

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS -1963-A

2

NPS-53-83-0006

# NAVAL POSTGRADUATE SCHOOL Monterey, California





TABLES FOR A NEW MULTIVARIATE

GOODNESS-OF-FIT TEST

Toke Jayachandran Richard Franke

Approved for public release: distribution unlimited

Prepared for:

Naval Postgraduate School Monterey, CA 93940

83 04 05 078

#### NAVAL POSTGRADUATE SCHOOL Monterey California 93940

Rear Admiral J.J. Ekelund Superintendent

David A. Schrady Provost

Reproduction of all or part of this report is authorized.

This report was prepared by:

Associate Professor Department of Mathematics

Associate Professor

Department of Mathematics

Reviewed by:

Department of Mathematics

Released by:

Dean of Research

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Date Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM		
	3. RECIPIENT'S CATALOG NUMBER		
NPS-53-83-0006 AD-A 126 290			
4. TITLE (and Subtitle)	S. TYPE OF REPORT & PERIOD COVERED		
Tables For A New Multivariate	Interim 1983		
Goodness-Of-Fit Test	6. PERFORMING ONG. REPORT NUMBER		
7. AUTHOR(e)	S. CONTRACT OR GRANT NUMBERY		
Toke Jayachandran			
Richard Franke			
3. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM EL EMENT, PROJECT, TARK		
Naval Postgraduate School	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS		
Monterey, CA 93940			
11. CONTROLLING OFFICE NAME AND ADDRESS	February 1983		
Naval Postgraduate School Monterey, CA 93940	13. NUMBER OF PAGES		
	15. SECURITY CLASS, (of this report)		
14. MONITORING AGENCY NAME & ADDRESS(II different from Controlling Office)	Unclassified		
	154. DECLASSIFICATION/DOWNGRADING		
17. DISTRIBUTION STATEMENT (of the obstract entered in Block 20, if different fro	er Report)		
IS. SUPPLEMENTARY NOTES			
goodness-of-fit, multivariate, Foutz test, Monte			
We present tables of critical values for a new m			

## TABLES FOR A NEW MULTIVARIATE GOODNESS-OF-FIT TEST

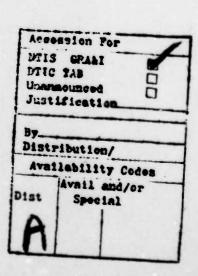
By

R. Franke and T. Jayachandran Department of Mathematics Naval Postgraduate School Monterey, CA 93940

#### ABSTRACT

We present tables of critical values for a new multivariate goodness-of-fit test introduced by Foutz. Some details of our improved asymptotic approximation and evaluation of its accuracy are given.





#### 1. Introduction

Foutz [1] proposed a new goodness-of-fit test for fitting univariate as well as multivariate distributions. He showed that the null distribution of the test statistic,  $F_n$ , does not depend on (1) the hypothesized distribution, or (2) the number of components in the random vector under study. An integral representation for the null CDF of  $F_n$  was provided. Closed form expressions for this null distribution are quite difficult to obtain, even for small sample sizes. The alternative has been to approximate the distribution by a normal distribution with mean  $e^{-1}$  and variance  $(2e^{-1} - 5e^{-2})/n$ ; this, however, does not appear to provide a good approximation to the percentiles of the null distribution of  $F_n$  for moderate sample sizes.

The authors[2] compared the  $F_n$ -test with the Chi-squared test and the Kolmogorov-Smirnov test and found that the  $F_n$ -test does have higher power when fitting certain types of distributions. Another investigation by the authors and Linhart [3] examined the power of the  $F_n$ -test when fitting a multivariate normal distribution; the test did well in detecting mean shifts and variance shifts. We therefore believe that the  $F_n$ -test is a definite alternative to the Chi-squared and Kolmogorov-Smirnov tests when fitting univariate distributions and it is just about the only available test for fitting multivariate distributions. However, the test is not very convenient for applications due to the difficulty in obtaining accurate

critical values. This paper fills the gap by providing tables of approximate pecentiles of the null distribution of  $F_n$ .

### 2. Description of the Fn-Test

The procedure for calculating the test statistic  $F_n$  is the following. Given a random sample  $\underline{x}_1$ ,  $\underline{x}_2$ , ...,  $\underline{x}_{n-1}$ , from a continuous multivariate distribution, the sample space is partitioned into n statistically equivalent blocks. Let  $h_1(\underline{x})$ ,  $h_2(\underline{x})$ , ...,  $h_{n-1}(\underline{x})$  be any n-1 "cutting functions" such that  $h_k(\underline{x})$  has a continuous distribution, k=1, 2, ..., n-1, and let  $k_1$ ,  $k_2$ , ...,  $k_{n-1}$  be a permutation of 1, 2, ..., n-1. Let  $\underline{x}(k_1)$  be the sample vector corresponding to the  $k_1$ <sup>th</sup> order statistic of  $h_k(\underline{x})$ ,  $i=1,2,\ldots,n-1$ . The initial partition of the sample space into two blocks is defined by

$$B_1 = \{\underline{x} | h_{k_1}(\underline{x}) < h_{k_1}(\underline{x}(k_1)) \}, \text{ and } B_2 = B_1^c.$$

The cutting function  $h_{k_2}(\underline{x})$  is then used to partition  $B_1$  (if  $k_2 < k_1$ ) or  $B_2$  (if  $k_2 > k_1$ ) into two subblocks in a similar fashion. When all the cutting functions are exhausted the sample space will have been partitioned into n statistically equivalent blocks,  $\beta_1$ ,  $\beta_2$ , ...,  $\beta_n$ . A convenient choice for the cutting functions in the univariate case is the identity function. In the multivariate case letting  $h_k(\underline{x}) = \underline{x}^{(j)}$ , the j<sup>th</sup> component of  $\underline{x}$  (for various j), appears to work well. More details on partitioning the sample space into statistically equivalent blocks and some examples can be found in [3].

Once the statistically equivalent blocks are determined, a computational formula for the test statistic  $\mathbf{F}_{\mathbf{n}}$  for the hypothesis that the samples are from a specified distribution H is

$$F_{n} = \sum_{i=1}^{n} \max \left[0, \frac{1}{n} - D_{i}\right],$$

where 
$$D_i = P[\underline{X} \in \beta_i | H]$$

The integral representation for the null CDF of  $F_n$  results in the following closed form expressions for n=3, 4, and 5.

$$P[F_{3} \le x] = \begin{cases} 6x^{2} & 0 < x \le \frac{1}{3} \\ 1 - 3(\frac{2}{3} - x)^{2} & \frac{1}{3} < x \le \frac{2}{3} \\ 1 & x > \frac{2}{3} \end{cases}$$

$$P[F_{4} \le x] = \begin{cases} 20x^{3} & 0 \le x \le \frac{1}{4} \\ -20x^{3} + 18x^{2} - \frac{9}{4}x + \frac{1}{16} & \frac{1}{4} < x \le \frac{1}{2} \\ 1 - 4(\frac{3}{4} - x)^{3} & \frac{1}{2} < x \le \frac{3}{4} \end{cases}$$

$$P[F_{5} \le x] = \begin{cases} 70x^{4} & 0 \le x \le \frac{1}{5} \\ -105x^{4} + 80x^{3} - 12x^{2} + \frac{16}{25}x - \frac{1}{125} & \frac{1}{5} < x \le \frac{2}{5} \\ 45x^{4} - 80x^{3} + \frac{228}{5}x^{2} - \frac{176}{25}x + \frac{31}{125} & \frac{2}{5} < x \le \frac{3}{5} \\ 1 - 5(\frac{4}{5} - x)^{4} & \frac{3}{5} < x \le \frac{4}{5} \end{cases}$$

It does not appear to be possible to generate a closed form expression for the CDF of  $\mathbf{F}_n$  in the general case. Foutz's large sample normal approximation is given by

(1) 
$$P[F_n \le x] \equiv \phi \left[ \frac{n(x - e^{-1})}{((2e^{-1} - 5e^{-2})n)^{1/2}} \right]$$

where  $\phi$  is the standard normal CDF. To check the accuracy of this approximation in our earlier study [2], we generated samples of size n-1 = 20, 30, and 50 from a uniform distribution on [0,1] and tested the hypothesis that the the samples are in fact from that distribution. The empirical significance levels in 80,000 replications are given in Table 1.

Nominal Significance Level	<u>n-1</u>	20	<u>30</u>	<u>50</u>
0.10		0.0757	0.0800	0.0859
0.05		0.0372	0.0399	0.0428
0.01		0.0082	0.0083	0.0093

Table 1
Empirical Significance Level
(Based on 80,000 replications)

It can be seen that the observed significance levels are consistently smaller than the nominal values by about 10-20%. We therefore proposed the use of Monte Carlo critical values, which were based on 25,000 replications. These values are given in Table 2 and the corresponding observed significance levels, based on 225,000 subsequent repetitions, are given in Table 3.

Significance	<u>n-1</u>				
Level	<u>20</u>	<u>30</u>	<u>50</u>		
0.10	0.42714	0.41903	0.40816		
0.05	0.44865	0.43553	0.42116		
0.01	0.48659	0.46579	0.44487		

Table 2
Monte Carlo Critical Values
(Based on 25,000 replications)

Nominal Significance		n-l	
Level	20	30	50
0.10	0.1006	0.9700	0.1003
0.05	0.0486	0.0486	0.0498
0.01	0.0103	0.0101	0.0102

Table 3
Empirical Significance Level
(Based on 225,000 replications)

The above findings lead us into a search for an improved approximation for determining the percentiles of the null distribution of  $F_n$ . We found that allowing the mean and variance to be functions of the sample size leads to greatly improved approximations. While it is difficult to give precise error bounds on the percentile values, our computational experience indicates about a four decimal place accuracy. This leads to rejection rates with errors in the fourth decimal place, usually. Comparing the error in the rejection rates for the asymptotic approximation (1) given by Foutz, our approximation is better by a factor of 10 or more.

#### 3. Modified Normal Approximation

The data for the approximation of the null distribution of the Foutz statistic was obtained by Monte Carlo methods. For a given sample size n-1, sequences of n-1 uniformly distributed numbers where generated using the IMSL\* random number generator GGUBS. The Foutz statistic was then computed and tabulated into one of 200 equilength intervals. This process was replicated 25,000 times. The entire set consists of the empirical cumulative distribution functions obtained from this data for 60 sample sizes, n-1 = 2(1)40, 40(2)70, and 70(5)100. Potentially this yields as many as 12000 pieces of data, however if only intervals with nontrivial data in them are counted, this is reduced to about 4700.

A data fitting problem with 4700 points is not easily handled unless a linear model is accepted. We do not know the behavior of the distribution as the sample size gets large, so we were reluctant to impose a form with only linear parameters, especially in sample size. We decided on attempting a correction to the asymptotic approximation given by Foutz.

After some experimentation with various types of corrections, it was decided the most reasonable was to include correction terms in the argument of the asymptotic approximation. In order to make the computation feasible it was decided to fit the data in a two pass scheme; first the null distribution for each sample

<sup>\*</sup>International Mathematics and Statistical Libraries, 7500 Bellaire Drive, Houston, TX 77036

size was approximated as below, and then the parameters in these approximations were fit by functions of sample size.

The precise form of the approximation was through the argument of a normal distribution, which was taken to be of the form

$$(a + b n(x-e^{-1}) + c (x-e^{-1})^2) / \sqrt{(2^{-1} - 5e^{-2})n}$$

Because we are strongly interested in the inverse CDF, the data was weighted at each point by the centered difference from the Monte Carlo data, which then resulted in a greater weight on the part of the curve with a large slope. The results of this least squares process yielded a table of values of a, b, and c versus sample size (actually we consider them as functions of n = sample size + 1). We observe that the amount of scatter increases as n increases. There tends to be even more scatter with higher powers of (x-e<sup>-1</sup>). For this reason it was decided to weight the smaller sample sizes more heavily, and a weight of 1/n was adopted. Since the data is more dense for smaller sample sizes this results in considerably less weight for the large sample sizes, although we feel the trend is still properly modelled and that our approximation is considerably better than the asymptotic approximation for very large sample sizes, say even up to 1000.

In the second stage of the process the coefficients a, b, and c was chosen to allow a rate of decay (or growth) of the coefficients to be dictated by the data. Thus we fit a, b and c with functions of the form  $A + Bn^{C}$ .

For the terms which are constant and linear in  $(x-e^{-1})$  the exponent was negative, however, for C(n) the exponent was positive, indicating that the term grows (somewhat slower than linearly) with sample size. We do not consider this as bothersome, however, since the linear term in  $(x-e^{-1})$  has already (due to the form of the asymptotic approximation) been included with a factor that grows linearly with sample size.

The overall result of this nonlinear least squares approximation is the approximate CDF involving the nine parameters,

(2) 
$$P[F_{n} < x] \quad \phi \quad [(g(x)/\sqrt{(2 e^{-1} - 5 e^{-2})n})] ,$$
where  $g(x) = a(n) + b(n) \cdot n \cdot (x-e^{-1}) + c(n) \cdot (x-e^{-1})^{2}$ , and
$$a(n) = 0.2089 + 0.1876 \cdot n^{-1.4416},$$

$$b(n) = 1.0015 - 0.05672 \cdot n^{-0.7377},$$

$$c(n) = 0.3049 - 0.5912 \cdot n^{0.8927}.$$

In order to test our results, two different approaches were taken. First, the number of rejections for previously run tests were available for sample sizes of n-1 = 20, 30, and 50, at (approximately) the 0.10, 0.05, and 0.01 levels. By computing the derivative of the approximate CDF, equation (2), and making a correction along the tangent line, we were able to estimate the anticipated rejection rate that would occur with our present approximation. This data was accumulated over 225,000 replications, and is given in Table 4. The main entry

is the anticipated rejection rate when using the results of our approximation, above. As a point of comparision with Foutz's asymptotic approximation, we include the corresponding rates for it in parenthesis. Second, to test the approximation for a smaller, as well as an intermediate sample size, we computed the Foutz statistic for 300,000 uniformly distributed samples of sizes 10 and 40, and tabulated them at intervals of .0001 in the range of interest. The results of these calculations are shown in Table 5 for the 0.10, 0.05, and 0.01 levels.

Nominal Significance	n-1			
Level	20	30	50	
0.10	0.0994 (0.0764)	0.1002 (0.0801)	0.1007 (0.0840)	
0.05	0.0496 (0.0385)	0.0500 (0.0402)	0.0505 (0.0420)	
0.01	0.0098 (0.0085)	0.0095	0.0098	

Table 4
Anticipated Rejection Rates
From Approximate Critical Values
(Based on 225,000 replications)

Nominal Significance	80 TH - W 10 ME	n-1
Level	10	40
0.10	0.0989 (0.0687)	0.0998 (0.0824)
0.05	0.0481 (0.0349)	0.0491 (0.0087)
0.01	0.0086 (0.0069)	0.0098 (0.0087)

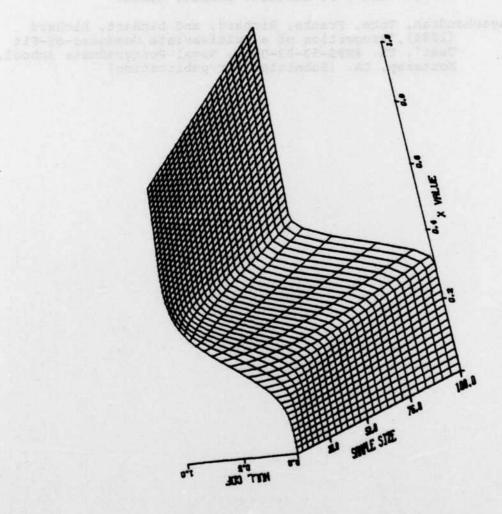
Table 5
Empirical Significance Levels
(Base on 300,000 replications)

As is shown by the tables, we expect the error in the rejection rates due to use of our approximate percentiles to be smaller by a factor of 10-20 for the 0.20 to 0.05 level than they are for Foutz's normal approximation. At the extreme tails, our approximation is not as good as at the more moderate levels, but is still a worthwhile improvement over the asymptotic approximation.

Table 6 lists some upper percentiles of the approximate CDF given by Equation (2) for sample sizes 5(1)100, 100(10)200, and 200(100)1000. The exact values are given for n-1=2, 3, and 4. Since we expect the entries to have about 4 digit accuracy, linear interpolation for intermediate sample sizes will have comparable accuracy. Linear interpolation in the percentiles is not accurate, and other percentiles should be calculated from equation (2). It is interesting to observe the "surface" of the null CDF in a perspective plot, as in Figure 1. Of course, only discrete slices exist; the cross section lines in the direction of sample size are an artifact of the plotting package. The convergence toward a sharp rise of the CDF in the vicinity of  $x - e^{-1}$  as sample size increases is very apparent.

#### REFERENCES

- 1. Foutz, Robert V. (1980), "A Test for Goodness-of-Fit Based on an Empirical Probability Measure", Annals. of Statistics 8, 989-1001.
- Franke, Richard, Jayachandran, Toke (1983), "An Empirical Investigation of the Properties of a New Goodnessof-Fit Test", J. Statist. Comput. Simul.
- 3. Jayachandran, Toke, Franke, Richard, and Linhart, Richard (1980), "Properties of a Multivariate Goodness-of-Fit Test", TR. #NPS-53-82-0001, Naval Postgraduate School, Monterey, CA. (Submitted for publication)



'igure 1: Null CDF of Fn

n-1	0.80	0.85	0.90	0.95	0.975	0.99	0.995
234567890123456789012345678901234567890123456789012345678901 11111111111222222222222333333333344444444	78989624981109293110929311092507422224469333109999112573337411109276881109211194987777889896214411109276881110921119498777788787878787878787878787878787878	68546547561638243773374618928557163213594975444681594962076 38976310022505298777880247047160505162840629629630742964197520 433333332222221154333210000000000000000111000000999988877777 444444444444444444444444444	930466629419550629354016722410386683978938410024727308767794683844334944455185243210579241038668383978938441002472730876779466655554344444444444444444444444444444	724661367547842818889122944130287188286817557163223605197779 779362700672015188889122944130287188286817557163223605197779 7793694940774186531986433953159999223222222222222222222111111111 779369494077418631986433953159999223222222222222222221111111111	8078754485996108160177681735526746462277561739275572866816298802	380484031285736265776088136773734634116461015213731248410022 92714313700796606777678881364674321123457913731248410022 065555555555555555555544444444444444444	487784443337731459332467349642959194401783251114080410015187926-8217717653467500423885457388642675703828441853251114080410015187926-666659555555555555554949878770777777777777777777777777777777

Approximate percentage points for the null distribution of the routz statistic (Note: sample size is n-1)

\* from exact distribution

n-1	0.80	0.85	0.90	0.95	0.975	0.99	0.995
123456789012345678901234567890123456789000000000000000000000000000000000000	00000000000000000000000000000000000000	5444444578136926059495062952974298643109889924525351067431354 6664668643169266059495062952974298643109889924525351067431354 79996959555549746431986542952322222218176542191002247179777777777777777777777777777777777	1482627300753100990023357925815948388406253355415475272777543 4404333322223197531009995428159483888406253355415475272777543 44043333222231015130897531599754258159888819888879478875478777777777777777777777777	14939631988891369271628418531098887889013331370217883003904 65554444333185207411964196428418531098877766666421964345753344003904 1111111111111000999988888777406666421098757334491103 144444331222211110009999888888888888888888888888	6163109013604951864322345703604949517429198558133960665177	60642222470517533223469261629631099999124697618101233392935971 2776661617283333333395173951739517395173333322233469999912469761810999999951246976551 37776665144333333333333333332222222222222222222	209025952099026040742111246937273964209969161397526044761514816044833729752609447661514434241100998877766596544333332221110099969161327297669967186514434444444444444444444444444444444444

Approximate percentage points for the null distribution of the Foutz statistic (Note: sample size is n-1)

#### DISTRIBUTION LEST

1.	Defense Technical Information Center Cameron Station Alexandria, Virginia 22217	2
2.	Dudley Knox Library Code 0142 Naval Postgraduate School Monterey, California 93940	2
3.	Dean of Research Code 012 Naval Postgraduate School Monterey, California 93940	2
4.	Department of Mathematics Naval Postgraduate School Monterey, California 93940	
	Professor C.O. Wilde, Code 53Wm Chairman	1
	Professor R. Franke, Code 53Fe Professor T. Jayachandran, Code 53Jy	10 15
5.	Dr. Richard Lau Office of Naval Research 1030 East Green Street Pasadena, California 91106	1
6.	Professor R.E. Barnhill Department of Mathematics University of Utah Salt Lake City, Utah 84112	. 1
7.	Professor G.M. Nielson Department of Mathematics Arizona State University Tempe, Arizona 85281	1
8.	Chief of Naval Research ATTN: Mathematics Program Arlington, Virginia 22217	2
9.	Professor Robert V. Foutz Department of Mathematics VPI	1
	Blacksburg, Virginia 24061	

