PROBABILITY CALCULATIONS ON A DIGITAL COMPUTER

ROBERT K. OTNES

MEASUREMENT ANALYSIS CORPORATION

FOREWORD

This report was prepared by the Measurement Analysis Corporation, Los Angeles, California, for the Aero-Acoustics Branch, Vehicle Dynamics Division, AF Flight Dynamics Laboratory, Wright-Patterson Air Force Base, Ohio, under Contract AF33(615)-1314. This research is part of a continuing effort to obtain significant information on sound environment simulation and dynamic response to acoustic excitation under the Research and Technology Division, Air Force Systems Command's exploratory development program. The Project No. is 4437, "High Intensity Sound Environment Simulation," and Task No. 443706, "Advanced Instrumentation Study for Sonic Fatigue Experimental Work," Mr. W. K. Shilling, III was the project engineer.

The contractor's report number is MAC 402-07. The manuscript was released by the authors April 1965 for publication as an Air Force Flight Dynamics Laboratory Technical Report.

This report covers work conducted from February 1964 to April 1965.

This technical report has been reviewed and is approved.

Walter J Mykytow

Walter J. Mykytow
Asst. for Research & Technology
Vehicle Dynamics Division
AF Flight Dynamics Laboratory

ABSTRACT

The purpose of this report is to discuss computer implementations of tests for normality. Some of the underlying statistical considerations and random process theory are briefly reviewed to provide a background for subsequent material. Methods for digitally computing basic statistical parameters, such as the mean, variance, skewness, kurtosis, etc., are given, along with procedures for computing the sample probability density function. Problems arising from the digital and discrete nature of the data are discussed, as well as sample sizes required for the tests. The central part of the report consists of a detailed discussion of the computer implementation of the chi-square goodness-of-fit test as applied to testing for normality, followed by computer program flow charts, suitable for use in coding the procedure for a digital computer. Sufficient detail has been provided so that a programmer unfamiliar with the material should be able to write a program which will make maximum use of the time which is frequently lost during periods of input and output of data from the computer. The report concludes with an appendix which lists standard approximations for some of the mathematical functions required in the calculations.

CONTENTS

PROBABILITY CALCULATIONS ON A DIGITAL COMPUTER

1. INTRODUCTION

The purpose of this document is to

- Review certain statistical parameters needed for the χ^2 goodness-of-fit test to test the probability density function of vibration and acoustic data for normality.
- Provide detailed methods for calculating these parameters and the χ^2 test statistic to give techniques suitable for implementation on a digital computer.

While some of this material is available in standard statistical treatises, the emphasis in those works is usually on small samples. Furthermore, the instructions for making the necessary computations usually are suitable only for hand or desk calculations. When large quantities of data are to be processed, or the procedures must be automated, the method of implementing the tests must be adapted accordingly.

Properly interpreted, the material contained herein may be employed as a guide for the design of a computer program for probability calculations, and as a reference in the use of the program after its completion.

2. REVIEW OF CERTAIN STATISTICAL CONCEPTS

The material given in this section is intended only as a review of the definition of some basic statistical parameters, and is included so that the reader may make use of material in subsequent sections without having to refer to other documents. For a deeper understanding of the subject, Reference 1 is recommended.

2.1 RANDOM PROCESSES

The precise definition of a random process requires an elaborate mathematical scaffolding due to the generality of the variables which may be examined. Fortunately, in common engineering practice, the pathological functions considered by mathematicians do not exist. Hence, several simplifying assumptions can be made without impairing the validity of the results. These are

1. that the functions being considered are continuous, and
2. that the random process underlying the function is ergodic and stationary.

The remainder of this subsection will be spent in an intuitive discussion of the meaning of assumption 2.

Suppose that a large number of identical random noise generators were turned on at some remote time in the past, and left to run. Associated with the output of all of the generators is a function $p(x, t)$ with the following characteristics. For a certain fixed time, say t_0, the probability that the output of the ith signal generator, $x_i(t_0)$, lies between values a and b is given by the integral of the probability density function taken between the limits a and b.

$$P\left[a \leq x_i(t_0) < b\right] = \int_a^b p(x, t_0)\, dx \qquad (1)$$

Note that the integration is performed with respect to the range of the random variable. Statistical moments are defined as follows.

$$\mu(t_0) = \int_{-\infty}^{\infty} x(t_0)\, p(x, t_0)\, dx \tag{2}$$

$$\sigma^2(t_0) = \int_{-\infty}^{\infty} \left[x(t_0) - \mu(t_0)\right]^2 p(x, t_0)\, dx \tag{3}$$

where $\mu(t_0)$ is the expected value (mean) and $\sigma^2(t_0)$ is the second moment about the mean (variance) of the random process at time t_0. If the random process is stationary, the parameters $\mu(t_0)$ and $\sigma^2(t_0)$ are independent of time. That is,

$$\mu(t_0) = \mu(t_1) \equiv \mu \tag{4}$$

and

$$\sigma^2(t_0) = \sigma^2(t_1) \equiv \sigma^2 \tag{5}$$

where t_0 and t_1 are arbitrary. As stationarity is assumed, the mean and variance hereafter will be written without the qualifying t_0.

The assumption of ergodicity permits ensemble averages to be replaced with time averages. In the noise generator example, if the generators are exactly alike, even though they are each producing different random variables, the output of one of the generators is sufficient to define the statistics for all of them. The expression for the mean, Eq. (2), may be replaced with

$$\mu = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t)\, dt \tag{6}$$

which is a time average based on a single record of the process.

2.2 SAMPLING

In the previous section, the functions discussed were defined on the interval $t = -\infty$ to $t = \infty$, and were not constrained except for continuity, etc. Two major constraints are placed on data functions by the method of testing during the course of a test:

1. only a finite time span of data is recorded, and
2. bandwidth of the data system is finite.

The second constraint will now be discussed in terms of "bit rate" for digital data systems.

If the system in question can record or transmit n binary bits per second per function, the bits could have been distributed to give varying sample rates and accuracies as follows.

$$S = \text{sample rate} = \frac{\text{bit rate}}{\text{bits/word}} = \frac{n}{q} \tag{7}$$

where, as usual, the sample rate is the number of samples per unit time taken of a function, and the data "word" is the digitized sample expressed as a binary number made up of q binary digits. Usually the number of bits per word, q, has been fixed by equipment design considerations during an early stage of development of the data system and unalterable thereafter. The bit rate and sample rate may be varied in the same manner in most of the common systems, but there is a definite upper limit for any individual system. The variation, of course, is quite large. If the digital data is to be recorded, this places yet another limitation, usually lower than that imposed by the rate of the analog to digital converters.

2.3 SAMPLED STATISTICAL PARAMETERS

In Section 2.1, expressions for the mean and variance of a probability density function were given. These and other parameters are frequently estimated for a sequence of sample data, in which case it is customary to use the adjective "sample" with the resulting calculations, and to modify the notation accordingly. For example, if the sequence $\{x_i\}$, $i = 1, 2, \ldots, N$, is a series of readings recorded during a test, then the sample mean, m, is computed by

$$m = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{8}$$

and the unbiased sample variance, s^2, is obtained from

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - m)^2 \tag{9}$$

The true values are μ and σ^2 defined by Eqs. (4) and (5).

Sample probability density functions may also be obtained from such a data sequence. Such sample functions are not unique for a given sample of data as are m and s^2, but depend upon the values of certain coefficients used in their derivation. The details of these computations will be postponed until Section 3.

2.4 CHI-SQUARE GOODNESS-OF-FIT TEST

A test which is often used to check the equivalence of a probability density for sampled data to some theoretical density function is called the chi-square goodness-of-fit test. The general procedure involves the use of the chi-square statistic as a measure of the discrepancy between an observed probability density function and the theoretical density function. A hypothesis of equivalence is then tested by studying the sampling distribution of chi-squared.

As before, consider a sample of N independent observations from the random variable x with the probability density function p(x). Let the observations be grouped into k equally wide intervals, called class intervals, which together form a frequency histogram. The number of observations falling within the ith class interval is called the observed frequency in the ith class, and will be denoted by N_i. The number of observations which would be expected to fall within the ith class interval if the true probability density function for x were $p_0(x)$ is called the expected frequency in the ith class interval, and will be denoted by F_i. Now the discrepancy between the observed frequency and expected frequency within each class interval is given by $(N_i - F_i)$. To measure the total discrepancy in each interval must be used since

$$\sum_{i=1}^{k} N_i = \sum_{i=1}^{k} F_i = N \tag{10}$$

which means that the sum of the discrepancies must equal zero. Using the sum of the squared discrepancies, the sample chi-square is obtained as follows.

$$X^2 = \sum_{i=1}^{k} \frac{(N_i - F_i)^2}{F_i} \tag{11}$$

Under suitable assumptions, this sample chi-square may be compared with the regular chi-square distribution, denoted by χ^2. [An actual numerical value will be written as $\chi^2_{n;\alpha}$.]

The distribution for χ^2, which is discussed in many references, depends upon the number of independent squared variables in χ^2 (the number of degrees-of-freedom n). The value of n is equal to k minus the number of

different independent linear restrictions imposed on the observations. There is one such restriction due to the fact that the frequency in the last class interval is determined once the frequencies in the first k - 1 class intervals are known. There is at least one additional restriction due to fitting the expected theoretical density function to the frequency histogram for the observed data. For the frequently occurring case where the expected theoretical density function is the normal density function, two restrictions are imposed because a mean and variance must be computed to fit a normal density function. Hence, in the instance where the chi-square goodness-of-fit test is used as a test for normality, the number of degrees-of-freedom for χ^2 is n = k - 3.

Having established the proper degrees-of-freedom n for χ^2, a hypothesis test may be performed as follows. Let it be hypothesized that the variable x has a probability density function $p(x) = p_0(x)$. After grouping the sampled observations into k class intervals and computing the expected frequency for each interval assuming $p(x) = p_0(x)$, compute X^2 as indicated in Eq. (11). Any deviation of p(x) from $p_0(x)$ will cause X^2 to increase. The region of acceptance is

$$X^2 \leq \chi^2_{n;\alpha} \tag{12}$$

If the computed value of X^2 is greater than $\chi^2_{n;\alpha}$, the hypothesis $p(x) = p_0(x)$ is rejected at the α level of significance. If X^2 is less than or equal $\chi^2_{n;\alpha}$, the hypothesis is accepted.

For the case where the chi-square goodness-of-fit test is used as a normality test, with a level of significance $\alpha = 0.05$, Table 1 provides an explicit guideline for selecting the number of class intervals.

TABLE 1

Minimum Optimum Number (k) of Class Intervals for Sample Size N when $\alpha = 0.05$

N	k
200	16
400	20
600	24
800	27
1,000	30
1,500	35
2,000	39
4,000	57
7,000	65
10,000	74
20,000	94
40,000	129
70,000	162
100,000	187
200,000	247
400,000	326
700,000	407
1,000,000	470
1,140,000	500

The above is based on the relation

$$k = 1.87(N - 1)^{2/5}$$

obtained from Reference 3.

This table cannot be used for all data. While satisfactory for digital processes developed from numerical random number generators, it may not be appropriate, except for providing an upper limit, for output data from an analog to digital convertor (ADC). Typical ADC's have a total number of discrete digital levels. This number is usually a power of two, common values being 64, 128, 256, 512, 1024, and 2048. Clearly, it is senseless to set up one hundred class intervals when there are only 64 digitizer levels, as at least 36 of the class intervals will be empty.

It would be preferable to have the same number of digitizer levels for all class intervals. This may be difficult to do in practice, as the digitizer itself may be biased, and the conversion of engineering data to digital counts may be nonlinear, resulting in a poor distribution of the possible data values to the equally spaced class intervals.

When processing digitized data (i.e., analog data processed through a coarse ADC), there must be a definite upper limit for k dependent on the number of digitizer levels itself. A safer procedure when processing large volumes of data is to set maximum k equal to one-fifth to one-tenth of the number of digitizer levels, so that each case need not be checked to see if the class intervals are biased.

The power of the test is decreased, however, when fewer than the optimum number of class intervals is used, and the chi-square test is more likely to accept the normality hypothesis when it is in fact false.

2.5 SKEWNESS AND KURTOSIS (MOMENTS) TEST OF NORMALITY

An alternative procedure to the chi-square goodness-of-fit test is to compute the normalized sample third and fourth moments about the mean (skewness and kurtosis, respectively), and examine the resulting numbers. As the normal distribution is completely determined by its first and second moments, the higher order moments turn out to be functions of the first two. In particular,

$$\alpha_3^* = \frac{\sum_{i=1}^{N} (x_i - \mu)^3}{\sigma^3 N} \qquad \text{(skewness)} \qquad (13)$$

$$\alpha_4^* = \frac{\sum_{i=1}^{N} (x_i - \mu)^4}{\sigma^4 N} \qquad \text{(kurtosis)} \qquad (14)$$

and $E\left[\alpha_3^*\right] = 0$, $E\left[\alpha_4^*\right] = 3$. Reference 4 lists values for the 1% and 5% significance levels of the sampling distributions for the sample estimates defined above. If both sample values are within the 5% deviation intervals, the sequence being examined may be accepted as normal at the 5% level of significance. This sample size given in the tables is for independent observations. Hence, an effective sample size of $N^* = 2BT$ should be used when entering the tables as opposed to the total number of digitized points. It is not possible to say under what conditions which test, χ^2 or moments, is the better test for normality. It would seem advisable to perform both and investigate the data further should they not agree.

A full explanation of the $N^* = 2BT$ expression is given in Reference 1; briefly, B is the noise bandwidth, T is the time span of the data, and N^* is the number of independent data points.

3. PROBABILITY DENSITY CALCULATIONS AND NORMALITY TEST ON A DIGITAL COMPUTER

The procedure for digitally generating the sample probability density function of a function $x(t)$ and testing it for normality can be arranged in a series of six steps:

i. computation of the sample mean and standard deviation

ii. sorting the data to produce the sample probability density function

iii. taking care of certain end point problems (to be discussed more fully below)

iv. computation of the normal distribution based on a sample mean and standard deviation

v. computing X^2, the sample estimate for χ^2

vi. testing the computed X^2 to see if it satisfies predetermined conditions

Suppose that as before the digitized realization of the random process is the sequence $\{x_i\}$, $i = 1, \ldots, N$. As in Section 2.3, the sample mean is

$$m = \frac{\sum_{i=1}^{N} x_i}{N}$$

and the sample standard deviation is

$$s = \sqrt{\frac{\left(\sum_{i=1}^{N} x_i^2\right) - Nm^2}{N - 1}}$$

the latter expression being a more convenient computational form.

The sample mean and standard deviation for a given sequence are unique. The sample probability density function on the other hand is not. It is determined by a and b (the interval of the range of x to be examined)

and the parameter k (the same k defined in Section 2.4). When a, b, or k are changed, the density function will also change. The reason for this is implicit in the manner in which the density function is computed. The interval $[a, b]$ is divided into k equally spaced subintervals, and the number of occurrences of x in each of them is tabulated. For a given subinterval, the number of occurrences of x within it, divided by N, is the sample estimate of the probability of occurrence of x in the subinterval.

To formalize this process, the histogram of x is obtained in the following manner. Let $\{N_j\}$ be the set of integers obtained by sorting x on the range $[a, b]$. Let $c = (b - a)/k$, and $d_j = a + jc$. Then $\{N_j\}$ is given by:

j	N_j
0	[Number of x such that $x \leq a$]
.	.
.	.
.	.
j	[Number of x such that $d_{j-1} < x \leq d_j$]
.	.
.	.
.	.
k	[Number of x such that $d_{k-1} < x \leq b$]
(k+1)	[Number of x such that $x > b$]

Figure 1 illustrates these quantities.

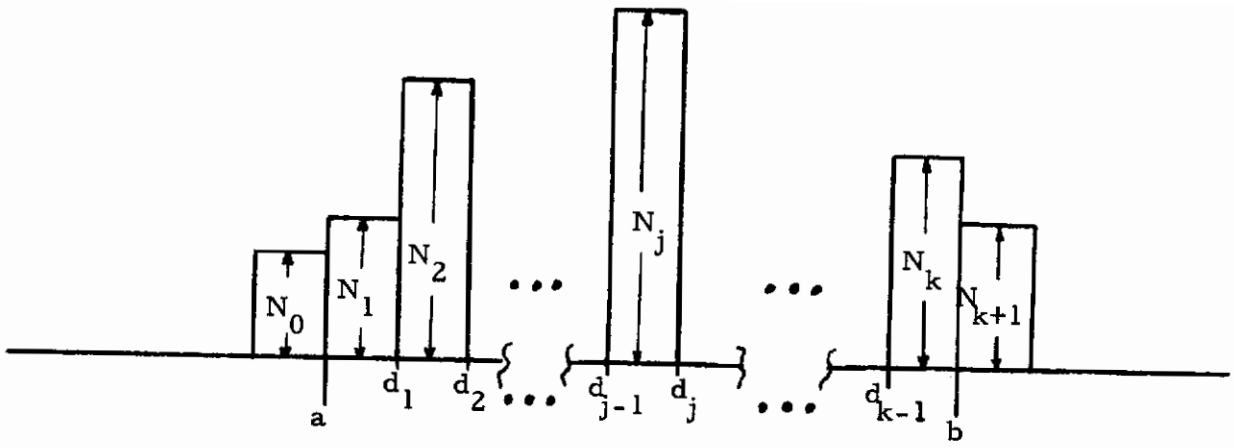

Figure 1. Illustration of Histogram Construction

One method of doing this sorting on a digital computer is to examine each x_i, $i = 1, 2, \ldots, N$, in turn. Then

i. if $x_i \leq a$, add one to N_0.

ii. If $a \leq x_i < b$, compute $L = \left[\frac{x_i - a}{c}\right]$; then select ℓ as the largest integer less than or equal to L and add one to N_ℓ. (This technique is usually easy to program on binary machines having floating point operations.)

iii. If $x_i > b$, add one to N_{k+1}.

Three forms of the sequence are used. The first is the histogram, which is simply the sequence $\{N_j\}$ without changes. The second is the function $\{P_j\}$, $j = 0, \ldots, (k+1)$, where

$$P_j = \text{Probability}\left[d_{j-1} < x \leq d_j\right]$$

$$= N_j/N$$

The third sequence, $\{p_j\}$, $j = 1, \ldots, k$, is the sample density function

$$p_j = \frac{N_j k}{N(b-a)}$$

which can be interpreted as the derivative of the distribution function at the midpoint of the given interval.

A normal distribution is completely determined by the mean and variance. Hence, if x_i is normally distributed, then N_j should be approximately equal to F_j, where

$$F_j = N \int_{(d_{j-1}-m)/s}^{(d_j-m)/s} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \quad ; \quad j = 1, \ldots, k$$

This integral can be evaluated numerically using standard techniques, such as Reference 3. If $\{N_j\}$ and $\{F_j\}$ are "nearly alike," then $\{x_i\}$ will be taken to be normally distributed. The chi-square test may be used to test the hypothesis that $\{N_j\}$ and $\{F_j\}$ are "nearly alike."

4. THE CHI-SQUARE TEST ON A DIGITAL COMPUTER

In order to carry out the χ^2 test for goodness-of-fit described below, elements of $\{F_j\}$ must be multiplied by the relative bandwidth B_r, where $B_r = N^*/N$ (see Section 2.5), and those values of $\{B_r F_j\}$ such that $B_r F_j < 5$ pooled with adjacent larger values. While this is easy to do visually, computer mechanizations are somewhat more difficult. One technique which has proved to be satisfactory in most cases is to find P such that $F_P > F_n$ for all j. Then, define the sequences Q_j and R_j by

$$\text{If } B_r F_j \geq 5, \text{ then } \begin{cases} B_r N_j + Q_j \longrightarrow Q_j \\ B_r F_j + R_j \longrightarrow R_j \end{cases}$$

$$\text{If } B_r F_j < 5, \text{ then } \begin{cases} 0 \longrightarrow Q_j \,,\; R_j \\ B_r N_j + Q_{j+1} \longrightarrow Q_{j+1} \\ B_r F_j + R_{j+1} \longrightarrow R_{j+1} \end{cases} \quad \text{for } j < P$$

or

$$\begin{cases} 0 \longrightarrow Q_j \,,\; R_j \\ B_r N_j + Q_{j-1} \longrightarrow Q_{j-1} \\ B_r F_j + R_{j-1} \longrightarrow R_{j-1} \end{cases} \quad \text{for } j > P$$

The sequences generated by this procedure are similar to $\{N_j\}$ and $\{F_j\}$, except that the numbers have been reduced by the bandwidth factor and the "tails" of distributions have been shifted towards the center.

Next, define the sequence $\{H_j\}$ by

$$H_j = \begin{cases} 1 & \text{if } Q_j \neq 0 \\ 0 & \text{if } Q_j = 0 \end{cases}, \quad j = 1, \ldots, k$$

X^2 and n, the number of degrees-of-freedom, are now defined by

$$X^2 = \sum_{j=1}^{k} \frac{(Q_j - R_j)^2}{R_j} H_j$$

and

$$n = \left(\sum_{j=1}^{k} H_j \right) - 3$$

The usual procedure in applying the χ^2 test is to preselect a parameter α, the level of significance of the test (frequently chosen values for α are .10, .05, and .01 with $\alpha = 0.05$ being the most common). $\chi^2_{n;\alpha}$ is obtained through one of several procedures. It is defined by the relationship

$$\alpha = \text{Prob}\left[X^2 > \chi^2_{n;\alpha} \right]$$

If $X^2 > \chi^2_{n;\alpha}$, then the hypothesis that x_i is normally distributed is rejected at the α significance level. On the other hand, if $X^2 \leq \chi^2_{n;\alpha}$, $\{x_i\}$ is accepted as being normally distributed. In this case, it is common practice to say that the acceptance of the hypothesis is at the $(1 - \alpha)$ confidence level. The quantity $\chi^2_{n;\alpha}$ is dependent on both α and n. It may be computed implicitly using various computer subroutines presently in existence, or for fixed α it may be stored as a table of numbers, as in Table 2. Common tables for $\chi^2_{n;\alpha}$ usually do not go beyond $n = 30$. If more

than 33 pockets are to be employed, resulting in $n > 30$, an approximation of the $\chi^2_{n;\alpha}$ distribution may be used.

$$\chi^2_{n;\alpha} = \frac{1}{2}\left(\sqrt{2n - 1} + x_\alpha\right)^2$$

where

$$x_\alpha = \left[x : \Phi(x) = 1 - \alpha\right]$$

and

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2} dt$$

A somewhat more complicated problem arises when α is not given, but is to be computed as a function of X^2 and n. This can be done by having a large table of α versus χ^2 and n stored as a part of the computer program, and then using a double interpolation routine to calculate α. If this procedure is too wasteful of computer space, a computer subroutine may be used to find α for a given X^2 and n. Such routines frequently contain a subset of instructions for computing the cumulative normal distribution, so that the computer storage used to contain the required instructions need not be too extensive.

TABLE 2

$\chi^2_{n;\alpha}$ for $\alpha = 0.05$ and $n = 1, \ldots, 30$

n	$\chi^2_{n;0.05}$	n	$\chi^2_{n;0.05}$
1	3.841	16	26.296
2	5.991	17	27.587
3	7.815	18	28.869
4	9.489	19	30.143
5	11.071	20	31.410
6	12.592	21	32.671
7	14.067	22	33.924
8	15.507	23	35.173
9	16.919	24	36.415
10	18.307	25	37.653
11	19.675	26	38.885
12	21.026	27	40.113
13	22.362	28	41.337
14	23.685	29	42.557
15	24.996	30	43.773

5. COMPUTER PROGRAM FLOW CHARTS

The computer program flow charts, Figures 2, 3, 4, and 5, show one standard scheme for program set-up. They include as a convenience the steps required to process some additional parameters not mentioned in the above discussion, such as the minimum and maximum values of x.

The arithmetical expressions on the charts, e.g., $x_i^2 + x^2 \longrightarrow x^2$, refer both to data values and storage locations. The above example would read: "the current value of x_i is squared and added to the contents of the storage location of the running total of x^2, and the results stored into that location."

The extension of the procedure to multifunction parallel processing is fairly obvious, so it is omitted from the charts in order to clarify them. The over-all flow of data is shown in Figure 2. The form of program arrangement in Figure 2 allows the user to recover some of the time frequently lost because of the slowness of input devices. Double buffering routines can be employed with the processing operating independently of the routines performing the input operations, thus allowing the input routine coding to be a sort of universal building block, usable with other applications or even separate parts of the same basic data processing program.

As a final comment on the flow charts, the boxes with printout or data input functions will probably require the largest part of the programming effort, as these items can mushroom into large tasks if speed of processing and clarity of output are requirements.

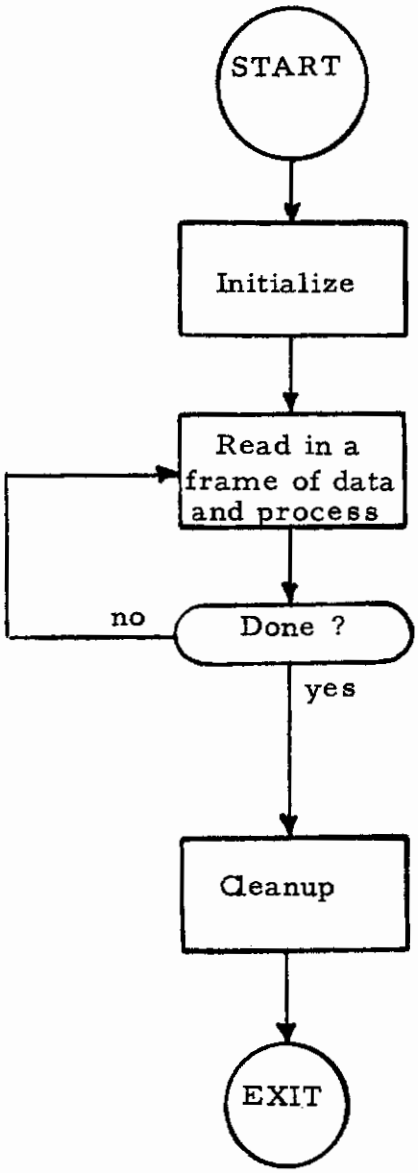

Figure 2 Over-all Flow Chart

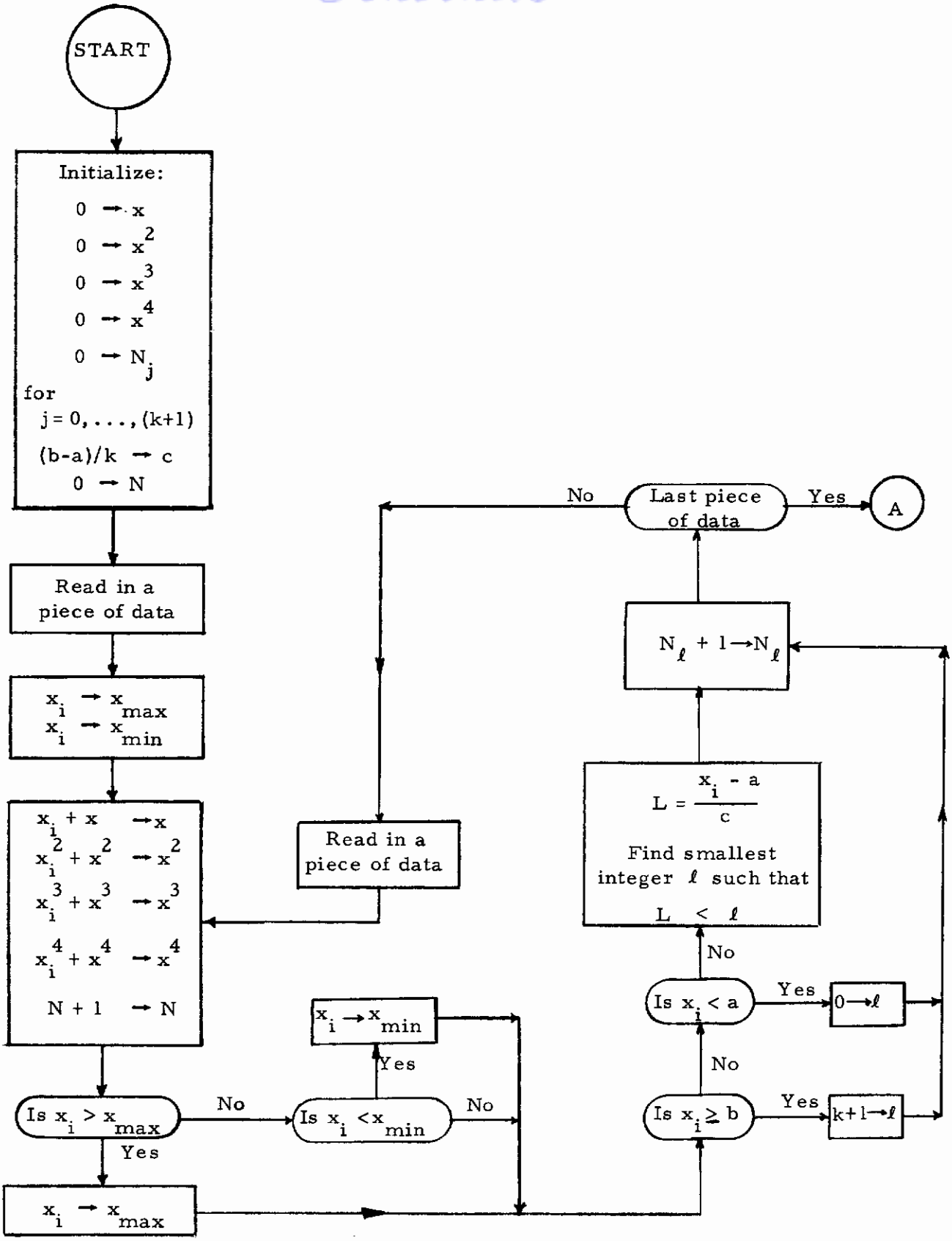

Figure 3. Basic Processing

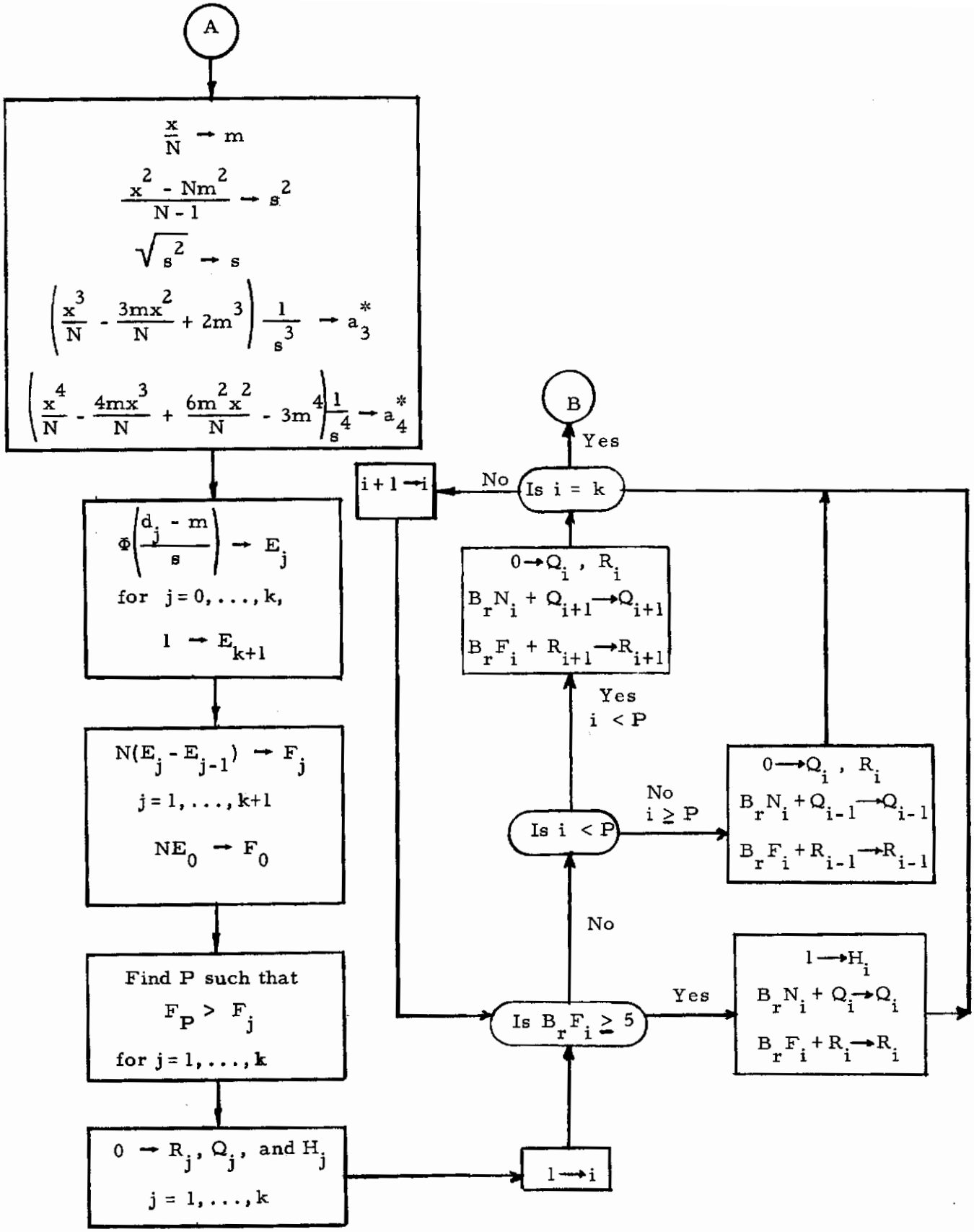

Figure 4 Final Processing

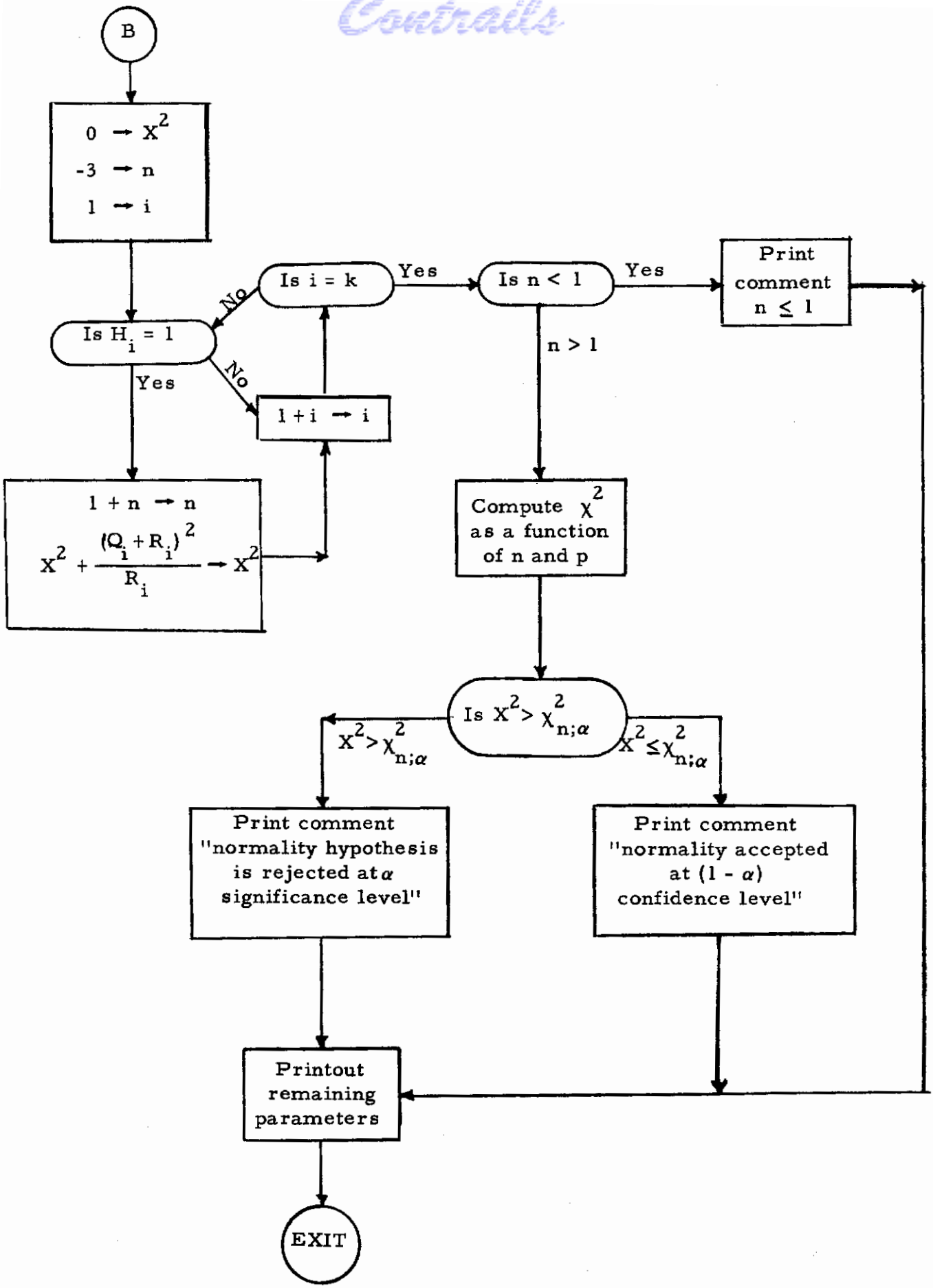

Figure 5 Final Processing 2

REFERENCES

1. Bendat, J.S., Enochson, L.D., Klein, G.H., and A.G. Piersol, "The Application of Statistics to the Flight Vehicle Vibration Problem," ASD TR 61-123, Aeronautical Systems Division, AFSC, USAF, Wright-Patterson AFB, Ohio. December 1961. (AD 271 913).

2. Hastings, C., Jr., Approximations for Digital Computers, Princeton University Press, Princeton, New Jersey. 1955.

3. Kendall, M.G., and A. Stuart, The Advanced Theory of Statistics, vol. 2, Hafner Publishing Company, New York. 1961.

4. Pearson, E.S. and H.O. Hartley, Biometrika Tables for Statisticians, vol. I, 2d Ed., Cambridge University Press, Cambridge, England, 1962.

APPENDIX

NUMERICAL SUBROUTINES

It is possible that some of the computer subroutines discussed in the main part of the write-up may not be available for a particular machine. The purpose of this Appendix is to provide some well-known numerical approximations for reference in case the routines must be coded.

The first of these is the common expression for $\Phi(x)$ as given in Reference 2. Let

$$f(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

$$\approx 1 - \frac{1}{\left[\sum_{i=0}^{4} a_i x^i\right]^4}$$

where

$$a_0 = 1$$

$$a_1 = 0.278393$$

$$a_2 = 0.230389$$

$$a_3 = 0.000972$$

$$a_4 = 0.078108$$

Define Φ by

$$\Phi(x) = \begin{cases} \frac{1}{2} + \frac{1}{2} f\left(\frac{x}{\sqrt{2}}\right) & x \geq 0 \\ \frac{1}{2} - \frac{1}{2} f\left(-\frac{x}{\sqrt{2}}\right) & x < 0 \end{cases}$$

Then

$$\frac{1}{\sqrt{2\pi}\ \sigma} \int_a^b e^{\frac{-(x-\mu)^2}{2\sigma^2}}\, dx \approx \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

The second set of relationships is two formulas which may be used in χ^2 computations. They yield α as a function of χ^2 and n, and are exact.

$$\alpha = \text{Prob}\left[X^2 > \chi^2_{n;\alpha}\right]$$

$$= 2\Phi(X) - 1 - 2\phi(X)\left[\sum_{r=1}^{(n-1)/2} \frac{X^{2r-1}}{1 \cdot 3 \cdot 5 \ldots (2r-1)}\right] \quad \text{for n odd}$$

$$= 1 - \sqrt{2\pi}\,\phi(X)\left[1 + \sum_{r=1}^{(n-2)/2} \frac{X^{2r}}{2 \cdot 4 \cdot 6 \ldots (2r)}\right] \quad \text{for n even}$$

where as usual

$$\phi(X) = \frac{1}{\sqrt{2\pi}}\, e^{-X^2/2}$$

For $n > 30$, the approximation given in Section 4 may be substituted for the above to save computer time.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R&D
(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Measurement Analysis Corporation 10962 Santa Monica Blvd. Los Angeles, California 90025	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

Probability Calculations on a Digital Computer

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Final

5. AUTHOR(S) (Last name, first name, initial)

Otnes, Robert K.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
August 1965	26	4

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
AF 33(615)-1314 b. PROJECT NO. 4437 c. Task No. 443706 d.	AFFDL-TR-65-75 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) MAC 402-07

10. AVAILABILITY/LIMITATION NOTICES

Qualified requesters may obtain copies of this report from DDC. This report has been furnished to CFSTI for sale to the public.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Air Force Flight Dynamics Laboratory Wright-Patterson AFB, Ohio 45433

13. ABSTRACT

The purpose of this report is to discuss computer implementations of tests for normality. Some of the underlying statistical considerations and random process theory are briefly reviewed to provide a background for subsequent material.

Methods for digitally computing basic statistical parameters, such as the mean, variance, skewness, kurtosis, etc., are given, along with procedures for computing the sample probability density function. Problems arising from the digital and discrete nature of the data are discussed, as well as sample sizes required for the tests.

The central part of the report consists of a detailed discussion of the computer implementation of the chi-square goodness-of-fit test as applied to testing for normality, followed by computer program flow charts, suitable for use in coding the procedure for a digital computer. Sufficient detail has been provided so that a programmer unfamiliar with the material should be able to write a program which will make maximum use of the time which is frequently lost during periods of input and output of data from the computer.

The report concludes with an appendix which lists standard approximations for some of the mathematical functions required in the calculations.

DD FORM 1 JAN 64 1473

UNCLASSIFIED
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Digital Computing Statistics Probability Density Functions						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.