

Calhoun: The NPS Institutional Archive DSpace Repository

# Tables for a new multivariate goodness-of-fit test 

Jayachandran, Toke; Franke, Richard H.
Monterey, California. Naval Postgraduate School
https://hdl.handle.net/10945/30238

## NPS-53-83-0006 <br> NAVAL POSTGRADUATE SCHOOL Monterey, California



TABLES FOR A NEW MULTIVARIATE<br>GOODNESS-OF-FIT TEST<br>Toke Jayachandran<br>Richard Franke

FedDocs
D 208.14/2
NPS-53-83-0006

## 

NAVAL POSTGRADUATE SCHOOL Monterey California 93940

Rear Admiral J.J. Ekelund Superintendent<br>David A. Schrady Provost

Reproduction of all or part of this report is authorized.
This report was prepared by:

17. DISTRIBUTION STATEMENT (Of the abotract ontered in Block 20, If different troan Report)
18. SUPPLEMENTARY NOTES
19. KEY WORDS (Continue on reverse alde if neceeeary and ldentlity by block number)
goodness-of-fit, multivariate, Foutz test, Monte Carlo, percentage points
20. ABSTRACT (Continus on reverse alde 18 necsesery and idently by block mumber)

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.
冨

## 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 [l] 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 $n u l l$ 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 $\left(2 e^{-1}-5 e^{-2}\right) / n$; 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 $\mathrm{F}_{\mathrm{n}}$-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 $\mathrm{I}_{\mathrm{n}}$-test is a definite alternative to the Chi-squared and KolmogorovSmirnov 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 $\mathrm{F}_{\mathrm{n}}$-Test

The procedure for calculating the test statistic $F_{n}$ is the following. Given a random sample $\underline{x}_{1}, \underline{x}_{2}, \ldots, \underline{x}_{n-1}$, from a continuous multivariate distribution, the sample space is partitioned into $n$ statistically equivalent blocks. Let $h_{1}(\underline{X}), h_{2}(\underline{X}), \ldots, h_{n-1}(\underline{X})$ be any $n-1$ "cutting functions" such that $h_{k}(\underline{X})$ has a continuous distribution, $k=1$, $2, \ldots, n-1$, and let $k_{1}, k_{2}, \ldots, k_{n-1}$ be a permutation of $1,2, \ldots, n-1$. Let $\underline{X}\left(k_{1}\right)$ be the sample vector corresponding to the $k_{1}$ th order statistic of $h_{k_{1}}\left(X_{i}\right), i=1,2, \ldots, n-1$. The initial partition of the sample space into two blocks is defined by

$$
\begin{aligned}
& B_{1}=\left\{\underline{X} \mid h_{k_{1}}(\underline{X})<h_{k_{1}}\left(\underline{X}\left(k_{1}\right)\right)\right\}, \text { and } \\
& B_{2}=B_{1}^{C} .
\end{aligned}
$$

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, $\beta_{1}, \beta_{2}, \ldots, \beta_{n}$. 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 $j$ th 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 $F_{n}$ 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 $D_{i}=P\left[\begin{array}{lll}\underline{X} & \varepsilon & \left.\beta_{i} \mid H\right]\end{array}\right.$
The integral representation for the null CDF of $\mathrm{F}_{\mathrm{n}}$ results in the following closed form expressions for $\mathrm{n}=3,4$, and 5.

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

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

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

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

$$
\begin{equation*}
P\left[F_{n} \leq x\right] \equiv \phi\left[\frac{n\left(x-e^{-1}\right)}{\left(\left(2 e^{-1}-5 e^{-2}\right) n\right)^{1 / 2}}\right] \tag{1}
\end{equation*}
$$

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

| $\mathrm{n}-1$ | $\frac{20}{}$ | $\underline{30}$ | $\underline{50}$ |
| :--- | :---: | :---: | :---: |
|  | 0.0757 | 0.0800 | 0.0859 |
|  | 0.0372 | 0.0399 | 0.0428 |
|  | 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, base? on 225,000 subsequent repetitions, are given in Table 3.

Significance
Level
0.10
0.05
0.01

| $n-1$ |  |  |
| :---: | :---: | :---: |
| $\underline{20}$ | $\underline{30}$ | $\underline{50}$ |
| 0.42714 | 0.41903 | 0.40816 |
| 0.44865 | 0.43553 | 0.42116 |
| 0.48659 | 0.46579 | 0.44487 |

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

Nominal
Significance
Level
0.10
0.05
0.01

| $\frac{n-1}{20}$ | 30 | 50 |
| :---: | :---: | :---: |


| 0.1006 | 0.9700 | 0.1003 |
| :--- | :--- | :--- |
| 0.0486 | 0.0486 | 0.0498 |
| 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 $n-l$ 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

[^0]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

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

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 $\left(x-e^{-1}\right)$. 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 $A+B n^{C}$.

For the terms which are constant and linear in $\left(x-e^{-1}\right)$ 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 $\left(x-e^{-1}\right)$ 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 $C D F$ involving the nine parameters,

$$
\begin{align*}
& P\left[F_{n}<x\right] \quad[ {\left[\left(g(x) / \sqrt{\left.\left(2 e^{-1}-5 e^{-2}\right) n\right)}\right]\right.}  \tag{2}\\
& \text { Where } g(x)= a(n)+b(n) n\left(x-e^{-1}\right)+c(n)\left(x-e^{-1}\right)^{2}, \text { and } \\
& a(n)=0.2089+0.1876 n^{-1.4416} \\
& b(n)=1.0015-0.05672 n^{-0.7377} \\
& c(n)=0.3049-0.5912 n^{0.8927}
\end{align*}
$$

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 $C D F$, 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

Level
0.10
0.05
0.01

| $\mathrm{n}-1$ |  |  |
| :---: | :---: | :---: |
| $\underline{20}$ | $\underline{30}$ | $\underline{50}$ |
| 0.0994 | 0.1002 | 0.1007 |
| $(0.0764)$ | $(0.0801)$ | $(0.0840)$ |

$$
\begin{array}{ccc}
0.0496 & 0.0500 & 0.0505 \\
(0.0385) & (0.0402) & (0.0420)
\end{array}
$$

$$
\begin{array}{cc}
0.0098 & 0.0095 \\
(0.0085) & (0.0086)
\end{array}
$$

$$
0.0098
$$

(0.0088)

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

| Nominal |
| :---: |
| Significance |
| Level |

0.10
0.05
0.01

|  | $n-1$ |
| :---: | :---: |
| 10 | $\underline{40}$ |

$$
\begin{array}{cc}
0.0989 & 0.0998 \\
(0.0687) & (0.0824)
\end{array}
$$

$$
0.0481 \quad 0.0491
$$

$$
(0.0349) \quad(0.0087)
$$

$$
\begin{array}{cc}
0.0086 & 0.0098 \\
(0.0069) & (0.0087)
\end{array}
$$

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 $C D F$ 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 $C D F$ 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.
2. Franke, Richard, Jayachandran, Toke (1983), "An Empirical Investigation of the Properties of a New Goodness-of-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)



TABLE 6

```
Appicximate percontage poinこs for, the mull
                dis=ribution of thミ Foutz staristシc
                            (No:e: sample size is n-1)
```



TABLE 6 （こつnネシャued）
Approximate percentage pozrts eor the null
dシsさribuにion of the Fourz seatiseこic
（Note：sample size is n－1）

1. Defense Technical Information Center ..... 2
Cameron Station
Alexandria, Virginia ..... 22217
2. Dudley Knox Library ..... 2
Code 0142
Naval Postgraduate School
Monterey, California 93940
3. Dean of Research ..... 2
Code 012
Naval Postgraduate School
Monterey, California 93940
4. Department of Mathematics
Naval Postgraduate School
Monterey, California 93940
Professor C.0. Wilde, Code 53Wm ..... 1
Chairman
Professor R. Franke, Code 53Fe ..... 10
Professor T. Jayachandran, Code 53Jy ..... 15
5. Dr. Richard Lau ..... 1
Office of Naval Research
1030 East Green Street
Pasadena, California 91106
6. Professor R.E. Barnhill ..... 1
Department of Mathematics University of Utah
Salt Lake City, Utah ..... 84112
7. Professor G.M. Nielson
Department of Mathematics
Arizona State University
Tempe, Arizona 85281
8. Chief of Naval Research ..... 2
ATTN: Mathematics Program
Arlington, Virginia 22217
9. Professor Robert V. Foutz
Department of Mathematics ..... VPI
Blacksburg, Virginia ..... 24061

DUOLEY KNOX LIBRARY

32768003417965


[^0]:    *International Mathematics and Statistical Libraries, 7500 Bellaire Drive, Houston, TX 77036

