Interfaces with Other Disciplines

DEA-based hypothesis tests for comparing two groups of decision making units

Rajiv D. Banker a,1, Zhiqiang (Eric) Zheng b,*, Ram Natarajan c,2

a 210C Speakman Hall, Fox School of Business, Temple University, Philadelphia, PA 19122, United States
b School of Management, SM33 University of Texas at Dallas, Richardson, TX 75080, United States
c School of Management, SM44 University of Texas at Dallas, Richardson, TX 75080, United States

ARTICLE INFO

Article history:
Received 21 August 2008
Accepted 18 January 2010
Available online xxxx

Keywords:
Data envelopment analysis
Hypothesis tests
Data generating process
Efficiency comparison

ABSTRACT

In this paper, we develop five statistical tests to compare the efficiencies of different groups of DMUs. We consider a data generating process (DGP) that models the deviation of the output from the best practice frontier as the sum of two components, a one-sided inefficiency term and a two-sided random noise term. We use simulation to evaluate the performance of the five tests against the Banker test (Banker, 1993) that were designed for DGPs containing a single one-sided error term. It is found that while the Banker tests are very effective when efficiency dominates noise, the tests developed in this paper perform better than the Banker test when noise levels are significant.

1. Introduction

Data envelopment analysis (DEA) has been used and continues to be used extensively in many settings to analyze the efficiency of organizations. Despite its widespread use, Schmidt (1985) classifies DEA as a non-statistical approach and states:

“I am very skeptical of non-statistical measurement exercises, certainly as they are now carried out and perhaps in any way in which they could be carried out. … I see no virtue whatever in a non-statistical approach to data.”

Recognizing the need for a statistical foundation for DEA, Banker (1993) provided a formal statistical basis by identifying conditions under which DEA estimators are statistically consistent and likelihood maximizing. He also developed hypothesis tests for efficiency comparison when a group of DMUs (decision making units) is compared with another.

In the past two decades, efficiency comparison of different groups has become an active area of research in the DEA literature. It has been applied to a variety of industries, ranging from software engineering, banking and insurance, sports to e-commerce etc. Charnes et al. (1981) compared the efficiency of DMUs with and without program follow-through. Banker et al. (1990) studied the efficiency of fast food outlets where a specific information system was installed versus those that did not have the system installed. Banker and Kauffman (1991) investigated software programmer productivity for projects with and without a structured development methodology. Golany and Storberg (1999) compared the operational efficiencies of bank branches with PIC (personal investment center) versus those without. Cummins et al. (1999) tested agency theoretic hypotheses by comparing DEA efficiencies of stocks with mutual property-liability insurers. Avikiran (2000) examined the efficiency changes of trading banks under deregulation and compared the efficiencies of trading banks with those of regional banks. Hahn and Kauffman (2002) compared customer online shopping efficiencies under two different web site designs. Lee et al. (2009) studied the efficiencies of different R&D programs. Lewis et al. (2009) applied DEA to evaluate the performance of the MLB (Major League Baseball) and examined the efficiencies of different teams in regular season and post-season. Staub et al. (2010) compared the economic efficiencies of banks in Brazil with those in Europe and found that the Brazilian banks had lower efficiencies than their European counterparts.

It is important to note that the above literature lacks consistent statistical tests for comparing efficiencies of two samples. The choice of which statistical test to use was often ad-hoc. Cummins et al. (1999) for example, used a regression-type of parametric test, i.e., by adding a dummy variable to indicate the groups and then regressing the efficiency scores (of the two groups) on the dummy variable. Golany and Storberg (1999) on the other hand argued that non-parametric tests such as the Mann–Whitney test were more appropriate since these tests do not make assumptions on the distribution of efficiency scores. Lee et al. (2009) also used non-parametric tests (the Kruskal–Wallis and Mann–Whitney U tests).
Recognizing the difficulty of choosing appropriate tests, Hahn and Kauffman (2002) simply adopted multiple tests in their study, including the two F-tests developed by Banker (1993) under two different distributional assumptions (half-normal and exponential) of the efficiency scores, and the non-parametric Mann–Whitney test.

The objective of this paper is to fill this gap by evaluating different hypothesis-testing methods for comparing the efficiencies of two DMU groups. In particular we compare the tests developed in this paper against the F-tests developed by Banker (1993), which are conditioned on a data generating process that characterizes the deviation of the actual output from the best practice frontier as arising from a single inefficiency term. In contrast, we model the deviation as the sum of two components, a one-sided inefficiency term and a two-sided random noise term possibly bounded above, analogous to composed error term formulations in parametric stochastic frontier models (Aigner et al., 1977; Meeusen and van den Broeck, 1977; Cook and Seiford, 2009). Under this stochastic framework, we develop estimation methods and statistical tests to compare the efficiency of two groups of DMUs.

Specifically, we examine five statistical tests including two parametric and three non-parametric ones. The two parametric tests considered here are a regression-based test and a t-test for comparing means of two groups. We identify conditions under which a simple regression-based test yields a consistent estimator of the difference in mean inefficiencies between the groups. We show that under appropriate conditions the two-sample t-test can be used to compare whether the mean efficiencies of the two groups are different. The three non-parametric tests include a median test, the Mann–Whitney test and the Kolmogrov–Smirnov test. All three tests are based on order statistics. We prove that the order statistics remain consistent in the presence of two-sided noises and show that all three tests yield consistent estimates. Finally, we carry out simulation experiments to evaluate the performance of these five tests against those developed by Banker (1993). We find that tests developed in this paper perform better than the tests in Banker (1993) when noise plays a significant role in the data generating process. On the other hand, the Banker tests are more effective when efficiency dominates noise.

The paper proceeds as follows. The next section describes our general framework and the data generating process that we consider. Section 3 and 4 contains the estimation methods and statistical tests based on DEA. Experiment setups and results are presented in Sections 5 and 6 respectively. Section 7 concludes.

2. Basic model and data generating process

Consider observations on \( j = 1, \ldots, N \) decision making units (DMUs), with each observation comprising a vector of outputs \( Y_j = (y_{j1}, \ldots, y_{jk}) \) and a vector of inputs \( X_j = (x_{j1}, \ldots, x_{jk}) \). The \( N \) observations are made up of sample observations from \( M \) distinct groups with the number of observations from the \( m \)th group designated as \( N_m \) and \( \sum_{m=1}^{M} N_m = N \). Thus, for instance, in Farrell’s (1957) original setting for efficiency evaluation, the outputs \( Y \) may represent a Farm’s production measured in tons of grain, the inputs \( X \) are labor, capital and materials, and the \( M \) distinct groups may be different geographical regions where different subsets of farms are located. The fraction of \( m \)th group DMUs in the population is \( p_m \), with \( \sum_{m=1}^{M} p_m = 1 \).

We describe our basic model for the case of a single output, \( y \), to maintain direct comparison with parametric stochastic frontier models and Banker (1993). The extension to the multiple outputs case is straightforward. For ease of exposition, without loss of generality, we also restrict our attention to the case of \( K = 2 \) i.e., we assume that there are two groups of DMUs in the entire population.

The model we specify includes the true production function \( \phi(X) \) and an error term \( \varepsilon \). The production function is monotone increasing and concave in \( X \), and relates the vector \( X \) to a single output \( y \) as specified by Eq. (1), depending on which group the DMU belongs to.

\[
  y = \phi(X) + \varepsilon_m, \quad m = 1, 2.
\]

The random variable representing the error \( \varepsilon_m \) is itself generated by the process

\[
  \varepsilon_m = v - u_m, \quad m = 1, 2,
\]

where the random variable \( u_m(>0) \) represents the technical inefficiency corresponding to the \( m \)th group and has a one-sided distribution in the range \([0, \infty)\) and \( v \) represents random noise and has a two-sided distribution. Further, the probability density functions \( f_{\text{um}}(u_m) \) and \( f_{\varepsilon}(v) \) are assumed to be independent of each other. Each stochastic variable is assumed to have finite variance. It is straightforward to verify that the p.d.f. of the composed error, \( \varepsilon_m \), is given by

\[
  f_{\varepsilon}(\varepsilon_m) = \int_{\varepsilon_m}^{\infty} f_{\varepsilon}(v)f_{\text{um}}(v - \varepsilon_m) dv.
\]

We denote the mean and variance of the random noise variable \( v \) as \( \mu \) and \( \sigma_v^2 \), respectively, while \( \sigma^2_i \) and \( \sigma^2_j \) as the variances of the inefficiency random variables \( u_i \) and \( u_j \), respectively. Further, we define \( u_i \) and \( u_j \) as the mean technical inefficiencies of the two groups in the population. Our main objective is to estimate the difference between the two mean technical inefficiencies and develop statistical tests to examine whether the mean technical efficiencies of the two groups are significantly different from each other.

3. DEA estimation

The theoretical production function \( \phi(\cdot) \) is usually not observed in applied research settings. It has to be estimated from sample data which is available only in terms of input–output pairs. Guidance provided by theoretical considerations may indicate only that the production function linking the input and output is monotone and concave. In such situations non-parametric techniques such as Data Envelopment Analysis may be used to estimate the production function.

Since the input–output relationship specified in (1) includes a two-sided error term which may result in a negative value for the output, it is not always possible to apply DEA directly to (1) and estimate the production function \( \phi(\cdot) \). Therefore, we adopt a transformation of the production function similar to Gostach (1998) and Banker and Natarajan (2004). We define:

\[
  \phi(X) = \phi(X) + V^{\text{MAX}},
\]

where a positive value \( V^{\text{MAX}} \) (representing the maximum possible value could be generated from \( V \)) is added to the output. Thus, Eq. (1) is transformed to

\[
  y = \tilde{\phi}(X) - (V^{\text{MAX}} - \varepsilon_m), \quad m = 1, 2.
\]

This simple transformation guarantees that the production function \( \phi(\cdot) \) can be estimated with conventional DEA formulations using (5) because \( (V^{\text{MAX}} - \varepsilon_m) \geq 0 \). Further since \( \tilde{\phi}(X) \) is derived from \( \phi(X) \) by adding a positive constant, \( \tilde{\phi}(X) \) is also monotone increasing and concave.

We refer to \( u_m = (V^{\text{MAX}} - \varepsilon_m) \), which is the deviation of observed output \( y \) from the derived frontier \( \tilde{\phi}(X) \), as inefficiency. The inefficiency for a particular DMU \( j \) belonging to group \( m, u_m \), can be estimated from the relationship

\[
  \hat{u}_m = \tilde{\phi}(X_m) - y_m,
\]
where the frontier output \( \hat{\phi}(X_{jm}) \) is estimated by using the following DEA program which makes use of the input–output information for the entire pool of \( N \) observations:\(^5\)

\[
\phi(X_{jm}) = \text{Max} \{ \phi \mid \sum_{k=1}^{N} \lambda_k y_k = \phi \sum_{k=1}^{N} \lambda_k x_k \leq x_j, \forall i = 1, \ldots, I; \sum_{k=1}^{N} \lambda_k = 1; \lambda_k \geq 0 \}.
\]

(7)

The inefficiency estimator \( \hat{u}_{jm} \) is a consistent estimator of \( u_{jm} \) (Banker, 1993). As observed earlier, these concepts extend directly to the multi-outcome case. See, also, Banker et al. (2001) for statistically consistent estimation of general monotone and concave or convex functional relationships.

4. Statistical tests

In this section, we examine two parametric and three non-parametric tests for comparing two groups of DMUs, when the error term is composed of both inefficiency and noise.

4.1. Parametric tests

We first evaluate an OLS-based regression and a t-test for comparing means of inefficiency.

4.1.1. OLS regression

The first statistical test we propose is based on the OLS regression of the inefficiency on a dummy variable \( z \) that takes a value of 1 if a particular DMU belongs to group 1 and 0 otherwise. Cummins et al. (1999) have applied this test for comparing two groups of DMUs, without considering the error term. For the purpose of this regression we define \( \beta_0 = V_{\text{MAX}} - (U_1 - 1) \) and \( \beta_1 = (U_2 - U_1) - (V - \hat{u}_{jm}) \). Since \( \hat{u}_{jm} = (V_{\text{MAX}} - \hat{e}_{jm}) \), we can write

\[
\hat{u}_{jm} = \beta_0 + \beta_1 z + \delta_{jm},
\]

(8)

where \( \beta_0 = \hat{u}_2 - \hat{u}_1 \). Evidently, the error term \( \delta_{jm} \) in (8) has a zero mean and a finite variance. Further, if \( \sigma_1^2 = \sigma_2^2 = \sigma_z^2 \), i.e., the variances of the inefficiency random variables for the two groups are equal,\(^4\) then the variance of the error term \( \delta_{jm} \) is the same for both groups and is equal to \( \sigma_z^2 + \sigma_\gamma^2 \). Therefore, if \( \hat{u}_{jm} \) is known, OLS estimation of (8) using the entire sample of \( N \) observations yields consistent estimators of the intercept \( \beta_0 = V_{\text{MAX}} - (U_1 - 1) \) and slope \( \beta_1 = \hat{u}_2 - \hat{u}_1 \) (Schmidt, 1976).

Since the true inefficiency value is not known, we replace \( \hat{u}_{jm} \), the dependent variable in (8), with the corresponding DEA estimator, \( \hat{u}_m \). It is intuitive that as far as evaluating the difference in mean inefficiencies is concerned, the use of \( \hat{u}_{jm} \) instead of the true \( u_{jm} \) preserves consistency. We prove that this is indeed the case. This leads to our first proposition.

**Proposition 1.** Assume the variances of the inefficiency random variables for the two groups are equal. The OLS estimators \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) of \( \beta_0 \) and \( \beta_1 \), respectively, in \( \hat{u}_{jm} = \hat{\beta}_0 + \hat{\beta}_1 z + \hat{\delta}_{jm} \) are consistent estimators of \( V_{\text{MAX}} - (U_1 - 1) \) and \( \hat{u}_2 - \hat{u}_1 \), the difference in mean inefficiency between groups 1 and 2. Further \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are normally distributed asymptotically with means \( V_{\text{MAX}} - (U_1 - 1) \) and \( \hat{u}_2 - \hat{u}_1 \), and variances \( \left( \sigma_z^2 + \sigma_\gamma^2 \right)/N_p \) and \( \left( \sigma_z^2 + \sigma_\gamma^2 \right)/N_p \), respectively, where \( p_1 \) and \( p_2 \) are the fractions of group 1 and group 2 DMUs in the population.

\(^5\) Note that this ensures both groups (of DMUs) face the same frontier. Therefore, the tests developed in the next section are not readily applicable to the situations when different groups face different frontiers.

\(^4\) This assumption holds under the null hypothesis where we assume the two groups are statistically equal.

4.1.2. t-test

Since the OLS regression yields an estimator of the mean inefficiencies when the logarithm of the inefficiency is regressed on a group dummy variable, it is logical to expect that a two-sample test such as the two-sample Student’s t-test may be appropriate in comparing the mean inefficiencies across the two groups. It is, however, not possible to use the standard two-sample t-test directly to compare the mean inefficiencies because \( \hat{u}_{jm} = (V_{\text{MAX}} - \hat{e}_{jm}) \), by construction, takes values in the range \((0, \infty)\) and the two-sample t-test requires that the random variables compared follow a (two-sided) normal distribution.

We assume that the random variables \( \hat{u}_{jm}, m = 1, 2 \) follow a log-normal distribution with parameters \( \eta_m \) and \( \omega_m \). Recall that we had earlier assumed that the variances of the inefficiency random variables \( u_1 \) and \( u_2 \) are equal, i.e., \( \sigma_1^2 = \sigma_2^2 = \sigma_\gamma^2 \) in the design of the OLS-based test; we retain that assumption.

**Lemma 1.** When the mean inefficiencies of the two groups are equal i.e., \( \bar{u}_1 = \bar{u}_2 \), the random variables \( \ln(\hat{u}_1) \) and \( \ln(\hat{u}_2) \) are i.i.d. normal with mean \( \eta = \eta_1 = \eta_2 \) and standard deviation \( \omega = \omega_1 = \omega_2 \).

Given Lemma 1, a test of the equality of the means of \( \ln(\hat{u}_1) \) and \( \ln(\hat{u}_2) \) is equivalent to a test of the equality of the mean inefficiencies i.e., a test of \( \bar{u}_1 = \bar{u}_2 \). The equality of the means of \( \ln(\hat{u}_1) \) and \( \ln(\hat{u}_2) \) can be tested using the Student’s t-test since \( \ln(\hat{u}_1) \) and \( \ln(\hat{u}_2) \) are i.i.d. Normal under the null hypothesis that \( \bar{u}_1 = \bar{u}_2 \) and the variances of \( \ln(\hat{u}_1) \) and \( \ln(\hat{u}_2) \) although unknown, are equal. The test statistic which is distributed as a Student’s t-variate with \( N_1 + N_2 - 2 \) degrees of freedom is calculated as

\[
\bar{t} = \frac{T_1 - T_2}{S/\sqrt{N_1 + N_2}},
\]

(9)

where

\[
T_1 = \frac{1}{N_1} \sum_{j=1}^{N_1} \ln(\hat{u}_1), \quad T_2 = \frac{1}{N_2} \sum_{j=1}^{N_2} \ln(\hat{u}_2), \quad S = \sqrt{\frac{1}{N_1 + N_2 - 2} \left( \sum_{j=1}^{N_1} \ln(\hat{u}_1) - T_1^2 + \sum_{j=1}^{N_2} \ln(\hat{u}_2) - T_2^2 \right)^{0.5}}.
\]

(10)

Since \( \hat{u}_{jm} \) is estimated using DEA, the DEA estimator \( \hat{u}_{jm}, m = 1, 2 \) is used in place of \( \hat{u}_{jm} \). We denote the resulting approximations as \( \bar{t}, \bar{T}_1, \bar{T}_2 \) and \( S \), respectively. Specifically

\[
\bar{t} = \frac{\bar{T}_1 - \bar{T}_2}{\bar{S}/\sqrt{\bar{N}_1 + \bar{N}_2}}.
\]

(11)

Since the bias in the DEA estimators approaches zero for large samples it is reasonable to expect that statistical tests based on \( \bar{t} \) is appropriate for evaluating the difference in mean inefficiencies across groups, provided the sample size is large. We verify that this is indeed true.

**Proposition 2.** For large samples, the distribution of \( \bar{t} \) approaches that of \( t \), which is distributed as a Student’s t-variate with \( N_1 + N_2 - 2 \) degrees of freedom. Further, the difference between the mean inefficiencies of groups 1 and 2 can be consistently estimated as

\[
\exp(0.5t^2) \times \left\{ \exp(\bar{T}_1) - \exp(\bar{T}_2) \right\}.
\]

Consider the special case when the distribution of the noise random variable collapses to a point located at the origin. It is obvious that the regression-based tests as well as the two-sample t-test are applicable in this special case to compare the mean inefficiencies across the two groups.

Please cite this article in press as: Banker, R.D., et al. DEA-based hypothesis tests for comparing two groups of decision making units. European Journal of Operational Research (2010), doi:10.1016/j.ejor.2010.01.027
4.2. Non-parametric tests

Non-parametric tests have been commonly used in the DEA literature (Grosskopf and Valdmanis, 1987; Golany and Storberg, 1999; Hahn and Kauffman, 2002). Most of these tests are based on order statistics, including the three commonly used tests – the median test, the Mann–Whitney test and the Kolmogorov–Smirnov test. However, in the presence of noise, it is unclear whether the order of DEA inefficiency estimates remains consistent with that of true inefficiencies. Below we prove that all the three tests considered in this study are consistent and discuss the asymptotic properties of these tests.

4.2.1. The median test

The median test (Gibbons 1985) is used for testing the null hypothesis that there is no difference between the medians of two groups of DMUs. Denote $M_1$ and $M_2$ as the median inefficiency of $u_1$ and $u_2$ for group 1 and group 2, respectively. Let $M$ be the median of the combined sample of the two groups. Then the null hypothesis is $H_0: M_1 = M_2 = M$. It is easy to verify (see the proof in the Appendix for Lemma 2) that the corresponding estimates $\hat{M}_1$, $\hat{M}_2$ and $M$ for $\hat{u}_{\text{in}}(m = 1, 2)$ are also consistent.

**Lemma 2.** $\hat{M}_1$, $\hat{M}_2$ and $M$ derived from DEA estimated $\hat{u}_{\text{in}}(m = 1, 2)$ are consistent estimators.

To develop the median test, we introduce some additional notation. We denote $n_1$ and $n_2$ as the number of times that $u_1$ and $u_2$ are less than $M$, respectively, and $P_1$ and $P_2$ as the probabilities that $u_1$ and $u_2$ are less than $M$, respectively, i.e., $P_1 = P(\hat{u}_1 < M)$ and $P_2 = P(\hat{u}_2 < M)$. It is obvious that the null hypothesis $M_1 = M_2$ is equivalent to $P_1 = P_2$. According to Gibbons (1985, p. 132), the random variables $n_1/N$ and $n_2/N$ are consistent point estimates of the parameters $P_1$ and $P_2$, that is, $\hat{P}_1 = n_1/N$ and $\hat{P}_2 = n_2/N$. For notational simplicity, let the weighted probability be $P = n_1P_1N_1 + n_2P_2N_2/N_1 + N_2$ and $\hat{P} = n_1\hat{P}_1 + n_2\hat{P}_2/N_1 + N_2$. We have

**Proposition 3.** When the median inefficiencies of the two groups are equal ($P_1 = P_2$), the random variables $n_1/N$ and $n_2/N$ are consistent estimators for $P_1$ and $P_2$. Additionally, $P_1 - P_2$ is asymptotically a standard normal distribution with mean $N_1N_2/P_1N_1N_2 + P_2N_2/N_1 + N_2$, and standard deviation $\sqrt{P(1 - P)(1/N_1 + 1/N_2)}$. Furthermore, the median test is a standard normal Z-test where $Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1 - \hat{P})(1/N_1 + 1/N_2)}}$.

The Z-test can also be viewed as the following: when $N_1$ and $N_2$ are relatively large, $P_1 - P_2$ is normally distributed with mean $P_1 - P_2$ and variance $\hat{P}(1 - \hat{P})(1/N_1 + 1/N_2)$. Its asymptotic distribution is also standard normal as stated in Lemma 3.

**Lemma 3.** Asymptotically, $\hat{Z}$ follows a standard normal distribution.

4.2.2. Mann–Whitney’s U-test

Mann–Whitney’s U-test assesses whether one of two random variables is stochastically larger than the other one (Mann and Whitney, 1947). For the problem considered here, the Mann–Whitney statistic is defined as the number of times $u_1$ precedes $u_2$ in the combined and ordered sample of the two groups $i = 1 \ldots N_1$ and $j = 1 \ldots N_2$. Define a random variable

$$\hat{D}_j = \begin{cases} 1 & \text{if } u_1 < u_j, \\ 0 & \text{otherwise}. \end{cases}$$

Then Mann–Whitney’s U statistic is

$$\hat{U} = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \hat{D}_j.$$  

Mann and Whitney (1947) show that for large samples of $N_1$ and $N_2$ (for $N_1$ and $N_2$ as small as 6), $U$ is normally distributed with mean $N_1N_2/2$ and variance $N_1N_2(N+1)/12$, where $N = N_1 + N_2$. Thus the large-sample test statistics is

$$\hat{Z} = \frac{\hat{U} - N_1N_2/2}{\sqrt{N_1N_2(N+1)/12}}$$

whose distribution is approximately standard normal. We prove that this statistic remains consistent when $\hat{u}_{\text{in}} = \sqrt{\text{MAX}} - v + u_{\text{in}}$ incorporates noise.

**Proposition 4.** When the inefficiency variables $u_{\text{in}}(m = 1, 2)$ and the noise variable $v$ are independent, the order of DEA estimate $\hat{u}_{\text{in}}$ remains consistent with the true order of $u_{\text{in}}$, and as a result, the Mann–Whitney’s U test remains consistent.

Further it is also straightforward to verify that the asymptotic distribution of $\hat{Z}$ is also standard normal as follows:

$$\text{plim } \hat{Z} = \text{plim } \frac{\hat{U} - N_1N_2/2}{\sqrt{N_1N_2(N+1)/12}} = \frac{\text{plim } \hat{U} - N_1N_2/2}{\sqrt{N_1N_2(N+1)/12}} = \frac{U - N_1N_2/2}{\sqrt{N_1N_2(N+1)/12}} = Z.$$  

4.2.3. The Kolmogorov–Smirnov test

Kolmogorov–Smirnov’s test statistic computes the maximum vertical distance between $F(\hat{u}_1)$ and $F(\hat{u}_2)$, the empirical distributions of $\hat{u}$ for groups 1 and 2, respectively. This statistic, by construction, takes values between 0 and 1 and a high value of this statistic is indicative of significant differences in inefficiency between the two groups. Readers are referred to Gibbons (1985, p. 127) for further details.

The Kolmogorov–Smirnov test, denoted as $\hat{D}$, is also derived from $u_{\text{in}}$. From the proof of Proposition 4, we know the order of $\hat{u}_{\text{in}}$ is consistent with that of $u_{\text{in}}$. Thus $\hat{D}$ is also consistent with the true $D$ estimated from the ranked $u_{\text{in}}$. Asymptotically, for any real value $\hat{Z}$, $\hat{D}$ is distributed as follows (Smirnov, 1939):

$$\lim_{N_1N_2 \to \infty} P(\frac{N_1N_2}{N_1 + N_2} \leq \hat{D}) = 1 - e^{-2\hat{Z}^2}.$$  

4.3. The Banker’s F-test

In addition, the two F-tests developed by Banker (1993) are also applicable. Next, we briefly describe the rationale of the tests. Suppose both $u_1$ and $u_2$ are distributed as half-normal over the range $(0, \infty)$ with parameters $\rho_1$ and $\rho_2$, respectively. Then under the null hypothesis that there is no difference between the two groups (i.e., $\rho_1 = \rho_2$), the test statistic is calculated as $\sum_{i=1}^{N_1}(\hat{u}_{i1}/N_1)/\sum_{j=1}^{N_2}(\hat{u}_{j2}/N_2)$ and is distributed as $F$ with $(N_1, N_2)$ degrees of freedom.

If both $u_1$ and $u_2$ are distributed as exponential over $(0, \infty)$ with parameters $\lambda_1$ and $\lambda_2$, respectively, then under the null hypothesis that $\lambda_1 = \lambda_2$ (i.e., there is no difference between the two groups), the test statistic is calculated as $\sum_{i=1}^{N_1}(\hat{u}_{i1}/N_1)/\sum_{j=1}^{N_2}(\hat{u}_{j2}/N_2)$ and evaluated relative to the critical value of the $F$ distribution with $(2N_1, 2N_2)$ degrees of freedom.

Assuming a data generating process that characterizes the deviation of the actual output from the best practice frontier as arising from a single inefficiency term, Banker and Chang (1995) evaluated the performance of the Banker (1993) test against the Welch T-test and Mann–Whitney tests that had been used traditionally in the DEA literature. They found that the Banker (1993) tests, in general, performed significantly better than the traditional tests. However, it is an open question whether the Banker (1993) test possesses enough power when noise is an integral part of the data generating process.
process. We present simulation results that yield insights into the performance of the various tests as a function of the noise in the environment.

5. Simulation design

We use simulation to assess the different hypothesis-testing methods considered in this paper. We adopt type I error as the primary criterion to evaluate the tests since all these tests are based on the premise that the two groups follow the same inefficiency distribution. We further use type II error as an ancillary criterion to shed light on the comparative power across the six tests considered. The design of simulation largely follows Banker and Chang (1995), Banker and Natarajan (2008). We specify the detailed procedure of the DGP below.

5.1. Choice of production technology

We consider the following cubic polynomial function as Banker and Natarajan (2008):

\[ y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3. \tag{17} \]

The input variable \( x \) is generated from the uniform distribution over the interval \([1, 4] \). The coefficients \( \beta_0, \beta_1, \beta_2, \beta_3 \) determine the properties of the production of technology and we use the following values: \( \beta_0 = -37, \beta_1 = 48, \beta_2 = -12, \beta_3 = 1 \). The choice of values for the parameters characterizing the composed error distribution is based on guidance from prior literature (Olson et al., 1980; Banker, 1993; Banker et al., 2004; Banker and Natarajan, 2008) that have used DGP involving composed error distribution.

Note that the above choices guarantee that the production function is a monotonically increasing and concave function when \( x \in [1, 4] \). We adopt this production function for the two groups (of DMUs) considered in this paper. Adding the error term, the complete production functions are:

\[ y_{1i} = -37 + 48x_{1i} - 12x_{1i}^2 + x_{1i}^3 + \nu_i - u_{1i} \quad \text{if } i \in \text{group 1}, \]
\[ y_{2i} = -37 + 48x_{2i} - 12x_{2i}^2 + x_{2i}^3 + \nu_i - u_{2i} \quad \text{if } i \in \text{group 2}. \tag{18} \]

5.2. Choice of the inefficiency distribution for \( u_{ai} \) and the random noise variable \( \nu \)

We follow Banker and Chang (1995), Banker and Natarajan (2008) and specify the inefficiency distributions as a half-normal distribution. In particular, we generate two different panels of data to compute the Type I and Type II errors separately. The first panel (the main panel) is designed to detect type I error which is the rate at which a test accepts H0 when it’s true. Here, we generate the inefficiencies for both groups from the same half-normal distribution \( N(0, 1) \), with a mean of 0.7978. The second panel is intended to detect type II error, which is the rate at which a test accepts H0 when it is untrue. In this case, the inefficiencies of the two groups are generated from different distributions. The first group is still generated from \( |N(0, 1)| \), but the second group is generated from \( |N(0, 0.5)| \) with a mean inefficiency of 0.3989.

We assume that the noise \( \nu \) (of both groups) comes from a normal distribution \( N(0, \sigma^2_\nu) \). We generate the noise variable, \( \nu \), from a two-side truncated normal distribution with a mean of zero and upper and lower bounds of \( 6\sigma_\nu \) and \( -6\sigma_\nu \), respectively. Note that truncating at six sigma results in a less than 0.001% potential loss of data. We set \( V_{max} \) (introduced in Eq. (4)) to be \( 6\sigma_\nu \) depending on the actual value of \( \sigma_\nu \).

For each test, we vary the noise level \( \sigma_\nu \) in order to examine its effect on the performance of individual tests. Specifically, we define the noise ratio \( r = \sigma_\nu / \sigma_{u_i} \), to reflect the relative magnitude of the noise to inefficiency. The ratio \( r \) takes four values: 0, 0.25, 1 and 5.

5.3. Sample size

We generate the two groups of DMUs following the designs specified above. For simplicity (and without loss of generality), we generate equal-sized samples for the two groups. Three levels – 30, 100 and 500 – are used for sample size, to represent relatively small, medium and large sizes.

The above three steps complete the DGP process. A sample of 100 generated DMUs is plotted in Fig. 1. The \( x \) value is a random draw from the uniform distribution between \([1, 4]|\). The \( y \) value is derived from the specified production function and is scaled down by taking the natural logarithm. The middle line represents the theoretical production function when no inefficiency or noise terms are considered. Each dot represents a sample datum generated through the DGP specified above. Fig. 1 clearly indicates that two-sided noise exists in the data.

5.4. Evaluation criteria

In each iteration, after the two groups of DMUs are generated, we apply the DEA BCC formulation (as specified in Eq. (7)) to estimate the inefficiencies for both groups. Then, we apply the six statistical tests at the 0.05 significance level to compare the two groups. To summarize, the tests are: (1) the \( F \)-test for means, (2) the OLS regression, (3) the Median Z-test, (4) the Mann–Whitney U test, (5) the Kolmogorov–Smirnov D test, and (6) the Banker’s \( F \)-test.

A thousand iterations are used to smooth out randomness. We then evaluate the six tests by computing the rate of the type I and Type II errors for all the tests.

6. Experimental results

The results are organized to answer the following three questions of interest:

1. How do the six tests compare with each other with respect to different sample sizes?
2. What’s the effect of noise? In particular, how does the noise affect the comparison between the tests developed in this paper and the test developed in Banker (1993)?
3. How closely does the asymptotic distribution of each test approach its theoretical one?

![Fig. 1. Hundred simulated DMUs based on the specified DGP.](Image 317x55 to 557x188)
6.1. The effect of sample size

Table 1 summarizes the results as a function of sample size. To reduce the number of possible combinations, we fix the noise ratio to 0.25 which represents a moderate noise level. Each value in Table 1 represents the corresponding error rate under each scenario and is computed on the basis of 1000 simulations.

Observe that Type I error has a tight range from 4.3% to 8.0% across the six tests and three different levels of sample sizes, with the exception of the Banker’s F-test at size 30 (where the error rate is 18.2%). This indicates that the tests perform well in general (compared to the benchmarking 0.05 p-value). The two parametric tests perform relatively better than the three non-parametric ones for sizes 30 and 100. When the sample size is large (500), the Mann–Whitney test stands out. However there is no uniform pattern across different sizes. It is a little surprising that the two parametric tests perform better with smaller sample sizes (e.g., a size of 30) and worse at larger sizes (e.g., 500). This could be a result of the standard error in the simulated samples. As a rule of thumb, larger samples tend to have smaller standard errors and thus a small difference (across the two samples of DMUs) is more likely to be significant. Thus it is more likely for larger sample size to exhibit type I error. Our simulation results show that the performances of the two parametric tests deteriorate as the sample size increases. The three non-parametric tests are more invariant to sample sizes as shown in Table 1. The Banker’s F-test turns out to only work well with larger sample sizes. The error rate at size 30 reaches 18.2% but the performance improves dramatically as the size increases (for size 500 it reaches 4.5%). This is consistent with Banker (1993, p. 1273) which states that large sample size is required for the F-test to asymptotically approach the theoretical distribution.

For Type II error, all the six tests perform better as the sample size increases. For the same reason discussed above, a difference (of two samples of DMUs) is more likely to be detected as significant when sample size increases. And thus the type II error rate is expected to decrease with large sizes. Table 1 also shows that the three non-parametric tests exhibit higher Type II errors than the two parametric tests, with the Kolmogrov–Smirnov test having the highest error rate.

6.2. The effect of noise

We vary the noise ratios from low to high in order to examine the effect of noise on the comparative performance across the tests. Four levels – 0, 0.25, 0.5 and 1 – are used. We fix the sample size to be 100, representing a normal size. All the experiments are repeated over 1000 iterations. Table 2 summarizes the results.

Overall, for Type I error, as the noise level increases, all the tests perform worse. However, Banker’s F-test is more sensitive to noise. When there is no noise (r = 0), the F-test only commits an error rate of 3.1%. But its performance deteriorates drastically as the noise ratio goes higher. With a noise ratio 5, the F-test yields a staggering error rate of 35.6%. The other five tests are more robust to noise. The two parametric tests in particular are less sensitive to noise (with a moderate sample size of 100).

Table 2 also shows that noise has a bigger impact on Type II error. When the inefficiency dominates noise (noise ratio < 1), all tests except the Kolmogrov–Smirnov test show sufficient test power (which is defined as 1-Type II error). However, when noise dominates inefficiency (noise ratio > 1), the powers of all the tests decrease dramatically. The two parametric tests are found to have lower Type II rates than the non-parametric tests.

6.3. Asymptotic distributions

In Section 4, we derived the asymptotic distributions for the hypothesis testing methods considered. An interesting question is how well the distribution of each statistic approaches the theoretical one. This can be empirically verified as follows. We first simulate two groups of DMUs following the DGP described in Section 5. Then, we apply all six tests to the resulting data and derive the statistics for each test. We repeat the same procedure for 1000 iterations with a relatively large sample size of 500 to evaluate the asymptotic properties of the tests. After computing 1000 statistics for each of the six tests, we compare its distribution with the theoretical one. A sample of the theoretical data, simulated from the corresponding theoretical distribution, is used as a benchmark. For example, t is asymptotically distributed as t. We thus generate 1000 data points according to the theoretical t-distribution. Finally, we check whether the two samples – one simulated and one from the theoretical distribution – indeed follow the same distribution, for all six tests.

For the purpose of comparing two distributions, we apply the quantile-quantile (Q–Q) plot technique (Chambers et al., 1983; Johnson and Wichern, 1998), a graphical technique for determining if two data sets come from populations with a common distribution. If the underlying distributions from two datasets are the same, then the Q–Q plots should be straight lines. We aim to test whether the underlying distributions of our tests differ significantly from their theoretical ones.

Fig. 2 presents the six Q–Q plots, with the theoretical distribution on the x-axis and the empirical ones on the y-axis. The straightness of the curves reflects the closeness of the two distributions. The correlation coefficient test is often used to measure the straightness of the Q–Q plot (Johnson and Wichern, 1998, p. 193). We computed the correlation coefficients based on 50 random data points sampled from the Q–Q plot. The correlation coefficients are 0.999, 0.999, 0.996, 0.996, 0.993 and 0.996 for the Student’s t-test, OLS, the median test, Mann–Whitney’s U test, etc.
the Kolmogorov–Smirnov test and the Banker’s F test, respectively. None of these are significant at 0.05 level (the critical value is $R^{*}_{0.05} = 0.977$). This indicates that there is no significant difference between the distributions governing the simulated data and their corresponding theoretical distributions.

6.4. Robustness check

In this section, we conduct several additional experiments to check whether our findings are robust to other specifications of DGP, in particular with different production functions, different inefficiency and noise distributions.

We first examine one additional production function – the Cobb-Douglas production function with one input and one output – in the form of $y = (x - \beta)^a$ with $x$ as the input and $y$ as the output. We follow Banker and Chang’s (2006) DGP for the specifications of $x$ and the two parameters $\alpha$ and $\beta$. We generate the input $x$ randomly from an independent uniform distribution over the interval $[10,20]$. We set $\alpha = 5$ and the coefficients $\beta$ are generated from an independent uniform distribution over the interval $[0.4,0.5]$. Since $\beta$ is less than one, the production function satisfies the BCC model’s assumption of a concave production function. The inefficiency for each observation is generated from a half-normal distribution $\mathcal{N}(0,1)$.

We fix the sample size to be 100 and the noise ratio to be 0.25, representing a normal size and a moderate noise level, and examine type I error. The result is reported in Table 3 under the row “Cobb-Douglas Production Function”. All six tests perform reasonably well (i.e. all Type I error rates are smaller than, but not far from 5%). And their performances are comparable within a tight range from 3.0% to 4.9.

We then examine a different inefficiency distribution – an exponential distribution with parameter 0.7978 – following Banker and Chang (1995). The other DGP remains the same as what was specified for detecting Type I error in Section 5. As before, the sample size is 100 and the noise ratio is 0.25. Note that the Banker’s F-test

### Table 2

<table>
<thead>
<tr>
<th>Noise ratio</th>
<th>T-test for mean</th>
<th>OLS (%)</th>
<th>Median test (%)</th>
<th>Mann–Whitney test (%)</th>
<th>Kolmogrov–Smirnov test (%)</th>
<th>Banker F-test (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>0</td>
<td>4.8</td>
<td>4.7</td>
<td>5.9</td>
<td>4.1</td>
<td>6.7</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>5.1</td>
<td>4.9</td>
<td>6.7</td>
<td>5.9</td>
<td>6.7</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5.2</td>
<td>5.0</td>
<td>6.9</td>
<td>5.5</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>5.5</td>
<td>5.3</td>
<td>7.9</td>
<td>6.5</td>
<td>8.7</td>
</tr>
<tr>
<td>Type II</td>
<td>0</td>
<td>2.1</td>
<td>2.1</td>
<td>9.7</td>
<td>6.0</td>
<td>19.3</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>4.3</td>
<td>4.3</td>
<td>10.6</td>
<td>6.7</td>
<td>24.3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>14.0</td>
<td>14.0</td>
<td>22.1</td>
<td>12.0</td>
<td>26.4</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>18.2</td>
<td>18.2</td>
<td>28.6</td>
<td>18.7</td>
<td>32.0</td>
</tr>
</tbody>
</table>

Fig. 2. Q–Q plots for the six distributions.
statistic under an exponential distribution becomes $\sum_{j=1}^{N_2}(u_j^2/N_2)/\sum_{j=1}^{N_2}(u_j^2/N_2)$ with $(2N_1, 2N_2)$ degrees of freedom (Banker, 1993). The results, reported under the row “Exponential Inefficiency Distribution”, remain consistent in that the two parametric tests outperform the non-parametric tests, though all six tests perform slightly worse than in the case of the half-normal distribution (see Table 1 at size 100).

Lastly, we examine the case when both inefficiency and noise are exponentially distributed. The DGP is the same as in the case above, except that the noise is distributed as an exponential distribution with parameter 0.1995 ($=0.25 \times 0.7978$). The result is shown in the last row of Table 3. The three non-parametric tests appear to be less sensitive while the Banker’s $F$-test is affected the most. This is consistent with our prior findings.

7. Conclusion

We provide a statistical basis for using DEA-based estimators to compare differences in the inefficiency distributions for two groups of decision making units (DMUs). Our estimation methods and statistical tests are developed under a data generating process (DGP) that models the deviation of the output from the best practice frontier as the sum of two components, a one-sided inefficiency term and a two-sided random noise term. The estimation procedures and tests proposed here are simple, intuitive and exploit the fact that production functions fitted by DEA asymptotically retrieve the true production functions.

Our simulation results confirm that DEA estimators can help detect efficiency differences across groups of DMUs. The simulation experiments also help us evaluate the performance of the tests developed in this paper against those developed by Banker (1993). We find that tests developed in this paper perform better than the tests in Banker (1993) when noise plays a significant role in the data generating process. On the other hand, the Banker tests are effective when efficiency dominates noise. All six tests considered in this paper yield robust asymptotic properties — their empirical distributions approach the theoretical ones. Overall, this paper adds to the sparse but growing literature that has sought to provide statistical rigor to DEA-based estimation procedures. It also provides a theoretical base for the use of DEA in the test of multiple groups of DMUs.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2010.01.027.

References


ARTICLE IN PRESS