Selection Procedures for Order Statistics in Empirical Economic Studies

William C. Horrace*
Department of Economics
University of Arizona
P.O. Box 210108
Tucson, AZ 85721-0108
Voice: (520) 621-6230
Fax: (520) 621-8450
whorrace@u.arizona.edu

August 2000

Abstract: In a presentation to the American Economics Association, McCloskey (1998) argued that "statistical significance is bankrupt" and that economists' time would be "better spent on finding out How Big Is Big". This brief survey is devoted to methods of determining "How Big Is Big". It is concerned with a rich body of literature called *selection procedures*, which are statistical methods that allow inference on order statistics and which enable empiricists to attach confidence levels to statements about the relative magnitudes of population parameters (i.e. How Big Is Big). Despite their prolonged existence and common use in other fields, selection procedures have gone relatively unnoticed in the fields of economics, and, perhaps, their use is long overdue. The purpose of this paper is to provide a brief survey of selection procedures as an introduction to economists and econometricians and to illustrate their use in economics by discussing a few potential applications. Both simulated and empirical examples are provided.

* Thanks to Gordon Tullock for providing fodder for this paper and for some excellent comments and discussions. All errors and omissions are my own.

JEL Calssification: C10, C12
Keywords: ranking and selection, order statistics, statistical inference.

1. Introduction

In her presentation to the American Economics Association, Deirdre McCloskey (1998) argued that "statistical significance is bankrupt" and that economists' time would be "better spent on finding out How Big Is Big".[1] Absent a philosophical debate on the merits of her arguments, one cannot deny that empirical economic research is often concerned with comparing the size of parameters across various populations or studies. For instance, economists are concerned with relative growth rates across countries, wage discrimination across industries, and technical efficiency across production units. Indeed, the proliferation of panel data sets and the econometrics of panel data, have provided applied economists with an arsenal of tools to perform these types of comparative studies easily for large numbers of populations.

For example, Seale (1990) uses panel data and a fixed effect regression specification to estimate and rank the technical efficiency of twenty-five Egyptian tile manufacturers. Interest centers on determining which tileries are most efficient. Fields and Wolff (1995) use the Current Population Survey to estimate a cross sectional log-wage equation. Parameter estimates are used to calculate and rank the gender wage gap across various industry classifications. Haurin (1989) uses the National Longitudinal Survey to estimate a women's leisure demand equation to determine the dynamic effects of disruptions in spousal income. The four disruptions explored include: "death of spouse", "divorce/separation from spouse", "spouse becomes unemployed" and "spousal health worsens". It is determined that "divorce/separation from spouse" has the largest

[1] My apologies to Dr. McCloskey for quoting her out of context. What McCloskey was actually arguing is that all frequentist notions of hypothesis testing are "bankrupt", and they are bankrupt in the sense that given a large enough sample size, any parameter can be found statistically significant. In the end, the inferences presented here are based on these same frequentist notions that McCloskey debunks, and they

effect on a woman's leisure demand. Mowery (1983) compares the survival rates of three classes of firms (large, small, and large with research facilities) during the 1930's. He estimates and ranks several coefficients that determine a firm's probability of survival.

In these examples an implied goal of the research may be to make statements about the relative magnitude of the populations, such as "firm *j* is the most efficient," or "industry *n* is more discriminatory than industry *q*," or "spousal disruptions *a* and *b* have the smallest effect on leisure demand". From a statistical standpoint, however, these statements are meaningless without an associated confidence level. For example, it is only meaningful to assert that "industry *r* has the smallest wage gap *with 95% confidence*". There is a body of statistical literature called *ranking and selection procedures* devoted to these types of inferences. The purpose of this paper is to describe some of these procedures that may be of use to economists.[2]

It is important to distinguish at the outset the difference between *ranking* procedures and *selection* procedures; the latter are likely to be more relevant to economists and are consequently the focus of this paper. Suppose that there are N populations, and population i has parameter value α_i, and that there are T observations from each population so that we can calculate unbiased sample estimates for each parameter, $\hat{\alpha}_i$. Let the population ordering of the parameters be $\alpha_{[N]} \geq \alpha_{[N-1]} \geq \ldots \geq \alpha_{[1]}$, and let $\hat{\alpha}_{(i)}$ be the sample mean from the population with mean $\alpha_{[i]}$. That is, $\alpha_{[N]} > \alpha_{[N-1]}$ does not necessarily imply that $\hat{\alpha}_{(N)} > \hat{\alpha}_{(N-1)}$ due to sampling variation, so that the

---

are not strictly immune to her sample size criticism. However, the coinage "How Big Is Big" is just too tempting to ignore.

[2] Per Dudewicz and Koo (1982), as of 1982 there were 1188 known publications, theses and technical reports on the subject of ranking and selection. Of these only one, Burdick et al. (1967), concerned the field of economics.

ordering in the sample may not correspond to the ordering in the population. *Ranking procedures* are techniques for controlling the probability that the ordering of the sample estimates is indeed the ordering of the population parameters. These are typically employed in the "design of experiments" literature to allow experimenters to *ex ante* select experimental sample sizes which ensure that a pre-selected probability of a correct sample ordering is attained. For example, if a bio-statistician is concerned with ranking three drug treatments in terms of their relative effectiveness in combating a particular disease, she administers the drug treatments to a sample of patients, measures their performance on each patient, and calculates a sample average of performance for each treatment. Based on these sample averages, she ranks the treatments as "good", "better" and "best". Ranking procedures enable the bio-statistician to pre-select appropriate sample sizes (numbers of patients to test) to ensure that her sample ranking is true at a predetermined confidence level. In economics, where "experiments" are typically not designed, the use of ranking procedures seems dubious. As such, they will not be addressed.[3]

Selection procedures, on the other hand, are techniques for identifying a select group of populations with the largest (smallest) population parameters at a pre-specified confidence level. To continue our example, suppose that our bio-statistician forgot to perform an ex ante ranking procedure and was faced with the *ex post* results of her experiment: "good", "better", "best". Suppose further that she recognizes the limitation of her experiment; she knows that even though "Treatment A" is the best in the sample, perhaps it is not the best in truth (in the population). Selection procedures allow her to

[3] My apologies to Vernon Smith and the Economic Science Lab at the University of Arizona, where economic experimental design is alive and well.

attach confidence levels to her ranking results such as, “Treatment A is best with 80% confidence” or “with 95% confidence Treatments A and C are best and Treatment B is not”. Complete details of these procedures are discussed in the following section, but, suffice it to say that an economist might use these techniques to make statements such as "the countries with the slowest growth rate is either *j* or *r* with probability 0.95" or "the most efficient firm is firm *j* with probability 0.85." That is, an economist might use selection procedures to determine “How Big Is Big”.

Due to the nature of order statistics, both ranking and selection procedures are necessarily simultaneous multivariate inference procedures. That is, to make statements about the relative size of population parameters (as these procedures do), it is necessary to *simultaneously compare* each parameter to *each* of the other parameters while controlling for the overall confidence level of the statement. That is, for the bio-statistician to know that Treatment A is the best, she must know *simultaneously* that Treatment A is better than Treatment B *and* that it is also better then Treatment C. To make joint probability statements one could use the Bonferroni inequality, but it becomes too conservative after only a few simultaneous statements. The Bonferroni inequality might be appropriate for our bio-statistician and her two simultaneous statements, but it would be much too conservative for an economist who is interested in ranking the efficiencies of 171 Indonesian rice farms or the growth rates of 50 U.S. states. Due to this implicit simultaneity, ranking and selection procedures are closely related to a branch of multivariate decision theory called *multiple comparison procedures.*

Multiple comparison procedures are techniques for constructing simultaneous confidence intervals for differences *between* population parameters and have recently

been used to perform inference in economic applications. See Horrace and Schmidt (1996, 1999) and Horrace (1999). For instance, Horrace and Schmidt (1999) employ a multiple comparison procedure called *multiple comparisons with the best* (MCB) to construct simultaneous confidence intervals on parameters, $\alpha_{[N]} - \alpha_i$, $i = 1, ..., N$, and to select the populations with the largest α_i at a pre-specified confidence level. Horrace and Schmidt (1999) provide a detailed survey and a few new theorems for MCB. Horrace (1999) uses MCB to uncover the ranking uncertainty of an order statistic of labor market wage gaps across various industry classifications. All of these papers are concerned with performing tests of "bigness" and indirectly use some of the techniques outlined herein.

This paper is a brief survey of selection procedures intended for an audience of applied economists to encourage their study and application. The next two sections provide the survey. Section 4 illustrates these procedures with two brief (but informative) empirical examples. Section 5 provides the summary and conclusions.

2. Selection Procedures: The Independent Case

2.1 Overview

Selection procedures are usually traced back to Bechhofer (1954). However, because they are only one approach in multiple decision theory, their roots are often attributed to Abraham Wald in the 1940's. A comprehensive, early survey is provided in Dudewicz and Koo (1982), and an excellent textbook treatment with extensive references is given in Gupta and Panchapakesan (1979). Much of the material presented here is drawn from these last two sources.

As eloquently detailed in Dudewicz and Koo (1982, p.12), statistical science has

historically had a "preoccupation with existence of effects." Questions such as, "does smoking cause cancer?", or "do seatbelts reduce accident severity?" as well as tests of hypotheses of the same are often the focus of empirical investigation. However, with the pioneering work of Bechhofer (1954) statistical techniques were developed to answer comparative questions concerning the relative magnitudes of populations or treatments. Responding to such questions as, "a farm may be planted with any of several different varieties of wheat, which one has the highest yield?" or "heat treated steel may be produced with several different additives, which such steel is strongest?" became the goal of certain branches of applied statistical research. It is questions of the latter variety that selection procedures attempt to answer: questions of "bigness". This section details two selection methods under the assumption that the populations of interest are independent and have equal variance; the methods are most readily applicable to controlled experiments. While independence of populations is unlikely in economic applications, these methods provide the basis for the dependent case to be discussed in section 3.

Selection procedures can be divided into two types: *indifference zone selection procedures* and *subset selection procedures*. *Indifference zone* selection procedures were originally due to Bechhofer (1954) and were later considered by Fabian (1962) and Desu (1970). *Subset selection procedures* were due to Gupta (1956, 1965). Both types of procedures are discussed in this paper. The basic framework for either type of selection procedure is as follows. Let κ_1, ..., κ_N be N independent populations with cumulative distribution functions $F(y,\alpha_1)$, ..., $F(y,\alpha_N)$, respectively. Typically, α_i is a population parameter such as the population mean. We assume that F is a normal distribution function. This assumption can be relaxed but is the most common distributional

assumption made and, for the purposes of economics, is usually appropriate for inference. Denoting the ordered parameters as $\alpha_{[N]} \geq \alpha_{[N-1]} \geq \ldots \geq \alpha_{[1]}$, the purpose of selection procedures is to use sample estimates of the parameters to make some sort of selection from the N populations concerning the α_i, while controlling the probability of making a correct selection. Typical selections might be: the one population with the largest (smallest) α_i, a subset of populations with the largest (smallest) α_i, or a subset of populations that includes the $k < N$ populations with the largest (smallest) α_i. The present survey is by no means exhaustive; it therefore only includes the most basic results in the literature and only those results that make sense in economic or econometric applications. We therefore restrict attention to the problem of selecting the population or subset of populations with the largest (smallest) α_i. The interested reader is referred to Gupta and Panchapakesan (1979), for a more complete discussion of the literature.

2.2 Indifference Zone Selection.

Let the α_i be the unknown population means, and assume that the populations have standard variance σ^2. Consider the problem of selecting that population with the largest mean, $\alpha_{[N]}$. Given independent random samples of size T from each population, we can calculate independent unbiased estimates of the population means, $\hat{\alpha}_1$, ..., $\hat{\alpha}_N$. Let $\hat{\alpha}_{(i)}$ be the sample mean from the population with mean $\alpha_{[i]}$. That is, $\alpha_{[N]} > \alpha_{[N-1]}$ does not necessarily imply that $\hat{\alpha}_{(N)} > \hat{\alpha}_{(N-1)}$. Selecting the population associated with $max_i\ \hat{\alpha}_i$ as our estimate of $[N]$, then

$$\Pr\{\text{population } [N] \text{ is selected}\} = \Pr\{\ \hat{\alpha}_{(N)} \geq \hat{\alpha}_{(N-1)}, \text{ for } i = 1, \ldots, N\text{-}1\}$$

$$= \int_{-\infty}^{\infty} \prod_{i=1}^{N-1} \Phi\left\{ z + \frac{\sqrt{T}}{\sigma}\left(\alpha_{[N]} - \alpha_{[i]}\right)\right\} d\Phi(z).$$

As $\alpha_{[N]} > \alpha_{[i]}$ is not known, we select some probability P*$\in (N^{-1}, 1)$ and some threshold δ^* such that,

$$\int_{-\infty}^{\infty} \Phi^{N-1}\{z+h\} d\Phi(z) \geq P^* \qquad h = \frac{\sqrt{T}}{\sigma}\delta^* \tag{1}$$

Tables for various values of *N*, *h* and *P** are contained in Bechhofer (1954) and Dudewicz and Koo (1982). Gupta (1963) and Milton (1963) tabulate $2^{-1/2}h$ for various values of $N \leq 25$ and *P**. For a complete listing of tabulations see Gupta and Panchapakesan (1979, section 23.2). The threshold δ^* partitions the parameter space of the α into two zones: the *preference zone*, where $\alpha_{[N]} - \alpha_{[N-1]} \geq \delta^*$, and the *indifference zone*, where $\alpha_{[N]} - \alpha_{[N-1]} < \delta^*$, hence the name of the procedure. For controlled experiments one can select δ^* and *P** then calculate the necessary sample size *T* to ensure that the population associated with $max_i\ \hat{\alpha}_i$ is [*N*] with at least probability *P**. (This is akin to the ranking procedure briefly mentioned in the introduction.) Conversely, for a given *T* we might derive an operating characteristic curve of the procedure as the set of all (δ^*, *P**) that satisfy equation 1. That is, for a given data set from the *N* populations that satisfy the assumptions of the problem, we can make probability statements such as, "we have selected the population with the largest α_i with probability at least P* when the difference between the largest and second-largest α_i is at least δ^*." As an example, if the bio-statistician's criterion for effectiveness is "days until cured", then she might use the subset selection procedure to state "treatment A is the best treatment with at least 90% probability when the difference between the best and

second best treatment is 5 days" or "treatment A is the best treatment with at least 95% probability when the difference between the best and second best treatment is 10 days".

Of course, the "independence of the populations" and "known variance" assumptions are unlikely to hold in economics applications, and the usefulness of the aforementioned probability statement is suspect, because we must arbitrarily select δ^*.[4] However, the present discussion is merely pedagogical, and more "realistic" discussions are presented in the sequel. This is the fundamental indifference zone procedure due to Bechhofer (1954), also considered by Tamhane and Bechhofer (1977, 1979). Modifications to the fundamental procedure are too numerous to detail here, but include procedures for σ^2 unknown due to Bechhofer et al. (1954) and Dunnett and Sobel (1954); differing unknown variances: σ_i^2 due to Dudewicz and Dalal (1975); selecting $\alpha_{[N-r]}$ through $\alpha_{[N-k]}$ populations $N > k > r$ due to Bechhofer (1954); and selection based on smallest σ_i^2 due to Bechhofer and Sobel (1954), to name but a few. For a complete bibliographic listing see Dudewicz and Koo (1982) or Gibbons (1982). We now discuss subset selection procedures which are more naturally applicable to economic analysis. Here, we maintain the independence assumption, but relax it in section 3.

2.3 Subset selection.

[4] In most of economics sample data are pooled and a single regression equation is estimated for all populations. Population distributions are implicitly assumed to be correlated in some fashion, and pooling the data improves estimation efficiency. Consequently, the population specific parameter estimates are usually correlated. It is in this sense that an independence assumption seems dubious. However, certain branches of economics (such as the labor market discrimination literature) actually split the sampled data into several populations and run separate regressions (e.g. into males and females or into whites and blacks). In these instances, an independence assumption across parameter estimates may be reasonable, and this procedure applicable.

Now assume that σ^2 is unknown, but can be estimated by s^2 based on ν degrees of freedom. Gupta (1956, 1965) showed that if one selects a subset of the *N* populations, *S*, according to the rule:

$$S = \{i : \hat{\alpha}_i > \max_{j \neq i} \hat{\alpha}_j - T^{\lambda}_{N,\nu,\rho} s\sqrt{2}T^{-1/2}\} \tag{2}$$

then $\Pr\{[N] \in S\} \geq 1 - \lambda$, where $T^{\lambda}_{N,\nu,p}$ is the solution in *t* of,

$$\int_0^{\infty}\left[\int_{-\infty}^{\infty}\Phi^{N}\left\{\frac{u\rho^{1/2}+tx}{(1-\rho)^{1/2}}\right\}d\Phi(u)\right]dF_{\nu}(x) = 1-\lambda. \tag{3}$$

F_{ν} is the c.d.f of a $\sqrt{\chi^2/\nu}$ random variable, and $\rho = 0.5$. $T^{\lambda}_{N,\nu,\rho}$ is the upper-λ percentage point of a multivariate Student *t* distribution with common correlation coefficient ρ. Tables for $T^{\lambda}_{N,\nu,\rho}$ are contained in Dunnett and Sobel (1954, 1955), Cornish (1954) and most recently in Bechhofer and Dunnett (1986). For tabulations of critical points of the limiting multivariate normal distribution (i.e. $\nu \to \infty$) see Odeh (1982) and Horrace (1998).

Equation (2) allows us to make inference statements such as, "the subset *S* contains the largest population with probability at least 1 - λ". If our bio-statistician selected $\lambda = 0.05$, she might find that *S*={A, C} and could make the statement, "Treatments A and C are best with 95% probability", or if she selected $\lambda = 0.10$, she might find that *S*={A} and could make the statement, "Treatments A is best with 90% probability". This type of confidence statement seems to make more sense for economic analysis than does the indifference zone statement, because we are not required to select the threshold δ^*. For instance, in the estimation of stochastic frontier models, where α_i could represent the technical efficiency of the i^{th} firm, the subset *S* would contain all firms

that are technically efficient at the (1-λ)$\times$100% confidence level. For example, in the empirical section we estimate the technical efficiency of ten Texas electric utilities, and find that the utility with the highest efficiency estimate is firm 5 and that with the second-highest is firm 3. The estimation results might lead us to erroneously conclude that firm 3 is inefficient relative to firm 5. However, using a selection procedure (to be described in the next section), it is asserted that "firms 5 and 3 are the efficient with 95% confidence, the other eight are not".[5] This is a very powerful probability statement which precludes us from jumping to conclusions about the results of the analysis.

As was the case with the indifference zone procedure, modifications to the basic subset selection result are too numerous to list here, but include a procedure for unequal sample sizes due to Gupta and Huang (1974); a procedure for selection based on $|\alpha_i|$ due to Rizvi (1971); a procedure for selection in terms of variance due to Gupta and Sobel (1962) and myriad other procedures for non-normal distributions. Again the interested reader is referred to Dudewicz and Koo (1982) or Gibbons (1982).

2.4 The Complications of Economic Data

The foregoing procedures suffer from assumptions that will generally preclude their use in economic empirical analyses and that manifest directly in the calculation of the appropriate critical values for the inference. Specifically, the calculability and relative simplicity of equations (1) and (3) hinge directly upon the assumption of independence of populations and on *N* being small; requirements which may not hold in economic applications. These features are discussed below and in subsequent sections.

First, the assumption of independence of populations is typically not relevant in economic analysis. In economic applications, where covariates are commonly employed

[5] McCloskey might say that, "utilities 5 and 3 are Big, and the rest are not with 95% probability".

and where experiments are not "controlled" orthogonality is the exception. For instance if the α_i were N slope parameters from a cross sectional regression analysis, then estimates $\hat{\alpha}_i$ are typically correlated through the exogenous variables, so that equations (1) and (3) would not apply.[6] As we shall see, these simple probability integrals will be replaced with N-dimensional probability integrals with intractable covariance structures. Second, economic field data sets can have extremely large values of N. When N is greater than 50, the probability integrals of equations (1) and (3) become difficult to calculate numerically. Therefore, even in the presence of orthogonality populations, these selection procedures may be difficult to perform because critical value tables may not exist.

Both of the preceding complications can be overcome, if we are willing to forego numerical solution of the probability integrals and replace it with simulation. Using simple computer algorithms it is a straight-forward task to artificially generate critical values that satisfy the probability statements regardless of whether independence is violated or the number of populations is large. For example, using simulation techniques Horrace (1998) generates critical values, $T^{\lambda}_{N,\infty,\rho}$, for values of N as high as 500. Notice that these critical values are for the equicorrelated case (e.g. when populations are orthogonal). We detail this simulation technique for finite v and a general correlation structure for the populations in the following section.

3. Selection Procedures: The General Case

3.1. Preliminaries.

If the populations, κ_i, are correlated with some unknown covariance structure, these correlations will manifest themselves as correlations among the estimates of the α_i.

[6] An example of this particular situation is provided in the sequel.

Again let $\hat{\alpha}_i$ be an unbiased estimate of the α_i. Let the covariance matrix of the $\hat{\alpha}_i$ be the ($N \times N$) matrix $\hat{\Omega}$, based on v degrees of freedom. Let $\hat{\omega}_{sr}$ represent the element in the s^{th} row in the r^{th} column of $\hat{\Omega}$, $s = 1, ..., N$, $r = 1, ..., N$. Given this specification we generalize the subset selection procedure of equations (2) and (3). However, before embarking on a discussion of this generalization we first develop the distributional theory that allows generalization of the probability statements of equation (1) and (3) to the case where covariance structures are non-spherical and unknown.

Let the random vector $Z = (Z_1, ..., Z_p)$ have a p-variate standard normal distribution, i.e. $E(Z_i) = 0$ and $Var(Z_i) = 1$. Let the covariance matrix of Z equal Ω and its correlation matrix equal R. Let U be distributed independently of Z as a χ^2 random variable with v degrees of freedom. Let $T_i = Z_i(U/v)^{-1/2}$. Then $T = (T_1, ..., T_p)$ has a p-variate Student t distribution with correlation matrix R and v degrees of freedom. The joint density of T is given by

$$f(t_1,....t_p, R, v) = \frac{\Gamma((p+v)/2)}{(v\pi)^{p/2}\Gamma(v/2)}(\det R)^{-1/2}(1+T'R^{-1}T/v)^{-(p+v)/2}.$$

Define the critical value $T^{\lambda}_{p,v,R}$ as the solution in t of the equation

$$\mathrm{P}\left\{\max_i |T_i| \le t\right\} = \int_{-t}^{t}...\int_{-t}^{t} f(t_{1,....,}t_{p,}R,v)dt_{1,....,}dt_p = 1-\lambda \qquad (5)$$

When the Z_i are independent, the correlation matrix is the identity matrix and the probability integral of equation (5) reduces to that of equation (3) with $p = N$, and for moderate values of N, equation (3) can be solved numerically. In this case the variates are said to be equicorrelated and solutions to equation (5) are commonly tabulated as $T^{\lambda}_{p,v,\rho}$. Of course, economic data rarely admit independent structure of variates and tabulations

of $T^{\lambda}_{p,v,R}$ would be clearly impractical. Without the equicorrelated structure, numerical solution of equation (5) is cumbersome, particularly when *p* is large. However, simulation of $T^{\lambda}_{p,v,R}$ is rather straight-forward:

1. *Perform a Choleski decomposition of* Ω *into Q, such that Q'Q =* Ω.
2. *Generate p independent standard normal variates:* $Z_m' = [Z_{1m}, ..., Z_{pm}]$.
3. *Generate an independent chi-squared random variable, U, with ν degrees of freedom.*
4. *Calculate* $T_m = Q'Z_m(U/v)^{-1/2}$, *a p-dimensional t variate with correlation matrix R.*
5. *Find* $Y_m = max \mid T_m \mid$, *the maximum element of* T_m.
6. *Perform steps 2, 3, 4 and 5 for m = 1, ..., M.*
7. *Calculate a* $(1-\lambda) \cdot 100$ *percentile from* Y_m, $m = 1, ..., M$. *This simulated value serves as a consistent estimate of* $T^{\lambda}_{p,v,R}$.

As $M \to \infty$, the simulated value approaches the solution in *t* of equation (5). Horrace (1998) provides an algorithm for determining confidence intervals for the coverage probability, $(1-\lambda)$. Since the limiting distribution of a multivariate Student *t* variate is a normal variate, for large values of *v* one can skip steps 3 and 4 and let $T_m = Q'Z_m$ in steps 5 through 7.[7]

3.2 Multiple Comparisons with a Control.

The first step toward generalizing the subset selection procedure of equation 2 is to discuss a multiple comparison procedure called multiple comparisons with a control, (MCC), initially due to Dunnett (1955). Let the k^{th} population be regarded as a control.

We construct simultaneous (1-λ)×100% confidence intervals on α_k - α_i, i = 1, ..., k-1, k+1, ..., N. These intervals are given by:

$$\alpha_k - \alpha_i, \in \left\{L_i^k, U_i^k\right\} \quad i = 1,\ldots,k-1,k+1,\ldots,N. \tag{6}$$
$$L_i^k = \hat{\alpha}_k - \hat{\alpha}_i - T^{\lambda}_{N-1,\nu,R}(\hat{\omega}_{kk} + \hat{\omega}_{ii} - 2\hat{\omega}_{ik})^{1/2}$$
$$U_i^k = \hat{\alpha}_k - \hat{\alpha}_i + T^{\lambda}_{N-1,\nu,R}(\hat{\omega}_{kk} + \hat{\omega}_{ii} - 2\hat{\omega}_{ik})^{1/2}$$

The interpretation of these intervals is straight-forward. To construct confidence intervals around a population parameter (in this case a set of population parameters) use the sample estimate of the parameter $(\hat{\alpha}_k - \hat{\alpha}_i)$ plus or minus an allowance term consisting of the product of a critical value ($T^{\lambda}_{N-1,\nu,R}$) and a standard error $(\hat{\omega}_{kk} + \hat{\omega}_{ii} - 2\hat{\omega}_{ik})^{1/2}$. The critical value is based on the correlation matrix of the α_i. It should be noted, however, that the parameters of interest here are *not* the α_i, rather the α_k - α_i, $i \neq k$. Therefore, the critical value, ($T^{\lambda}_{N-1,\nu,R}$), should *not* come from the estimated covariance matrix of the $\hat{\alpha}_i$ (i.e. $\hat{\Omega}$), but instead from the estimated covariance matrix of the $\hat{\alpha}_k$ - $\hat{\alpha}_i$ (i.e. $L\hat{\Omega}L'$), where L is a N-1 *negative* identity matrix with a column of ones inserted between the (k-1)th and k^{th} columns. That is, if α' = [α_1, ..., α_N] and $\alpha^{k\prime}$ = [α_k - α_1, …, α_k - α_{k-1}, α_k - α_{k+1},…,α_k - α_N], then L is a (N-1)×N matrix such that $L\alpha = \alpha^k$. Hence, to implement this technique the Choleski decomposition in step 1 of the critical value simulation algorithm should be $Q'Q = L\hat{\Omega}L'$. For the purposes of a generalized subset selection procedure, the salient feature of these intervals is the upper bound, U_i^k, which provides information on the relative magnitude of the k^{th} population parameter.

3.3 Generalized Subset Selection.

[7] GAUSS code is available from the author to generate both the normal and Student t critical values. Visit: www.u.arizona.edu/~whorrace/mcresources.html for the code.

Edwards and Hsu (1983) developed a subset selection technique that generalizes equation 2 to the case where the populations are not independent. The technique hinges on the existence of MCC intervals of equation 6. Edwards and Hsu (1983) show that if one selects a subset of the N populations, S, according to the rule:

$$S = \{k: U_i^k \geq 0 \text{ for } i = 1, ..., k-1, k+1, ..., N\} \quad (7)$$

then $\Pr\{[N] \in S\} \geq 1 - \lambda$. The interpretation of equation (7) is simple. If U_i^k are the MCC upper bounds with population k as the control and if all the $i \neq k$ upper bounds are large (non-negative), then the k^{th} population is one of the largest at the $(1-\lambda)\times 100\%$ confidence level. The subset, S, consists of all populations, k, that meet this criterion. To perform this inference one must construct (N-1) confidence intervals for each of the N populations. Therefore, for large N the number of confidence intervals become prohibitively large for hand-calculations.[8] Fortunately, for each of the N populations, the population can be eliminated from the subset, S, once any single MCC upper bound fails the upper bound criterion. Consequently, most analyses will not require strict calculation of $N(N-1)$ upper bounds, but some number less than this.

When, does a population fall into the subset, S? As described above, when its MCC upper bounds are all non-negative. This occurs when either *a*) the parameter estimate of the control, $\hat{\alpha}_k$, is large relative to the rest, *b*) the variance of the control estimate is large compared to its covariance with the rest of the populations (i.e. $\hat{\omega}_{kk} - 2\hat{\omega}_{ik}$) or *c*) the covariance structure of the ($\hat{\alpha}_k$ - $\hat{\alpha}_i$) or the values of N or λ are such that $T^{\lambda}_{N-1,\nu,R}$ is large. Case *a* is obvious: ignoring sampling error, large α_i tend to

[8] GAUSS code is available from the author to perform this procedure. Visit: www.u.arizona.edu/~whorrace/mcresources.html for the code.

produce large $\hat{\alpha}_i$. Cases *b* and *c* illustrate that even though $\hat{\alpha}_k$ is small, this does not mean that its population equivalent must also be small. Sampling variability and the covariance structure of the parameter estimates can cause a large α_k to produce a small $\hat{\alpha}_k$, and this anomaly can only be detected with a properly constructed inference procedure. Case *c* also embodies the multiplicity of the inference statement. When *N* is large we are making many individual comparisons *simultaneously*, so the rejection region of the multivariate sampling distribution must decrease and the critical values must increase to control for the overall error rate of the statement. Some examples follow.

4. Examples

To illustrate the utility of the subset selection procedure for economic applications, two analyses are provided: one based on simulated data and the other based on actual data.

4.1 Simulation Example.

A simulated data study was performed to highlight various features of the selection procedures that an empirical study could not. Consider the econometric specification:

$$y_i = \alpha_1 + \alpha_2 v_i + \alpha_3 w_i + \alpha_4 x_i + \alpha_5 z_i + \varepsilon_i \qquad i = 1,...N. \qquad (8)$$

where *u*, *v*, *w*, *x*, *y*, and *z* are data, α_i are parameters for estimation and ε_i is *iid* $N(0,\sigma^2)$. Interest centers on estimating the model's slope parameters and performing inference on their relative magnitudes. An example of such a specification in economics is a labor market wage regression where *y* = ln(*wage*) and the right-hand-side (RHS) variables are configured such that certain slope parameters represent wage gap estimates across various industries. An example of such an application is found in Fields and Wolf (1995). We are interested in knowing in a statistical sense which of the 5 slope parameters (wage

gaps) are the biggest. To this end, data on the RHS variables of equation 8 were simulated using a GAUSS uniform random number generator on the unit interval. Slope parameter values were selected as $\alpha_1 = 1$, $\alpha_2 = 2$, $\alpha_3 = 3$, $\alpha_4 = 4$, $\alpha_5 = 5$. Data on the y_i were then generated using the slope parameter values, the RHS data and ε_i "data" simulated from a GAUSS $N(0,1)$ random number generator. Three data sets were generated in this fashion with sample sizes $N = 25$, 50 and 100. Least-squares estimates of the parameter values were calculated for each of the three sample sizes and are reported in the second column of Tables 1, 2, and 3. Corresponding standard errors on the slope parameters are shown in column 3. The slope parameters estimates were correlated in the sample so the generalized subset selection procedure of equation (7) was performed to draw inferences on the slope estimates. First, critical values, $T^{\lambda}_{4,\nu,R}$, were simulated for each sample size and for each parameter, using λ=0.05, $\nu = N$-5 and the particular covariance matrix generated by each data set.[9] For each critical value simulation the simulation sample size, M, was set to 10,000. Individual critical values for $\lambda = 0.05$ are tabulated in column 4 of each table.

These critical values were used to construct the MCC upper bounds of equation (6) and ultimately the subset, S, of equation (7). The elements of the subset, S, are contained in Table 4 for each N. For $N = 25$ the subset consisted of indices 3, 4 and 5, implying that with at least 95% confidence the slope parameters α_3, α_4 and α_5 are the largest parameters. For $N = 50$ and $N = 100$ the subset consisted of indices 3 and 4,

[9] One critical value was need for each parameter estimate, because the generalized subset selection procedure requires calculation of a set of MCC upper bounds, U_i^k, for each parameter (in turn) as the control parameter. The covariance structure of the estimates is $\hat{\Omega}$ = Var($\hat{\alpha}$), $\hat{\alpha}' = [\hat{\alpha}_1,...,\hat{\alpha}_N]$. The

implying that α_4 and α_5 are the largest parameters at the 95% level or better. We do not know which index in the subsets is the largest, because sampling error confounds this determination. However, we can say that the slope parameters of the indices contained in the subsets are bigger than those not in the subsets. It should also be clear that as *N* increases the precision of the inference increases, since the cardinality of *S* decreases.[10]

As an additional experiment, critical values for the $N = 100$ data and $\lambda = 0.10$ were also simulated. The larger value of λ resulted in smaller critical values. These are tabulated in the fifth column of Table 3. The subset was again calculated based on the new, smaller critical values, and this time it was a singleton, $S = \{5\}$. (See the last row of table 4). The implication is that for $N = 100$, α_4 and α_5 are the largest parameters with probability 0.95, but α_5 is the largest with probability 0.90.

A few additional comments are in order. First, note that in Table 1 the ordering of the estimates of $\hat{\alpha}_3$ and $\hat{\alpha}_4$ are reversed in terms of their magnitudes. This illustrates how sampling variability can distort sample rankings of parameter estimates. However, the selection procedure captures this by selecting $S = \{3, 4, 5\}$. That is, the estimation might erroneously infer that $\alpha_3 > \alpha_4$, but the inference suggests otherwise: at the 95% level we cannot distinguish between α_3 and α_4, and that they (along with α_5) might all be the largest parameters. Second, the critical values in each table vary across the slope parameters. Had the parameter estimates admitted an equicorrelated structure the critical values all would have been identical. It is the difference in the variance and the

---

covariance structure for each critical value is then $L\hat{\Omega}L'$, with *L* being different for each of the five parameter estimates.

[10] This clearly demonstrates that as sample size increases, the differences among the population parameters become simultaneously "statistically different from zero". It is in this sense that these procedures can be

covariances of the estimates (lack of equicorrelation) that induces the different critical values. Third, in Table 3 the difference between the estimates $\hat{\alpha}_4$ and $\hat{\alpha}_5$ is large relative to the difference between estimates $\hat{\alpha}_3$ and $\hat{\alpha}_4$. It was not large enough, however to make $\hat{\alpha}_4$ significantly different from $\hat{\alpha}_5$ when $\lambda = 0.05$ (i.e. $S = \{4,5\}$). This was probably due to a high degree of noise in the simulated data as evidenced by the relatively large value of the estimate $s^2 = 1.2263$ compared to the true value $\sigma^2 = 1$. When the subset selection was performed with $\lambda = 0.10$, the difference between the two estimates was significant, as $S = \{5\}$.

4.2 Empirical Example.

Consider the Cobb-Douglas specification of the fixed-effects stochastic frontier model for a panel of ten privately owned Texas electric utilities, observed annually from 1966 to 1985:

$$E_{it} = \alpha_i + \beta_L L_{it} + \beta_k K_{it} + \beta_F F_{it} + \varepsilon_{it}, \qquad i = 1, .., 10; \quad t = 1, \ldots, 18;$$

where E = electrical output, L = labor, K = capital and F = fuel. This data set was originally analyzed by Kumbhakar (1994). Interest centers on estimating each α_i (a proxy for technical efficiency of the i^{th} firm) and ranking them to determine the most efficient firm in the sample. The model was estimated with the so-called "within" estimation technique. Slope estimates were β_L = -0.1291, β_k = 0.6275 and β_F = 0.5652. Estimates of the α_i for each firm are contained in Table 5. Based on the covariance structure of the $\hat{\alpha}_i$, the generalized subset selection procedure of equation (7) was performed at the 90% confidence level ($\lambda = 0.10$), producing a subset, S = {5, 3}.

---

considered tests of significance and are not immune to strict interpretation of McCloskey's criticism. However, our purpose is not to debate McCloskey on the merits of Neyman-Pearson testing procedures.

That is, with probability at least 0.90 the utilities 5 and 3 are the most efficient in the sample, and the rest of the firms are relatively inefficient. This is a powerful inference result.

5. Conclusions

This paper has introduced economists to ways of determining "How Big is Big". It has argued that questions of size may be relevant to economists and that these questions are usually not answered with any statistical rigor. Selection procedures have always provided a tool to answer these questions; they have just never been embraced by economists. It is clear from the empirical exercise that the solutions are now within reach. All that remains is to encourage their use within the discipline. As mentioned, studies have already been done that select the most efficient firm and the largest wage gaps across industries. It is interesting to speculate on other potential economic applications of the procedures. Selection of the countries with the largest growth rate, selection of the largest elasticities and selection of the most effective healthcare delivery system are all potentially interesting problems.

Table 1. N = 25

Parameter	Estimate	Standard Error	$T^{.05}_{4,20,R}$
$\alpha_1 = 1$	-0.7865	0.6085	2.677
$\alpha_2 = 2$	2.1434	0.8833	2.624
$\alpha_3 = 3$	4.5912	0.7196	2.682
$\alpha_4 = 4$	4.5030	0.6680	2.683
$\alpha_5 = 5$	5.6561	0.7609	2.627
$\sigma^2 = 1$	0.8833		

Table 2. N = 50

Parameter	Estimate	Standard Error	$T^{.05}_{4,45,R}$
$\alpha_1 = 1$	1.4389	0.4951	2.484
$\alpha_2 = 2$	1.6398	0.4881	2.551
$\alpha_3 = 3$	2.8878	0.4614	2.557
$\alpha_4 = 4$	3.2522	0.5166	2.560
$\alpha_5 = 5$	5.0651	0.4916	2.539
$\sigma^2 = 1$	0.9172		

Table 3. N = 100

Parameter	Estimate	Standard Error	$T^{.05}_{4,95,R}$	$T^{.10}_{4,95,R}$
$\alpha_1 = 1$	0.0889	0.4230	2.440	2.119
$\alpha_2 = 2$	3.2059	0.4135	2.482	2.195
$\alpha_3 = 3$	3.4010	0.4076	2.460	2.188
$\alpha_4 = 4$	3.7447	0.3991	2.523	2.205
$\alpha_5 = 5$	5.1070	0.3828	2.497	2.190
$\sigma^2 = 1$	1.2263			

Table 4. Subset, *S*

Sample Size	λ	Subset, *S*
$N = 25$	0.05	{3, 4, 5}
$N = 50$	0.05	{4, 5}
$N = 100$	0.05	{4, 5}
$N = 100$	0.10	{5}

Table 5. Texas Utility Order Statistic

Firm	$i = 5$	$i = 3$	$i = 10$	$i = 1$	$i = 8$	$i = 9$	$i = 2$	$i = 6$	$i = 7$	$i = 4$
$\hat{\alpha}_i$	-4.995	-5.083	-5.145	-5.176	-5.194	-5.211	-5.218	-5.236	-5.237	-5.267

References

Bechhofer, R.E. (1954). A single-sampled multiple decision procedure for ranking means of normal populations with known variances. *Annals of mathematical Statistics.* 25, 16-39.

Bechhofer, R.E. and C.W. Dunnett (1986), Tables of the percentage points of multivariate Student t distributions, R.E. Odeh and J.M. Davenport, eds. *Selected Tables in Mathematical Statistics* (Providence: American Mathematical Society).

Bechhofer, R.E., C.W. Dunnett and M. Sobel (1954). *Annals of Mathematical Statistics*, 25, 170-176.

Bechhofer and Sobel (1954). A single-sampled multiple decision procedure for ranking variances of normal populations with known variances. *Annals of mathematical Statistics.* 25, 273-289.

Burdick, D.S., T.H. Taylor and W.E. Sasser (1967). Computer simulation experiments with economic systems: the problem of experimental design, *Journal of the American Statistical Association*, 62, 1315-1337.

Cornish, E.A. (1954), The multivariate small t-distribution associated with a set of normal sample deviates, *Australian J. of Physics.* 7, 531-42.

Desu, M.M (1970) A selection problem. *Annals of Mathematical Statistics*, 41, 1596-1603.

Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association.* 50, 1096-1121.

Dunnett, C.W. and M. Sobel (1954), A bivariate generalization of Student's t-distribution with tables for certain special cases, *Biometrika* 41, 153-69.

Dunnett, C.W. and M. Sobel (1955), Approximations to the probability integral and certain percentage points of a multivariate analogue of Student's t-distribution, *Biometrika*, 42, 258-60.

Dudewicz, E.J. and S.R. Dalal (1975). Allocation of observations in ranking and selection with unequal variances. *Sankhya, Series B,* 37, 28-78.

Dudewicz, E.J. and J.O. Koo (1982). *The Complete Categorized Guide to Statistical Selection and Ranking Procedures. American Science Press,* Columbus, Ohio.

Edwards, D.G. and J.C. Hsu (1983). Multiple comparisons with the best treatment, *J. Amer. Stat. Assoc.* 78, 965-71. Corrigenda (1984), *J. Amer. Stat. Assoc.* 79, 965.

Fabian, V. (1962). On multiple decision methods for ranking population means *Annals of Mathematical Statistics,* 33, 248-254.

Fields, J. and E.N. Wolff (1995) Interindusty wage differentials and the gender wage gap. *Indust. And Labor Rel. Rev.* 49, 105-20.

Gibbons, J.D. (1982). Selection Procedures. In *Encyclopedia of Statistical Sciences*, Kotz, S. and N.L Johnson eds. New York : Wiley.

Gupta, S.S. (1956). On a decision rule for a problem of ranking means. *Institute of Statistics Mimeo Series* No. 150, University of North Carolina.

Gupta, S.S. (1963). On a selection and ranking procedure for gamma populations. *Annals of Mathematical Statistics.* 14, 199-216.

Gupta, S.S. (1965). On some multiple decisions (selection and ranking) rules. *Technometrics. 7*, 225-245.

Gupta, S.S. and W.T. Huang. A note on selecting a subset of normal populations with unequal sample sizes. *Sankhya, Series A*, 336, 389-396.

Gupta, S.S. and S. Panchapakesan (1979). *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations.* Wiley, New York.

Gupta, S.S. and M. Sobel (1962). On selecting a subset containing the population with the smallest variance, *Biometrika,* 49, 495-507.

Horrace, W.C. (1998). Tables of percentage points of the *k*-variate normal distribution for large values of *k*. *Communications in Statistics: Simulation & Computation,* 27, 823-831.

Horrace, W.C. (1999). On the ranking uncertainty of labor market wage gap estimates. Unpublished Manuscript. University of Arizona, Department of Economics.

Horrace, W.C. and P. Schmidt (1996). Confidence statements for efficiency estimates from stochastic frontier models. *J.of Productivity Analysis.* 7, 257-82.

Horrace, W.C. and P. Schmidt (1999). Multiple comparisons with the best, with economic applications. *Journal of Applied Econometrics*, Forthcoming.

Haurin, D.R. (1989) Women's Labor Market Reactions to Family Disruptions. *The Review of Economics and Statistics*, 71, 54-61.

Kumbhakar, S.C. (1996). Estimation of cost efficiency with heteroscedasticity: an application to electric utilities in Texas, 1966-1985. *Journal of the Royal Statistical Society, Series D*, 45, 319-335.

McCloskey, D (1988), Two vices: proof and Significance, Presented at the American Economics Association Winter Meetings, January 3, 1988. Chicago.

Milton, R.C. (1963). Tables of the equally correlated multivariate normal probability integral, Technical Report No 27, Department of Statistics, University of Minnesota, Minneapolis.

Mowery, D.C. (1983) Industrial Research and Firm Size, Survival, and Growth in American Manufacturing, 1921-1946: An Assessment. *Journal of Economic History*, 43, 953-980.

Odeh, R.E (1982), Tables of the percentage points of the distribution of the maximum absolute value of equally correlated normal random variates, *Communications in Statistics:, Series B* 11, 65-87.

Rizvi, M.H (1971) Some selection problems involving folded normal distributions, *Technometrics,* 13, 355-369.

Seale, J. L. (1990). Estimating stochastic frontier systems with unbalanced panel data: the case of floor tile manufactories in Egypt. *Journal of Applied Econometrics.* 5, 59-79.

Tamhane, A.C. and R.E. Bechhofer (1977). *Communications in Statistics, Series A* 6, 1003-1033.

Tamhane, A.C. and R.E. Bechhofer (1979). *Communications in Statistics, Series A* 8, 337-358.