmogorov-Smirnov test),
2. CHI (chi-square test),
3. CVM (Cramer-Von Mises test), or
4. MOM (moments test).

The corrected RUN statement should be reentered. This error is fatal in the batch version of the GOF program.

X. Appendix D - GOF Program Components

The GOF program is composed of a main program and 41 subprograms. It requires approximately 32,000 storage locations. It does not require the use of disk or tape storage during its operation. However, the user may select (by the ALTERNATE DATA DEVICE command) to have his data values read from disk, tape, or any other storage device acceptable to his operating environment.

Functionally, the GOF program can be divided into two categories--the language translator and the computational section. The language translator requires about 4000 storage locations. Upon execution of the GOF program, the main routine calls the language translator. The language translator retains control until all testing and operational commands are input and interpreted. Once the START command is interpreted, control is passed to the driver routine of the computational section. Control remains in this section until all testing is completed or until additional information is required from the user. The only additional information which might be required is the selection of the hypothesized distribution. A brief description of each subprogram is included in this section. The subprograms are documented in alphabetical order.

Program Component Descriptions

Language Translator Subprograms

Function CONVRT. The conversion of alphanumeric character codes into real numbers is handled by this subprogram.

Subroutine ERROR. All error messages required by the language translator section are output by this routine. Errors which are fatal to the GOF program execution are flagged and this subroutine terminates execution upon detection of a fatal flag.

Subroutine GC. GC is an acronym for "Get Character" which describes the only function of this subroutine. It returns non-blank characters to its calling program.

Function KOS. This service function searches a language command string until it either identifies a colon or a semicolon and returns the appropriate identification code.

Subroutine LANG. This subroutine is the main driver of the language translator. It interprets all GOF language statements and sets the program parameters necessary for controlling the operational flow of the computational section.

Function LOOKUP. LOOKUP searches its table of acceptable three-character command and keyword abbreviations and returns an integer code if the input characters are identified. It flags invalid three-character sequences.

Subroutine SCAN. This short subroutine reads and immediately prints each command line the user inputs.

Computational Subprograms

Subroutine CELL. If ungrouped sample observations are input by the GOF program user, this subroutine groups the data into the number of cells specified by the user or the program default value of 15. It computes the absolute observed frequency for each cell.

Subroutine CHICHI. The calculations for the chi-square goodness of fit test are performed by this subprogram.

Subroutine CHIPAS. Once the chi-square statistic is calculated, this subroutine checks the critical value tables and notifies the user of the acceptance or rejection of his null hypothesis.

Subroutine CVMPAS. The computed Cramer-Von Mises test statistic is compared to the appropriate critical value and the results of the test are printed by this subroutine.

Subroutine CVMTST. The Cramer-Von Mises test statistic is compute by CVMTST. If the GOF program is operating in batch mode, this subrout prints the intermediate calculation values.

Subroutine DUMPIT. Figure 27 contains an example of the output produced by this subroutine. DUMPIT is executed automatically for batch operations and on command for on-line executions of the GOF program.

Subroutine ENDRN. If the last class (or cell) of any run has less than five observations, and the chi-square test is to be applied, this subroutine consolidates the last class with the previous class and adjusts the necessary class boundaries.

Function ERF. The integration of the normal distribution function is approximated by the calculations in this subprogram.

Subroutine EST. If the parameters of the Weibull distribution are not known, this subroutine estimates them by a technique described in reference (25).

Subroutine EXPON. The theoretical frequency and cumulative theoretical frequency for each cell is calculated by EXPON for the exponential distribution.

Subroutine FIX. If the chi-square test is requested and each cell does not contain at least five expected observations, this subroutine assists in grouping adjacent cells until this criterion is met.

Subroutine FREE. This subroutine implements the free field format option for input values. It accepts numeric characters, blanks, and the characters "E", "+", "-", ".", and "," necessary for integer, floating, or scientific notation.

Subroutine GAMFAM. This subprogram calculates theoretical frequencies and theoretical cumulative frequencies for the gamma, Erlang-k, and chi-square distribution. It calls two related subprograms to assist in the calculations: GAMMA and TAMMA.

Function GAMMA. The GAMMA function is the driver routine for the computation of the gamma distribution integral. This function relies on the following four functions to calculate or approximate the gamma integral under various data conditions: GSER, GCHEB, GFRAC, and GAMNEG.

Subroutine GFMAIN. This subroutine is the driver program for the entire computational section. It is responsible for executing the procedures defined by the user through the GOF language commands. Most of the work of the GOF computational section is performed by the individual subprograms documented in this section. GFMAIN does contain the code which constructs and prints the appropriate hypothesis statement.

Subroutine HIST. The histogram of observed frequencies is printed by this subroutine.

Subroutine INPUT. All data values, whether grouped or ungrouped, are read by INPUT. It handles all format and counting options available to the user through the GOF language commands.

Subroutine KSPAS. Once the Kolmogorov-Smirnov test statistic is calculated, the KSPAS subroutine checks the computed value against the appropriate critical values and notifies the user of the results of the test.

Subroutine KSTEST. The calculations for the Kolmogorov-Smirnov goodness of fit test form the body of this subroutine. It prints intermediate computational values in the batch operating mode.

Subroutine LTRSFM. Any sample observations believed to have come from a log-normal distribution are transformed to their normal counterparts before any goodness of fit tests are applied. This subroutine handles the transformation of the sample observations.

Subroutine MOMENT. Biased or unbiased coefficients of skewness and kurtosis are calculated by this subprogram.

Subroutine MOMTES. If the user has requested a moments test for normality, the coefficients of skewness and kurtosis are compared to the correct critical values by this subroutine. It also prints the results of the moments test for the user.

Subroutine NORMAL. The theoretical expected frequencies and theoretical cumulative frequencies are calculated by NORMAL.

Subroutine POISON. POISON is responsible for the calculation of the theoretical frequencies and theoretical cumulative frequencies for the Poisson distribution.

Subroutine SAMPLE. For grouped and ungrouped data, SAMPLE computes the mean, variance, standard deviation, and coefficient of variation from the sample observations. For ungrouped data, it calls the MOMENT subroutine to compute either the biased or unbiased coefficients of skewness and kurtosis.

Function TAMMA. This short function is part of the gamma family of subroutines and functions necessary to calculate the expected theoretical frequencies and cumulative theoretical frequencies for the gamma distribution.

Subroutine TRESRT. Ungrouped sample observations are arranged into numerically ascending order by this tree sort routine.

Subroutine TRIANG. The TRIANG subprogram computes the expected theoretical frequencies and the cumulative theoretical frequencies for the triangular distribution.

Subroutine UNIFRM. The expected theoretical and cumulative theoretical frequencies are computed for the uniform distribution by this subroutine.

Subroutine WEIBUL. WEIBUL computes the expected theoretical and cumulative theoretical frequencies for the Weibull distribution.

Table 6 gives a quick reference of the various subprogram components of the GOF program. It divides them into those programs necessary for the language translator and those necessary for the computational portion. The only routine common to both areas is the ERROR subroutine. The numbers is parentheses following some subprograms are the numbers of the subprograms which they call. The main program of the GOF package is called GOF and calls only the LANG subroutine.

Table 6
GOF Subprogram Components

Language Translator Subprograms	Computational Subprograms	
1. CONVRT(2)	1. CELL(12)	19. GFRAC
2. ERROR	2. CHICHI(3)	20. GSER
3. GC(7)	3. CHIPAS	21. HIST
4. KOS(3)	4. CVMPAS	22. INPUT(9,13)
5. LANG(1-4,6,7)	5. CVMTST(4)	23. KSPAS
6. LOOKUP	6. DUMPIT	24. KSTEST(23)
7. SCAN(2)	7. ENDRN	25. LTRSFM
	8. ERF	26. MOMENT
	9. ERROR	27. MOMTES
	10. EST	28. NORMAL(8)
	11. EXPON	29. POISON
	12. FIX	30. SAMPLE(26)
	13. FREE	31. TAMMA
	14. GAMFAM(15,31)	32. TRESRT
	15. GAMMA(16,17,19,20)	33. TRIANG
	16. GAMNEG(20)	34. UNIFRM
	17. GCHEB	35. WEIBUL
	18. GFMAIN(1,2,5-7,10-12,14,21,22,24,25,27-30,32-35)	

XI. References

1. Breiman, Leo. Statistics: With a View Toward Applications. Boston: Houghton Mifflin Company, 1973.

2. Caulcott, Evelyn. Significance Tests. Boston: Routledge & Kegan Paul Ltd., 1973.

3. Clark, Charles T., and Lawrence L. Schkade. Statistical Methods for Business Decisions. Cincinnati, Ohio: South-Western Publishing Company, 1969.

4. Cochran, William G. "The χ^2 Test of Goodness of Fit." Annuals of Mathematical Statistics, 23:315-345, 1952.

5. Cox, D. R., and P. A. W. Lewis. The Statistical Analysis of a Series of Events. London: Methuen and Co., Ltd., 1966.

6. Darling, D. A. "The Kolmogorov-Smirnov, Cramer-Von Mises Tests." Annuals of Mathematical Statistics, 28:823-838, 1957.

7. Hahn, Gerald J., and Samuel S. Shapiro. Statistical Models in Engineering. New York: John Wiley and Sons, Inc., 1967.

8. Harnett, Donald L. Introduction to Statistical Methods. Reading, Massachusetts: Addison-Wesley Publishing Company, 1970.

9. Hastings, N. A. J., and J. B. Peacock. Statistical Distributions. New York: John Wiley and Sons, Inc., 1974.

10. Klugh, Henry E. Statistics: The Essentials for Research. New York: John Wiley and Sons, Inc., 1970.

11. Martin, Francis F. Computer Modeling and Simulation. New York: John Wiley and Sons, Inc., 1968.

12. Massey, Frank J., Jr. "The Kolmogorov-Smirnov Test for Goodness of Fit." Journal of American Statistics Association, Vol. IV, pp. 68-78, 1951.

13. Meyer, Paul L. Introductory Probability and Statistical Applications. 2d ed. Reading, Massachusetts: Addition-Wesley Publishing Company, 1970.

14. Miller, Irwin, and John E. Freund. Probability and Statistics for Engineers. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1977.

15. Ostle, Bernard. Statistics in Research. Ames, Iowa: The Iowa State College Press, 1954.

16. Phillips, Don T. "Applied Goodness of Fit Testing." OR Monograph Series #1, AIIE-OR-72-1, Norcross, Georgia: American Institute of Industrial Engineers, Inc., 1972.

17. Shannon, Robert E. Systems Simulation: The Art and Science. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1975.

18. Snedecor, George W., and William G. Cochran. Statistical Methods. 6th ed. Ames, Iowa: The Iowa State University Press, 1967.

19. Tsokos, Chris P. Probability Distributions: An Introduction to Probability Theory with Applications. Belmont, California: Duxbury Press, 1972.

20. Wadsworth, George P., and Joseph G. Bryan. Applications of Probability and Random Variables. 2d ed. New York: McGraw-Hill Book Company, Inc., 1960.

DISTRIBUTION LIST

NAVOCEANO TECH. NOTE
NO. 5000-1-79

DATE:

2 August 1979

SUBJECT:

A FORTRAN COMPUTER PROGRAM TO PERFORM GOODNESS OF FIT TESTING ON EMPIRICAL DATA

by Sue D. Guthrie, Dated June 1979

CLASSIFICATION:
Unclassified

NUMBER OF COPIES PREPARED:
Sixty (60)

COPY NO.		COPY NO.
1	(Internal)	
2	Code 00	
3	Code 02	
4	Code 3000	
5	Code 3400	
6	Code 3500	
7	Code 3700	
8	Code 5000	
9	Code 5200	
10	Code 6000	
11	Code 8103 (Library)	
12	Code 8600	
13		
14		
15		
16		
17		
18		
19		
20		

REMARKS: