Proceedings of the 2000 Winter Simulation Conference
J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.

ANALYSIS OF SIMULATION FACTORIAL EXPERIMENTS BY EDF RESAMPLE STATISTICS

Russell C. H. Cheng
Owen D. Jones

Faculty of Mathematical Studies
University of Southampton
Southampton, SO17 1BJ, UNITED KINGDOM

ABSTRACT

The output from simulation factorial experiments can be complex and may not be amenable to standard methods of estimation like ANOVA. We consider the situation where the simulation output may not satisfy normality assumptions, but more importantly, where there may be differences in output at different factor combinations, but these are not simply differences in means. We show that EDF statistics can provide a similar but potentially more sensitive analysis to that provided by ANOVA. Moreover we show that with the use of resampling, we can generate accurate critical values for tests of hypothesis under much weaker conditions than those required for ANOVA tests. The method is illustrated with an example based on an actual simulation experiment comparing two methods of operating a production facility under different production levels.

1 INTRODUCTION

We consider the analysis of data obtained from a factorial simulation experiment. The use of linear models for modelling data of this type and their study using analysis of variance is one of the most well known and used of statistical techniques. The simplicity of the assumptions and the flexibility of the method of analysis makes it a powerful and attractive approach.

However one situation where a linear model may not be adequate is when differences in the behaviour of the response variable Y at different factor level combinations are not explainable simply in terms of differences in the means, but requires a comparison of other features of a distribution, such as variance, or shape. Thus in what follows we wish to remove the usual requirements of normality and homoscedasticity assumed in a linear model.

Though we drop these two requirements, we shall however retain, as far as possible, the underlying geometric framework of the linear model. As with ANOVA, where differences in means are measured by squared differences, we shall develop an analogue where more general differences are measured by squared components.

Contenders for the statistic to use as a basis of our analysis are Empirical Distribution Function Integral Test (EDFIT) statistics of goodness-of-fit such as the Cramer-von Mises and Anderson-Darling statistics (The seminal paper is Anderson and Darling, 1952). These are well-known to provide good tests for detecting differences between different distributions. Such statistics are not as widely used as they should be, probably because their distributions, even under the null, are not simple to evaluate, being typically that of the weighted sum of chi-squared variables with one degree of freedom. Moreover the distribution changes significantly if parameters have to be estimated.

These difficulties can be overcome by the use of bootstrap and resampling methods which enable the distributions of the test statistics to be directly generated as part of the statistical analysis.

In this article we describe a method for using such EDFIT statistics for the analysis of factorial experiments. The main features of the method are:

(i) The method is essentially non-parametric, but provides a method of analysis that retains the attractive structural characteristics of Analysis of Variance. In terms of generality of application it thus matches that of ANOVA techniques.

(ii) The method is applicable under much weaker assumptions than that assumed in the linear model. In particular we drop the requirements of normality and homoscedasticity.

(iii) The method detects general differences between distributions. Thus, though it will detect differences between means, it will also detect differences in variance or shape. It is thus potentially much more sensitive than ANOVA.

(iv) Because resampling procedures are used in the analysis, results can be readily presented graphically, giving non-specialists a more natural feel and interpretation of findings.

We shall not develop the method in full generality but consider two cases: the one and two way models. It should be clear how the technique generalises from these two examples.

2 CLASSIFICATION MODELS

2.1 One Way Model

Consider the one way model with t treatments, and n_i observations made with treatment i. The standard ANOVA model of the observations is then:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, ..., t, \quad j = 1, ..., n_i$$

where the ε_{ij} are all identically distributed with $\mathrm{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$. This yields the ANOVA decomposition of this model as

$$CSS(total) = TreatSS + ResidSS \tag{1}$$

where

$$CSS(total) = \sum_i \sum_j (y_{ij} - \bar{y})^2$$

and

$$TreatSS = \sum_j n_i (\bar{y}_i - \bar{y})^2$$

are respectively the corrected total and treatment sums of squares The residual sum of squares is found by differencing, or directly from the formula

$$ResidSS = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2.$$

For our purpose, the key quantity is $TreatSS$. The EDFIT analogue of this quantity is obtained as follows.

We suppose that the observations under treatment i have distribution with Cumulative Distribution Function (CDF) $F_i(\cdot)$. Thus the (non-parametric) model is

$$y_{ij} \sim F_i(\cdot) \quad i = 1, ..., t, \quad j = 1, ..., n_i$$

Let

$$\tilde{F}_i(y) = n_i^{-1} \sum H(y - y_{ij})$$

where

$$\begin{aligned} H(y) &= 1 \quad \text{if} \quad y \geq 0 \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

be the Empirical Distribution Function (EDF) of the sample $\{y_{i1}, ..., y_{in_i}\}$ corresponding to treatment i. The total, combined, sample can be regarded as being drawn from the composite distribution with CDF

$$F(y) = \sum_{i=1}^{t} w_i F_i(y) \tag{2}$$

where $w_i = n_i / \sum_{j=1}^{t} n_j$. The EDF of the combined sample then has the attractive alternative analogous representation:

$$\bar{F}(y) = \sum_{i=1}^{t} w_i \tilde{F}_i(y).$$

The EDFIT statistic takes the form

$$T = \sum_{i=1}^{t} n_i \int \left(\tilde{F}_i(y) - \bar{F}(y)\right)^2 \psi[\bar{F}(y)] d\bar{F}(y). \tag{3}$$

This is the analogue of $TreatSS$. Here $\psi(\cdot)$ is a weight function. For simplicity of notation, in what follows we shall take the case $\psi(\cdot) = 1$. It will be clear that all the integral expressions given below can be generalised simply by replacing $d\bar{F}(y)$ by $\psi[\bar{F}(y)]d\bar{F}(y)$.

The sum T comprises terms each of which is the EDFIT statistic computed from the sample corresponding to treatment i, with the EDF $\bar{F}(y)$ obtained from the entire combined sample acting as the base distribution. The correspondence with (1) cannot be extended further because T, at least in the non parametric case, is the analogue not simply of $TreatSS$ but also of $CSS(total)$ as well. Thus the EDFIT equivalent of *ResidSS* is zero. We therefore cannot develop the usual F-ratio tests found in ANOVA.

[Note: A parametric form can be developed where parameters are estimated. This provides a closer analogy with the standard ANOVA case. However it is not clear that the EDFIT statistic possesses any advantages.]

To proceed we need to obtain the distribution of T under the null hypothesis:

$$H_0 : \ F_i(y) = F_0(y) \quad i = 1, ..., t.$$

Under the null, T has the form

$$T_0 = \sum_{i=1}^{t} n_i \int \left(\bar{F}_i(y) - \bar{F}(y)\right)^2 d\bar{F}(y) \tag{4}$$

where

$$\bar{F}_i(y) \quad i = 1, ..., t$$

are the EDFs of samples (of size n_i) each sampled from the same null distribution $F_0(.)$, and with $\bar{F}(y) =$

$\sum_{i=1}^{t} w_i \bar{F}_i(y)$. We show in the next section how the distribution of T_0 is easily obtained.

Though the EDFIT statistic, T, does not decompose as in (1) to give a usable *ResidSS*, it does however decompose in all other respects like a treatment sum of squares. For example, consider the case $t = 3$ where 'treatment i' corresponds to a numerical level of a single factor set at level $i = 1, 2, 3$, and that the design is orthogonal with $n_1 = n_2 = n_3 = n$. Then if T is found to be significantly different from zero, we can decompose it into two identifiable contrasts

$$T = T_1 + T_2$$

where

$$T_1 = \int \frac{n}{2} \left(\tilde{F}_3(y) - \tilde{F}_1(y)\right)^2 d\bar{F}(y)$$

and

$$T_2 = \int \frac{n}{6} [\tilde{F}_1(y) - 2\tilde{F}_2(y) + \tilde{F}_3(y)]^2 d\bar{F}(y).$$

2.2 Two Way Model

It is evident that the one way model of the previous section generalises to other ANOVA factor decompositions. As a further illustration we consider the two way replicated classification model with observations of the form

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \\ & \quad i = 1, ..., r, \quad j = 1, ..., c, \quad k = 1, ..., n. \end{aligned}$$

The sum of squares decomposition is

$$\begin{aligned} &\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2 \\ &= n \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{...})^2 + \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.})^2 \\ &= TreatSS + ResidSS \end{aligned}$$

where the first term on the right hand side is the corrected sum of squares due to treatments. This can be written in the form

$$\begin{aligned} &n \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{...})^2 \\ &= cn \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2 + rn \sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2 \\ &\quad + n \sum_i \sum_j (y_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 \end{aligned}$$

The corresponding EDF decomposition is

$$\begin{aligned} &n \sum_i \sum_j \int \left(\bar{F}_{ij}(y) - \bar{F}_{..}(y)\right)^2 d\bar{F}_{..}(y) \\ &= \int \Big\{ cn \sum_i [\bar{F}_{i.}(y) - \bar{F}_{..}(y)]^2 \\ &\quad + rn \sum_j [\bar{F}_{.j}(y) - \bar{F}_{..}(y)]^2 \\ &\quad + n \sum_i \sum_j [\bar{F}_{ij}(y) - \bar{F}_{i.}(y) \\ &\quad - \bar{F}_{.j}(y) + \bar{F}_{..}(y)]^2 \Big\} d\bar{F}_{..}(y) \end{aligned}$$

with the obvious interpretation that $\bar{F}_{ij}(y)$ is the EDF of the sample at factor combination (i, j), $\bar{F}_{i.}(y)$ is the EDF obtained by combining all the samples corresponding to row factor level i, $\bar{F}_{.j}(y)$ is the EDF obtained by combining all the samples corresponding to column factor level j, and $\bar{F}_{..}(y)$ is the EDF obtained by combining all samples.

3 RESAMPLING

3.1 Distribution Free Tests

The distribution of T_0 (ie the distribution of T under the null) can be derived theoretically, at least asymptotically. For the two sample case see Anderson (1962) and also Baumgartner et al (1998). However we shall not do this here but instead describe a simple resampling procedure that can be used to construct, to arbitrary specified accuracy, the *exact* null distributions. The reason why this is possible is because the proposed test statistics are actually distribution free, in the sense of not depending on the distributions from which the original samples are drawn. We show this next.

Consider a typical term

$$I_k = n_k \int \left(\bar{F}_k(y) - \bar{F}(y)\right)^2 d\bar{F}(y)$$

in (4). We have

$$\bar{F}(x) = \begin{cases} 0 & x < x_1 \\ \frac{i}{n} & x_i \le x < x_{i+1}, \quad i = 1, 2, ..., n-1 \\ 1 & x_n \le x \end{cases}$$

whilst, under the null, the subsample, of size n_k, is simply a random selection of n_k out of the n observations. Suppose that the jth (ordered) value of the subsample is the $i(j)$th value of the original. Then

$$\bar{F}_k(x_i) = \frac{j}{n_k} \quad \text{for } i(j) \le i < i(j+1). \tag{5}$$

If we write therefore

$$j(i) = j \quad \text{for} \quad i(j) \le i < i(j+1)$$

then we find that

$$I_k = n_k \int \left(\bar{F}_k(y) - \bar{F}(y)\right)^2 d\bar{F}(y) \qquad (6)$$

$$= n_k \sum_i \left[n^{-1} \left(\frac{i}{n} - \frac{j(i)}{n_k} \right)^2 \right]. \qquad (7)$$

Thus I_k depends only on the position of the observations of the subsample in the full sample, and so is distribution free, and it follows that T_0 is distribution free also.

This formula for I_k is a convenient form for calculation of a typical component of the test statistic.

3.2 Null distribution by Monte Carlo

The form of (5) shows that the distribution of the test statistic T (under the null) is easily constructed by Monte Carlo simulation. We break up the set $\{1, 2, ..., n\}$ into random subsamples of sizes n_k values, compute the I_k, and then form T from (3). We repeat this B times to get B values of T: $T^{(1)}$, $T^{(2)}$, ..., $T^{(B)}$. The EDF of these $T^{(j)}$ converges to the CDF of T_0. The sample $\{T^{(i)}, \ i = 1, 2, .., B\}$ can thus be used in the usual way to calculate bootstrap critical values to test the significance of T.

An alternative way of obtaining the distribution of T_0 is simply to combine the separate samples of the original data obtained under the different treatments into a single composite sample $\{y_{ij}, \text{ all } i, j\}$. Bootstrapping can be used to obtain bootstrap samples $\{y^*_{ij}\}$ from this $\{y_{ij}\}$. Thus we write

$$\bar{F}^*_i(y) \quad i = 1, ..., k$$

for the EDFs of these bootstrap samples, which are *all* drawn from $\{y_{ij}\}$, and $\bar{F}^*(y)$ for the combined EDF. Under the null the observations in $\{y_{ij}\}$ will all come from the same distribution F_0, and

$$T^*_0 = \sum_{i=1}^{k} n_i \int \left(\bar{F}^*_i(y) - \bar{F}^*(y)\right)^2 d\bar{F}^*(y). \qquad (8)$$

then constitutes the bootstrap version of T_0.

4 APPLICATION

As an example of the above discussion we consider data from a simulation experiment comparing two different methods of operating a certain piece of equipment processing a certain product. The output comprised 782 observations of (simulated) operator activity for each method, where each observation was on a simple integral scale ranging from 0 through 5 (0 indicating low activity, 5 indicating high activity). Some preliminary analysis indicated that the observations could be assumed to be independent. The observed activity level was also expected to be dependent on the production level which could be set at one of four levels. The two methods of operation were simulated under the same operating conditions, with production levels 1, 2, 3, 4 occuring a total of 50, 140, 346 and 246 times respectively.

Histograms of the observed activity levels for the different combinations of operating method ($i = 1, 2$) and production level ($j = 1, 2, 3, 4$) are plotted in Figure 1.

Use of ANOVA is clearly not appropriate in this case. A possible though not ideal method of analysis is to use the well known Friedman non-parametric test for a two-way layout (See for example Hollander and Wolfe 1973). This test allows a number of matched samples to be compared. Thus the samples have to be of the same size with corresponding observations in each sample matched. In our case we have two samples, each corresponding to an operating method. The two operating methods can thus be regarded as 'Treatments', with each sample corresponding to one treatment and forming a *column* (in Hollander and Wolfe's notation). The layout is as given in Table 1.

Table 1: Structure of Data in the Example

Operating Method 1	Operating Method 2
x_{11}	x_{12}
x_{21}	x_{22}
$\vdots$	$\vdots$
$x_{782,1}$	$x_{782,2}$

Each row of corresponding observations thus constitutes a *matched pair of observations*, with both observations made at the same production level. The test statistic is calculated from this pair of matched samples. (Details of the calculation are given in Hollander and Wolfe.) There are a large number of ties in the observations, and these are handled using the correction method given in Hollander and Wolfe. The test value obtained was 0.465. To determine the level of significance, we used bootstrap resampling to form 1000 bootstrap test values under the null. This was done by obtaining two samples as in the layout of Table 1, with each of size 782, using bootstrap resampling, only with *both* bootstrap samples obtained from the one original sample corresponding to operating method 1. To make sure that corresponding observations in each bootstrap sample were matched, each such pair was sampled from observations made at the same production level. This ensured that there was proper matching of production levels as in the original samples.

Figure 2 shows the bootstrap EDF of 1000 values of the Friedman test statistic calculated from paired samples formed in this way. This EDF yields an estimate of the

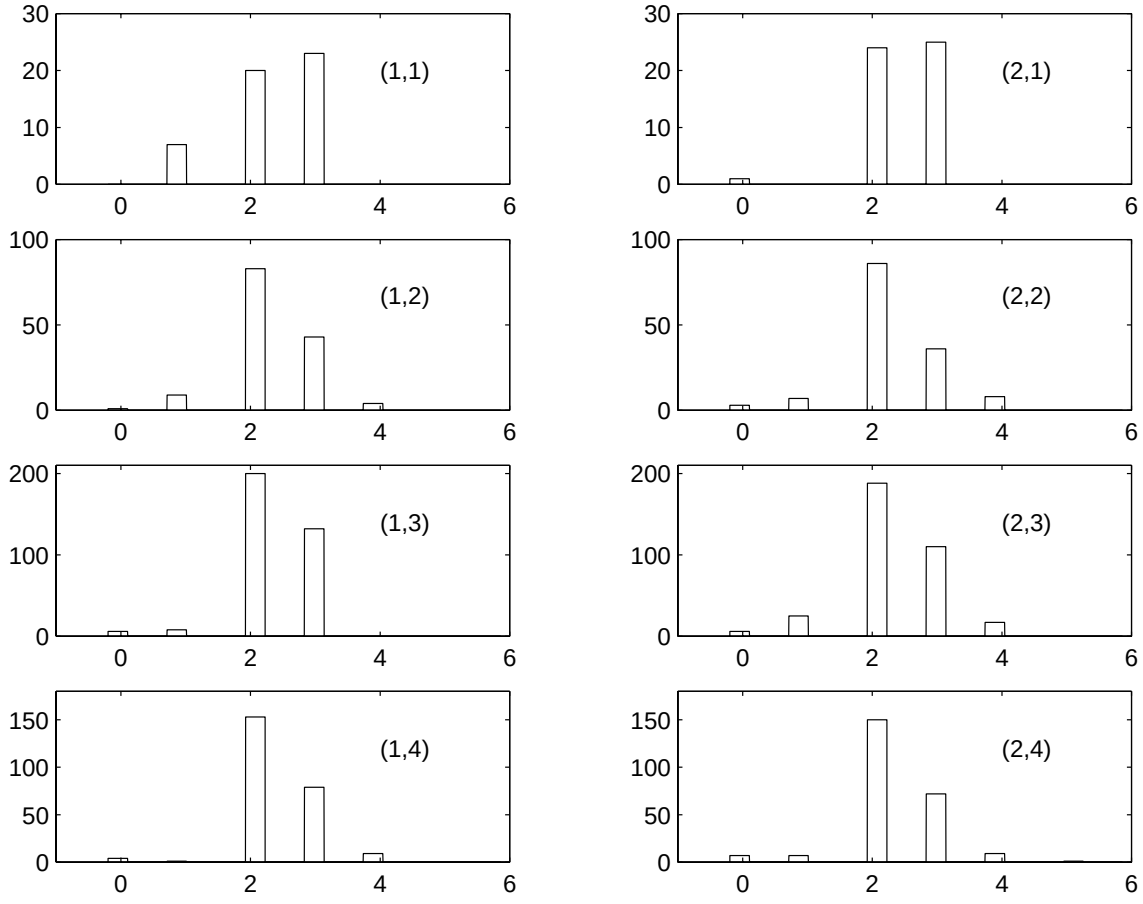

Figure 1: Activity Levels at Different Factor Combinations

critical value at the 90% level of significance as 2.806, and a value of 3.704 at the 95% level of significance. This latter value is depicted in Figure 2. The actual test value of 0.465 is therefore nowhere near significant at either of these levels.

Figure 2 also shows the EDF obtained by bootstrapping from *both* original samples. Thus, in this case, a pair of samples each of 782 observations is again formed, only with one sample obtained by bootstrapping from the first original sample (corresponding to operating method 1), and the other obtained by sampling from the second original sample (corresponding to the observations made under operating method 2). The test statistic was then calculated from this pair of bootstrap samples. The resulting EDF, $G_{Friedman}(y)$ say, thus estimates the true distribution of the test statistic as obtained in the *original* simulation. We can thus estimate the power of the Friedman test, at the 95% level of size, as being $1 - G_{Friedman}(3.704) = 0.139$. At the 90% level the power increases to 0.196. Both values are rather low.

Figure 3 shows the corresponding EDFs using the EDFIT test statistic

$$T = n \int \sum_{j=1}^{2} [\bar{F}_{.j}(y) - \bar{F}_{..}(y)]^2 d\bar{F}_{..}(y).$$

This test statistic tests for a difference between the two operating methods, whilst allowing for differences between production levels. The test value obtained from the original samples was 0.925. The resulting EDF formed under the null gave critical values of 0.747 at the 90% level, and 0.970 at the 95% level. The latter is marked in Figure 3. In this case the test value lies between the two and so is significant at the 90% level, but not at the 95% level.

The Figure also shows the estimate of the true distribution of T, $G_{EDFIT}(y)$ say, from which we can estimate the power of the test at the 95% level of size as being $1 - G_{EDFIT}(0.970) = 0.433$. This increases to 0.647 at the 90% level. Both values are more than three times that of the Friedman test.

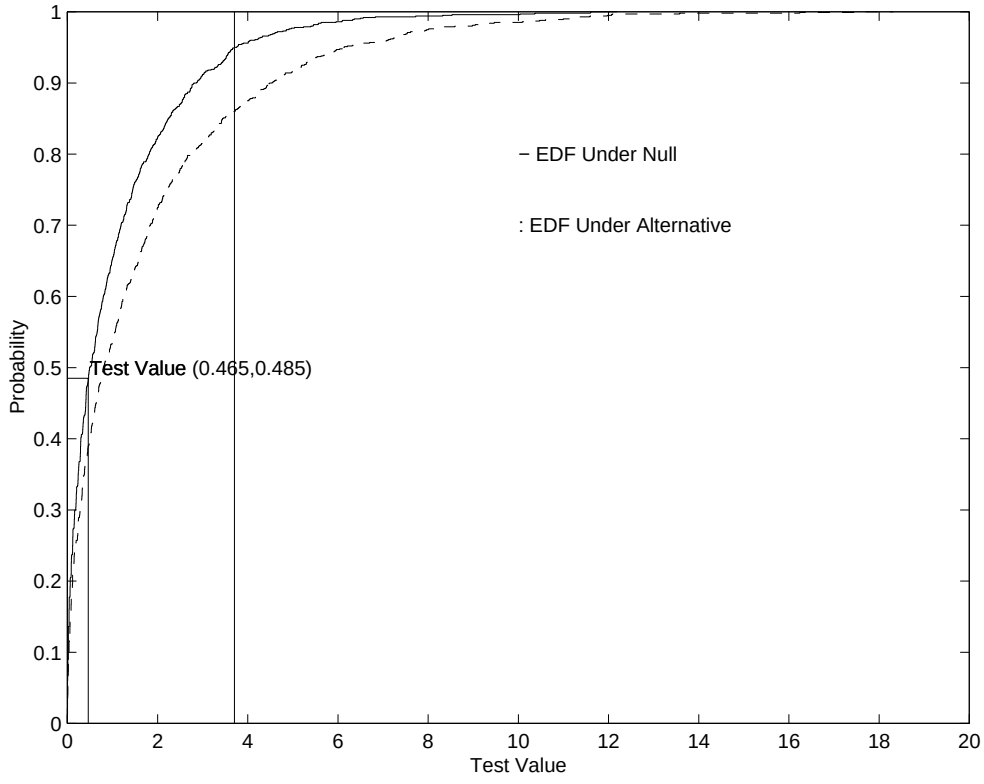

Figure 2: Bootstrap EDFs — Friedman Statistic

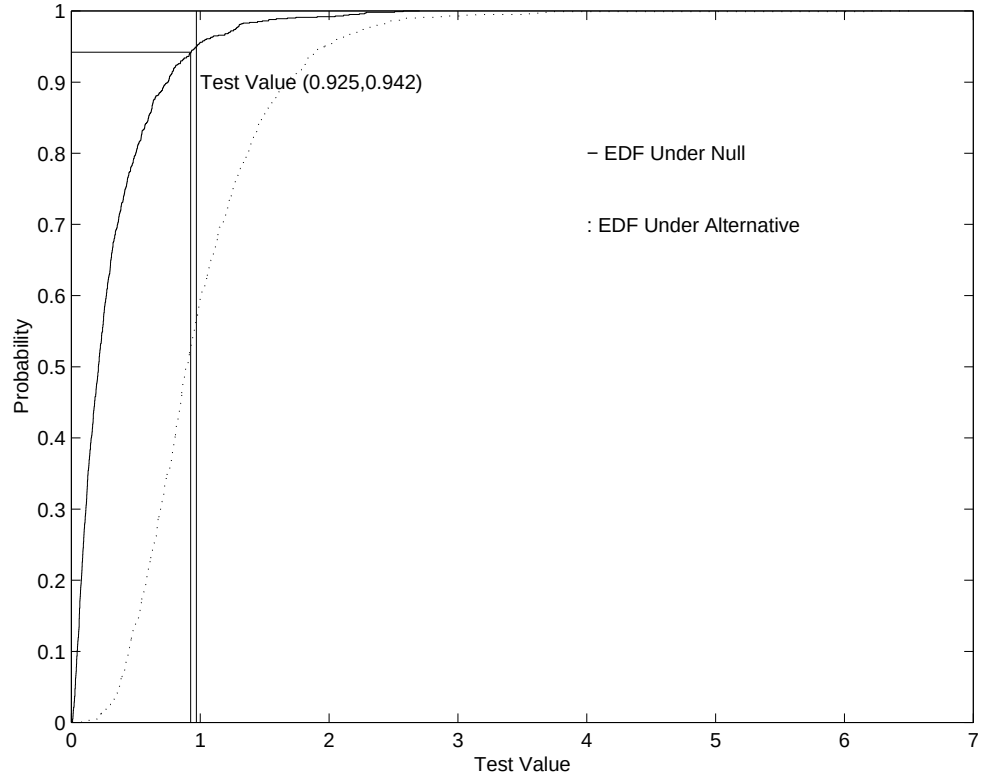

Figure 3: Bootstrap EDFs — EDFIT Statistic

5 CONCLUSION

In simulation factorial experiments where observations are replicated, the discussion suggests that the $EDFIT$ statistic can provide a much more sensitive test for distinguishing differences between responses at different factor combinations.

This is backed up by the example of Section 4, where a standard, normally quite powerful non-parametric test did not reveal differences between two treatments. In contrast the $EDFIT$ test indicated a significant difference. The reason is because there is a significant difference in the overall *variability* of the two samples even though the difference is not all that great between the overall means of the two samples. The sensitivity of the $EDFIT$ statistic to *any* difference between samples has therefore enabled this difference in variability to be detected.

REFERENCES

Anderson, T.W. 1962. On the distribution of the two-sample Cramer-von Mises criterion. *Annals of Mathematical Statistics* 33:1148–1159.

Anderson, T.W. and Darling, D. A. 1952. Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *Annals of Mathematical Statistics* 23:193–212.

Baumgartner, W., Weiss, P. and Schindler, H. 1998. A nonparametric test for the general two-sample problem. *Biometrics* 54:1129–1135.

Hollander, M. and Wolfe, D.A. 1973. *Nonparametric statistical methods.* New York: Wiley.

AUTHOR BIOGRAPHIES

RUSSELL C. H. CHENG is Professor of Operational Research at the University of Southampton. He has an M.A. and the Diploma in Mathematical Statistics from Cambridge University, England. He obtained his Ph.D. from Bath University. He is a former Chairman of the U.K. Simulation Society, a Fellow of the Royal Statistical Society, Member of the Operational Research Society. His research interests include: variance reduction methods and parametric estimation methods. He is Joint Editor of the *IMA Journal on Mathematics Applied to Business and Industry*.

OWEN D. JONES is a Lecturer in Operational Research at the University of Southampton. He is a Fellow of the Cambridge Commonwealth Trust and a member of the Australian Mathematical Society. His research interests are in stochastic analysis, in particular for self similar processes.