Selected Paper AAEA/AERE Summer Meetings, Long Beach, CA., July 2002

Simple Computational Methods for Measuring the Difference of Empirical Distributions: Application to Internal and External Scope Tests in Contingent Valuation*

Gregory L. Poe
(Corresponding Author)
Associate Professor
Department of Applied Economics and Management
Cornell University
454 Warren Hall
Ithaca, NY 14853
Ph: 607-255-4707
E-mail: GLP2@cornell.edu

Kelly L. Giraud
(Presenting Author)
Assistant Professor
Department of Resource Economics and Development
University of New Hampshire
312 James Hall
Durham, NH 03824
Ph: 603-862-4811
E-mail: kelly.giraud@unh.edu

John B. Loomis
Professor
Department of Agricultural and Resource Economics
Colorado State University
Fort Collins, CO 80523-1172
Ph: 970-491-2485
E-mail: jloomis@ceres.colostate.edu

September 25, 2001

* This paper has benefited from conversations with Michael Welsh, Ian Bateman, Richard Ready, and Christian Vossler, although any errors remain the responsibility of the co-authors. Funding for this project was provided by USDA Regional Project W-133 and the U.S. Fish and Wildlife Service. This paper was written in part while Poe was a visiting Fellow at the Jackson Environmental Institute (JEI) and the Centre for Social and Economic Research on the Global Environment (CSERGE) at the University of East Anglia, UK. Earl Ekstrand of the U.S. Bureau of Reclamation assisted in the survey design. Cornell University Applied Economics and Management Staff Paper 2001-05.

Simple Computational Methods for Measuring the Difference of Empirical Distributions: Application to Internal and External Scope Tests in Contingent Valuation

Abstract: This paper develops a statistically unbiased and simple method for measuring the difference of independent empirical distributions estimated by bootstrapping or other simulation approaches. This complete combinatorial method is compared with other unbiased and biased methods that have been suggested in the literature, first in Monte Carlo simulations and then in a field test of external and internal scope testing in contingent valuation. Tradeoffs between methods are discussed. When the empirical distributions are not independent a straightforward difference test is suggested.

Simple Computational Methods for Measuring the Difference of Empirical Distributions: Application to Internal and External Scope Tests in Contingent Valuation

I. Introduction

Applied economics has capitalized on the dramatic advances in the computational capacity of desktop computers. In some instances, however, the adoption of more computationally-intense methods has proceeded without adequate attention to underlying statistical foundations. Such is the case of simulated or bootstrapped distributions widely used in applied welfare economics, and measures of the differences in these distributions. Whereas there has been longstanding interest in applying and comparing alternative methods of empirically approximating distributions of economic parameters (e.g., Green, Hahn and Rocke, 1987; Krinsky and Robb, 1991; Li and Maddala, 1999), the extension of these methods to assessing the statistical significance of differences in parameter estimates has largely been disconnected from the relevant statistical theory. Simple computational methods widely used in economic applications, such as the non-overlapping confidence interval criterion or normality assumptions, have been demonstrated to be biased (Poe, Severance-Lossin and Welsh, 1994); unbiased empirical methods based on the theoretically-correct method of convolutions are computationally intensive, apparently exceeding the capacity of many researchers[1]. Perhaps reflecting the level of programming complexity, researchers in leading applied economics journals have continued to resort to statistically biased measures of differences (e.g., Berrens,

[1] For example, in adopting a sampling approach discussed later in this paper, Ready, Buzby and Hu (1996) note that discrete approximations of the convolution "are very difficult to calculate without specialized software" (p. 406). Similarly, in defending their use of a normality-based difference test, Rollins and Lyke (1998) suggest that the convolutions approach is not "intuitive".

Bohara, and Kerkvleit, 1997; Rollins and Lyke, 1998; Kotchen and Reiling, 1999; Reaves, Kramer, and Holmes, 1999; Hutchinson *et al.,* 2001).

The primary objectives of this paper are to: 1) Take advantage of recent expansions in computer speed and memory to develop a statistically unbiased measure of the convolution that can be readily applied in most statistical programs using simple do loops; and 2) To compare this "complete combinatorial" method with other methods of comparing differences of independent distributions that have been suggested in the literature. The sole limitation of this new approach is the size of memory and time taken for the necessary computations. However, the capacity of most current econometric software such as Gauss, SAS, STATA, TSP, and Shazam exceed these requirements. In addition, we demonstrate a computationally simple algorithm for comparing distributions that are not independently distributed.

This paper is organized as follows. In the next section, we review the statistical theory of differences in distributions. In Section III we summarize biased and unbiased measures of these differences that have been used in the literature and introduce a new complete combinatorial method that provides an exact measure of the difference between two distributions. These methods are applied in Section IV to known, independently drawn samples, while the fifth section provides results from a field study of internal and external scope in contingent valuation. We summarize and extend our findings in the final section.

II. Statistical Foundations: The Convolution

Our analysis is based on the assumption that the researcher has generated two empirical distributions, X and Y, with corresponding probability density functions, $f_X(x)$ and $f_Y(y)$. These distributions may depict the spread and density of welfare estimates (Kling, 1991), alternative methods of measuring non-market values (Loomis, Creel and Park, 1991), elasticities (Li and Madalla, 1999), efficiency measures (Weninger, 2001), or other parameters of interest to economics researchers. It is assumed that the research objective is to measure the precision of these parameter estimates as well as to test if these distributions are significantly different, in a manner analogous to a difference of means test.[2] The two distributions may be generated by non-parametric bootstrapping or jackknife procedures (Efron and Tibshirani, 1993), parametric bootstrapping (Krinsky and Robb, 1986), Monte Carlo simulations (Krinsky and Robb, 1991) or any number of resampling techniques. It is further assumed that these estimated distributions are a function of a number of parameters, and hence normality of the approximated parameter distribution is likely to be the exception rather than the rule.

Assuming that the two distributions are independently distributed random variables, the distribution of the difference, V = X-Y, of these two distributions is given by the subtractive variant of the convolution formula (Mood, Graybill and Boes, 1974):

---

[2] Implicitly, this assumes that the researcher is interested in descriptive statistics on the individual distributions as well as comparing these distributions. Alternatively, one could assume that the major objective is to estimate the difference directly. As such, it would be appropriate to build an estimate of the difference into the original estimation problem. While we acknowledge that this is a reasonable alternative, it should be noted that such an approach is limited to the comparisons that are directly estimated. It is often the case, however, that the researcher needs to compare a number of distributions. For example, Rollins and Lyke (1998) sought to compare a number of different levels of national park protection. Or, as in the case of the adding up or part-whole issue in contingent valuation (Bateman *et al*., 1997), it may be that the research issue is to combine and compare a number of distributions (e.g., does WTP(A)+WTP(B) = WTP(A+B)?). In these cases it is more facile to have independently estimated distributions and compare these distributions.

$$f_V(v) = \int_{-\infty}^{\infty} f_X(v+y) f_Y(y) dy \tag{1}$$

The associated cumulative distribution function at a specific value V' is:

$$F_V(V') = \int_{-\infty}^{V'} \int_{-\infty}^{\infty} f_X(v+y) f_Y(y) dy\, dv \tag{2}$$

The above equations are generally not transparent to non-statisticians. Nevertheless, the concept that these equations capture is both intuitive and simple. Basically, the convolution calculates the probability of each possible outcome, considering all possible combinations of the two independent distributions. The probability of outcomes is simply the sum of the products of each possible combination. For example, with a pair of red and blue dice, the probability of outcome red minus blue equals 5 is 1/36 (f_R (1)*f_B(6) = (1/6)*(1/6)), the probability of outcome red minus blue equals 4 is 2/36 (= f_R(1)*f_B(5) + f_R(2)*f_B(6) = (1/6)*(1/6) + (1/6)*(1/6) = 1/36 + 1/36), and so on. The convolution simply makes these calculations for all possible outcomes, for either continuous or discrete distributions.[3] Another way of looking at it, one that we will capitalize on later in the empirical section of this paper, is that the convolution essentially computes every possible combination, sorts these values, and forms a cumulative distribution. Returning to the example of the red and blue dice, the 1 on the red die would be combined with all possible (1 to 6) values on the blue die, the 2 on the red die would similarly be combined all possible blue die values and do on. In all a total of 36 combinations would be possible, each with equal probability, resulting in a sorted vector of (-5, -4, -4, …, 4, 4, 5). From this sorted vector a cumulative distribution function could be constructed.

[3] See the appendix of Poe, Severance-Lossin and Welsh (1994) for a further example.

When the distributions are not independent, the cumulative distribution function at V' is given as,

$$F_V(V') = \int_{-\infty}^{V'} \int_{-\infty}^{\infty} f_{X,Y}(v+y, y) f_Y(y) dy\, dv \tag{3}$$

where $f_{X,Y}(.)$ depicts the density of the joint distribution of X and Y. Thus when the empirical distributions are not independent, their jointness must be accounted for in their generation of the distributions as well as in the statistical tests that compare distributions.

III. Empirical Estimates of the Convolution

This section describes and assesses empirical methods of approximating the difference in distributions and estimating the one-sided significance (α) of the difference[4]. First we provide an examination of unbiased methods of estimating the subtractive convolution for independent distributions, focusing on two methods (empirical convolutions and sampling) that have been used in the literature and introducing a third, computationally simple method based on a complete combinatorial design. We then briefly describe and criticize two biased methods that have been used in the literature. The brevity here is motivated by the fact that a detailed criticism of the biased methods has already been provided in Poe, Severance-Lossin and Welsh (1994), but is useful to summarize here. We then turn to a discussion of an appropriate and simple empirical method of computing differences in distributions when the distributions are not independent.

In the paragraphs that follow, the text in parentheses after the title of each method indicates whether the method is intended for assessing the difference of independent or

joint distributions and whether the corresponding approach provides a generally unbiased or biased measure of the difference.

Empirical Convolutions Method (independent distributions, unbiased): The most immediate analogy to subtractive convolutions method presented in Equation 2 is a discrete, empirical convolutions approach as developed in Poe, Severance-Lossin and Welsh (1994). Letting "^" denote a discrete approximation of an underlying distribution, "min" identify the lowest possible value in a distribution, "max" identify the highest possible value in a distribution, and "Δ" represent a small increment, Equation (2) could be approximated by:

$$\hat{F}_{\hat{V}}(V') = \sum_{\min(\hat{X}-\hat{Y})}^{V'} \sum_{\min \hat{Y}}^{\max \hat{X}} \hat{f}_{\hat{X}}(v+y)\,\hat{f}_{\hat{Y}}(y)\Delta y \Delta v \qquad (4)$$

Implementation of this approach is computationally intensive, relying on convolutions programs that exist in existing software packages (e.g., Gauss) or requiring researchers to program the convolution themselves. Yet, this approach provides an approximate distribution of the difference that is exact up to the discontinuity imposed by the width of the increments used, and can be used to estimate the significance of the difference $\hat{X}-\hat{Y}$ by estimating $\hat{\alpha} = \hat{F}_{\hat{V}}(0)$. That is, as Δ approaches zero, the empirical convolution will approach the true difference of the two empirical distributions and $\hat{\alpha}$ will approach α. Unfortunately, increased precision dramatically raises computing power and time requirements, as will be demonstrated in Sections IV and V.

[4] Two-sided estimates of the difference can generally be obtained by doubling the one-sided significance level

Complete Combinatorial (independent distribution, unbiased): As suggested above, an alternative way of computing the convolution is to calculate every possible difference between the two empirical distributions, sort this difference, and create a cumulative distribution function. Letting I denote the number of observations in the simulated distribution X, and j denote the number of observations in the simulated distribution Y, then the distribution of the difference is given as:

$$\hat{X}_i - \hat{Y}_j \quad \forall i, j \tag{5}$$

The estimated i*j vector of the difference is then sorted, and the proportion of negative values is calculated to come up with an exact value for α. The Appendix provides a simple to do-loop routine for calculating this complete combinatorial approach.

Although exact, this approach is computationally intensive, requiring the capacity to store and sort extremely large vectors. For example, the typical number of bootstrap observations found in the applied economics literature appears to be 1,000 observations for each parameter. As such, a complete combination of two 1,000 element vectors results in a vector of one million (1,000*1,000) elements. While many econometric programs presently have the storage capacity to accommodate such large matrices, this capacity is quickly broached as i and j each increase beyond 1000 elements.

Repeated Sampling (independent distribution, unbiased): This approach builds upon the sampling method introduced in Ready, Buzby, and Hu (1996), which suggested that a random draw of 1000 paired differences be drawn from $\hat{X}$ and $\hat{Y}$. The Ready, Buzby and Hu approach was limited to a single set of differences. Here we suggest that the difference be conducted a number (N) of times and averaged across the N repetitions.

Letting $Rand_N(\hat{Y})$ indicate a random ordering of distribution $\hat{Y}$, then estimating the significance of the difference could be accomplished by randomizing the Y vector N times, calculating a subtraction $\hat{X}_i - (Rand_N(\hat{Y}))_i$ for each randomization, computing the number of negative difference values as a proportion of all differences, repeating this process for N random orderings, and calculating the average proportion of negative differences.

This approach has the advantage of simplicity, speed and minimum storage requirements relative to the complete combinatorial or convolutions methods. Nevertheless, by its very nature drawing a sample introduces sampling error, which, as we demonstrate below, could influence the determination of significance between two distributions.

Non-Overlapping Confidence Intervals (independent samples, biased): This technique, suggested in Park, Loomis and Creel (1991) and elsewhere, judges that two empirical distributions are significantly different at the α level of significance if their estimated individual (1-α) confidence intervals do not overlap. Although simple, Poe, Severance-Lossin and Welsh (1994) demonstrate that the significance level of this approach is overstated and hence biased. That is the true level of significance is smaller than $\hat{\alpha}$, with the degree of disparity depending on the shape of the individual distributions being compared.

Normality (independent samples, biased): While one could appeal to variations on the central limit theorem, the Slutsky theorems, and delta method approaches as a basis for assuming normality of parameters computed as functions of sample means (e.g.

Goldberger, 1991), in practice we have found that simulated distributions generally do not conform to normality[5]. Given this result and the basic motivation for estimating distributions using bootstrap and other resampling techniques, it seems odd, and unnecessary, to impose normality assumptions at this stage. To the extent that the normality assumptions do not correspond with the actual empirical distributions, the imposition of normality will lead to biased estimates of two distributions. In cases, where normality is found to hold, estimates of significance obtained via the unbiased techniques listed above should converge with normality-based approaches.

Paired Difference (jointly distributed samples, unbiased): When the estimated distributions are not independent, the statistical test is greatly simplified. As described in Poe, Welsh and Champ (1997), the estimate of the difference simply involves calculating

$$X_i - Y_j \quad \forall i = j. \tag{6}$$

This vector is then sorted and the significance of the difference is obtained by calculating the proportion of negative values as above.

The difficulty with this approach typically lies in generating the joint distribution. For bootstrapping or jackknifing procedures, this will involve providing a paired value of X and a paired value of Y estimated from each simulated data set. When a parametric (e.g., Krinsky and Robb., 1986) approach is used, then the correlation of the error term has to be accounted for when generating the values from the variance co-variance matrix. Similarly, Monte Carlo techniques need to account for the joint distribution in estimating paired values for X and Y.

IV. Independent Samples: Application to Known Distributions

[5] Efron (2000) similarly notes that these Delta Method techniques are still used "(sometimes unfortunately)

In this section we simulate two 1000 X 1 independent distributions from known single-parameter (c) Weibull distributions ($F_x(x)=1-e^{-x^c}$), adjusting the location of these distributions such that the two distributions would be approximately significantly different at the 5% level if they were normally distributed. We then compare the estimated significance of the difference and the time and memory required for each of the independent distribution techniques described above[6]. Throughout we use the complete combinatorial approach as the reference for the significance of the difference, as this method provides the exact difference of the two empirical distributions.

The upper portion of Table 1 provides the relevant statistics when each of the two distributions are approximately normal. This occurs when c=3.6 (Johnson and Kotz, 1970 p. 253). Under these conditions the significance level of the complete combinatorial method for this particular sample is 4.95, which would lead to a rejection of the equality of the two distributions using a significance level of α = 5%. This value is reasonably close to the corresponding significance level, 4.96, of the normality-based approach. Such a result is expected: the significance of the complete combinatorial method should converge with the normal-based approach if the underlying distributions approximate normality. Each of the convolution comparisons are similarly close to the combinatorial method, although the accuracy improves as the increment size diminishes. The mean of the 100 samples deviates somewhat from the complete combinatorial approach, with the average value (5.01) in this instance, marginally failing to reject the hypothesis of equality if an α=5% criteria is used. This is due to the fact that, although none of the samples are systematically biased, individual samples, as used in Ready,

---

even though we are now armed with more potent weaponry" (p. 1293).

Buzby, and Hu (1995), exhibit sampling error with $\hat{\alpha}$ ranging from 4.00 to 6.50. Hence some of the samples correctly reject the null hypotheses of equality between the two distributions at the 5% significance level, whilst other comparisons fail to reject the null at this level of significance. In contrast, adopting a non-overlapping confidence interval criterion which identifies the lowest $\hat{\alpha}\%$ at which the (1-$\hat{\alpha}$) confidence intervals do not overlap, is substantially biased. Using this method one would erroneously conclude that the two distributions are not significantly different at the 10% level.

The bottom portion of Table 1 repeats these method comparisons for two distributions exhibiting different levels of skewness. Here the exact difference of the two simulated distributions, as determined by the complete combinatorial approach, is 6.18%. This value is approximated by the small increment convolutions value. In contrast to the above discussion, the normality-based measure (4.93%) diverges from the actual significance of the difference as would be expected when the underlying normality assumptions are inappropriate. Importantly, the sampling method continues to exhibit sample error, with some estimated significance levels lying on the wrong side of the 5% level. Reflecting the previous analysis, the non-overlapping confidence interval criterion erroneously concludes that the two distributions are not significantly different at the 10% level.

Comparison of the entries in the last two columns indicate that each of the methods varies tremendously in terms of the computational time and memory required. Here we focus only on the unbiased methods, with the caveat that the biased methods are computationally efficient (but biased!). Within the category of unbiased methods the

---

[6] All calculations were performed on a Gateway Solo 9300, Pentium III 700MHz laptop.

mean of 100 samples requires negligible computational time and memory. In contrast, the complete combinatorial approach requires a tremendous amount of memory. The time and memory requirements for the empirical convolution vary with the width of the increments used. With increments of 0.001 the empirical convolution is efficient in terms of time and memory relative to the complete combinatorial approach. When finer increments are used (e.g., 0.0001) the time taken is substantially higher than the complete combinatorial method, although the memory required remains much lower.

In all, there appears to be a tradeoff between accuracy, time, and memory requirements across methods. Although computationally efficient, it is clear that the biased measures are indeed biased. Within the unbiased methods the repeated sampling method is relatively facile to program and computationally efficient, yet in instances when the actual difference of the two distributions is near a critical significance level, the sampling approach can lead to erroneous assessments of statistical significance because of sampling error. The empirical convolution method requires access to a convolutions program or sophistication in programming skills, can be computationally efficient if the increments are relatively coarse, and generally approximates the significance level complete combinatorial method. As greater precision is desired, say when the estimated value is close to a threshold significance level, this method requires increasing amounts of time and memory. As demonstrated in the Appendix, the complete combinatorial method is simple to program and provides and exact difference of two empirical distributions. However, it is time and memory intensive.

Having demonstrated these relationships in a controlled, Monte Carlo situation, we now turn to explore these relationships in field conditions.

V. Application to Internal and External Scope Tests in Dichotomous Choice

The Data

In order to test the hypothesis of significantly different WTP between an embedded good and a comprehensive good, data was obtained from two 1996 surveys of threatened and endangered species protection in the Southwestern United States. The comprehensive good is a program that would protect a set of 62 threatened and endangered (T&E) species. The embedded good is a program that would protect only one of those species, the Mexican spotted owl (MSO). The protection programs included designating Critical Habitat Units (CHUs) that totaled 4.6 million acres for the MSO and a total of 4.6 million acres plus 2,456 miles of rivers for the 62 T&E species. The protection program contained restrictions on human activities such as timber harvesting and dam operations within the units. The CHUs are located on the Colorado Plateau (Southwestern Colorado, Southern Utah, Northern New Mexico and Arizona). Other parts of the protection programs included scientific research and habitat improvement. For more information on the species and the protection programs, see Giraud, *et al*. (1999).[7]

Two survey treatments containing two dichotomous choice CVM questions were employed. The survey treatments were identical with the exception of the CVM question ordering. In other words, one treatment asked about the MSO program first, and then the 62 T&E species (what we shall refer to as the "bottom-up" format) while the other treatment reversed the order of the questions ("top-down" format). In total, 1600 surveys

were mailed to individuals across the United States, 800 for each survey treatment. Multiple mailings resulted in 383 returned surveys from the bottom-up sample and 369 returned surveys from the top-down sample. When eliminating the undeliverables, this is a 54% response rate.

The two dichotomous choice questions in each of the surveys are provided in Figure 1. Fourteen different bid amounts were used ($1, 3, 5, 10, 20, 15, 30, 40, 50, 75, 100, 150, 200, and 350). The bid amounts were based on a number of focus groups, pretests, and past Spotted Owl CVM studies. The bid amounts were systematically assigned to each survey, and each individual received the same bid amount for both WTP questions in a given survey.

Reflecting the concern that hypothetical responses over-state actual willingness to pay, each respondent was further asked the following payment certainty question:

> *On a scale of 1 to 10, how certain are you of your answer to the previous question? Please circle the number that best represents your answer if 1 = not certain and 10 = very certain.*
>
> *1 2 3 4 5 6 7 8 9 10*
> *not certain* <----------> *very certain*

Champ *et al.* (1997) have argued that this type of certainty question allows the researcher to calibrate hypothetical responses by recoding respondents who answered 'yes', but indicated relatively low levels of certainty on the follow-up certainty question as 'no's'.[8]

---

[7] The Giraud *et al.*(1999) paper differs form the present analyses in that no correction for certainty was used and the significance tests of internal scope failed to account for the correlation in errors in a bivariate analysis of dichotomous choice responses to the MSO and the 62 T&E questions.

[8] Champ et al. (1997) further argue that such a certainty question provides an appropriate lower bound on willingness to pay. However, Chilton and Hutchinson (1999) have demonstrated that this is not necessarily the case. Because the comparison of hypothetical and actual contributions for public goods is clouded by the fact that under-revelation is likely in the voluntary contributions mechanisms utilized in Champ et al.

Field validity tests of contingent valuation indicate that certainty levels of '10 and higher' (Champ *et al*., 1997), '8 and higher' (Champ and Bishop, 2001), and '7 and higher' (Poe et al. 2001), to be appropriate calibration levels for adjusting hypothetical willingness to pay down to actual donation levels. In a series of papers, Blumenschein *et al.* (1998) and Johannesson *et al.* (1998, 1999) similarly demonstrate that respondents who are fairly to definitely sure that their hypothetical 'yes' response would be a real 'yes' response most closely predicts responses in actual money situations.

The Models

Table 2 contains the maximum likelihood estimation output. Both independent (Probit) and joint (bivariate Probit) models were estimated in order to investigate differences between bottom-up and top-down effects. The endogenous variable in each regression is willingness to pay for a protection program, 1 if 'yes' and 0 if 'no'. Exogenous variables include a *Constant*, *TEKnow*, which is a combination of three knowledge-holding questions, results of opinion questions *Protect* and *ProJob* and the *Bid*. The knowledge holding response values range from 0 to 3, depending on how much the respondent had read or heard about various threatened and endangered species[9]. The Protect and ProJob variables combined opinion questions regarding resource extraction

---

(1997) and Champ and Bishop (2001), Poe et al. (2001) have argued that it is likely that such comparisons lead to over calibration, and, hence, underestimate true willingness to pay. As a result, Poe *et al.* suggest that lower levels of certainty in hypothetical payment should be used to calibrate hypothetical values to actual contributions. The use of a certainty level of '7 or higher' in this paper reflects this more conservative approach to calibration.

[9] TEKnow is the sum of three dummy variables in which YES = 1 and NO = 0. The three questions are as follows: 1. Have you read or heard about threats to the Mexican Spotted Owl in the Southwestern United States? 2. Have you read or heard about threats to the Northern Spotted Owl in the Northwestern United States? 3. Have you read or heard about threatened and endangered fish in the Colorado River?

and endangered species protection.[10] Respondents that valued jobs and resource extraction have a higher number for ProJob. Respondents that valued species protection have a higher number for Protect. The models were estimated as described in Poe, Welsh and Champ (1997), wherein the dependent variable was coded as a '1' if the respondents answered "yes" to the dichotomous choice bid variable and indicated a follow up certainty level of '7 or higher'. Otherwise, the dependent variable was coded '0'. This level of certainty corresponds was chosen because a hypothetical/actual validity test reported in Poe *et al.* (2001) found that a certainty level of '7 or higher' best corresponded with actual participation rates. Descriptive statistics for the variables used are provided in Table 2, with the maximum likelihood estimation results provided in Table 3.

The estimated coefficients are all significant and of the expected sign. TEKnow and Protect were positively correlated with willingness to pay. ProJob had a negative coefficient as did the dichotomous choice bid amount, indicating that the probability of a 'yes' response declines as the bid level increases. The correlation coefficient ρ, was positive and significant, demonstrating that the error terms of the 62 species and MSO response functions are positively correlated. This can be interpreted that the goods are regarded as substitutes. In comparison to previous research on multiple dichotomous questions in the same questionnaire, these correlation coefficients are relatively high

[10] *Protect* and *Projob* contain information from 6 of Likert-scale questions that ranged from 1 = Strongly Agree to 5 = Strongly Disagree. Questions i) and v) below were added together and multiplied by -1 to form ProJob. Questions ii), iii), iv), and vi) below were added together and multiplied by -1/2 to form Protect. The individual questions underlying these two variables are: i) "Businesses should be allowed to extract natural resources from Federal lands"; ii) "All species endangered due to human activities should be protected from extinction whether or not they appear important to human well being"; iii) "Plants and animals have as much right as humans to exist"; iv) "I am glad that the endangered species in the Four Corners Region are protected even if I never see them"; v) "If any jobs are lost, the cost of protecting a

(Alberini, 1995; Alberini *et al.,* 1997; Poe, Welsh and Champ, 1997). In part we attribute this to the fact that in this questionnaire the same dichotomous choice bid levels was asked of individual respondents for both levels of species protection, whereas previous research has varied this value across questions for the same individual. In addition, it is also likely that individuals view protecting MSOs and all 62 T&E Species as very similar commodities.

Table 3 reports the estimated mean willingness-to-pay values using the non-negative mean (Hanemann, 1984, 1989) and Krinsky and Robb (1986) parametric bootstrapping procedures. As depicted, the independent and joint distributions provide fairly similar results, with the only notable difference being that the joint distributions tend to be less dispersed. As such there are little efficiency gains from adopting a bivariate probit approach to estimate the individual mean willingness to pay distribution, a result that is similarly found in previous research (Alberini and Kanninen, 1994; Alberini *et al.*, 1997; Poe, Welsh, and Champ, 1997)

Table 4 extends the previous comparisons of methods to an external scope test in which the first values elicited in a questionnaire are compared across the top-down and bottom-up formats. That is, the MSO value, when it is asked first in the bottom-up question format, is compared to the 62 Species in the top-down format. To take advantage of additional information from the joint estimation, the values derived from the joint model were used in these comparisons. Each of the unbiased methods marginally reject the null hypothesis of equality between the mean willingness-to-pay values for the MSO and the 62 T&E Species protection levels and, hence, indicate scope sensitivity,

---

Threatened or Endangered Species is too large". vi) "Protection of Threatened and Endangered Species is a

although some individual samples in the mean of 100 samples approach would fail to reject the null hypothesis of equality because of sampling error. It should be noted that the normality-based approach and the non-overlapping confidence interval approach erroneously fail to reject the null hypothesis of equality between these two distributions.

The time and memory requirements also exhibit the same patterns as discussed previously. The mean of 100 samples approach is relatively efficient, requiring nominal time and memory. The complete combinatorial method requires substantial memory and time. The convolutions approach is fairly efficient in terms of time and memory for relatively wide increments (1 to 0.1), but the memory and time increase rapidly as the increment size declines. However, it should be noted that reasonable accuracy is obtained at moderate increment sizes.

Table 5 provides a comparison of independent and joint distribution tests of internal scope. That is, it compares the 62 Species and the MSO values within the bottom-up and top-down question formats. The first column of comparisons uses the independent estimates and a complete combinatorial approach to estimating the differences. The second column of comparison assumes that the two distributions are jointly distributed, and accounts for the correlation estimating the difference of the distributions. In this latter approach substantial efficiency gains are found by accounting for the correlated error term through application of Equations 3 and 5. As demonstrated, this paired samples approach decreases $\hat{\alpha}$ (i.e., increases the estimated significance of the difference) substantially. However, in this case accounting for correlation does not change the assessment of whether the null hypothesis of equality can be rejected. In the

---

responsibility I am willing to pay for".

bottom-up comparison, the null cannot be rejected in either the independent or the joint comparison. In the top-down version, the null is rejected regardless of method. It is interesting to note that this directional asymmetry in scope sensitivity has similarly been observed in laboratory experiments, and has been attributed to other-regarding and strategic behavior, gains-loss asymmetry, and other psychological motives (see Bateman *et al*., 2001).

VI. Conclusion

This paper has described and demonstrated methods that provide unbiased estimates of the significance of difference of two distributions. In situations where the two distributions are not independent, appropriate tests that rely on "paired" differences can easily be implemented. In such cases, the primary difficulty in implementing this method lies in generating the paired observations in a way that accounts for correlation in responses or in errors.

However, when the two distributions to be compared are independent, a number of unbiased options exist for assessing the statistical difference between two distributions. We have demonstrated that there are tradeoffs between these alternative methods, and it is likely that the benefits and costs of employing each method will vary by researcher and by situation[11]. Nevertheless, the following generalizations can be

[11] The following provides an example of a situational issue. Suppose the researcher is interested in combining estimates of differences and comparing them to some other value. For example, the adding up test for two different goods ((WTP A) + WTP(B) = WTP(A + B))frequently raised in contingent valuation discussions (e.g., Diamond, 1996) would require such a computation. In such instances the complete combination approach becomes quite unwieldy, as an additive complete combination of, say, 1,000 observations for WTP(A) and 1,000 observations for WTP(B), would result in a distribution of WTP(A) + WTP(B) consisting of a million observations. The comparison of this vector with a 1000X1 vector WTA(A+B) will thus involve a billion calculations. In contrast, the memory and calculations

made. For independent samples, if sufficient memory is available, the complete combinatorial approach provides an exact measure of the difference of two distributions, and would, on this basis, seem to be the preferred option. This method is also very easy to program. The more difficult to program empirical convolutions approach also provides precise estimates of the difference, provided that the increments used to approximate the distribution are relatively fine. Whereas the complete combinatorial method and the empirical convolutions method each require substantial memory, averaging 100 randomly paired differences of the distribution is computationally efficient. Yet, if the exact difference of the distributions is proximate to a significance threshold, sampling error introduces the possibility of a false rejection/acceptance of the null hypothesis.

---

associated with the convolutions approach depends simply on the spread of the distributions to be compared and the size of the increment used, and is likely to be computationally more efficient in evaluating such an equality. Alternatively, a research might adopt a 'short cut' method by sorting each of the possible vectors and identifying the subset of possible combinations that sum to zero or less.

References

Alberini, A. "Efficiency vs Bias of Willingness-to-Pay Estimates: Bivariate and Interval-Data Models." *J. of Environ. Econ. and Manag.* 29(2 1995): 169-180.

Alberini, A., M. Cropper, T-T Fu, A. Krupnick, J-T Liu, D. Shaw, and W. Harrington. "Valuing Health Effects of Air Pollution in Developing Countries: The Case of Taiwan." *J. of Environ. Econ. and Manag.* 34(2 1997): 107-126.

Alberini A. and B. Kanninen. "Efficiency Gains from Joint Estimation: When Does a Second Question Improve Estimation of the First." Unpublished paper presented at the Annual Meeting of the Western Econ. Assoc. (1994), San Diego.

Bateman, I. J., P. Cooper, M. Cole, S. Georgiou, D. Hadley and G. L. Poe. "An Experimental and Field Test of Study Design Effects upon Nested Contingent Values." *CSERGE Working Paper* (2001), Univ. of East Anglia, UK.

Bateman, I., A. Munro, B. Rhodes, C. Starmer and R. Sugden. "Does Part-Whole Bias Exist? An Experimental Investigation." *Econ. J.* 107(441 1997): 322-332.

Berrens, R., A. Bohara, and J. Kerkvleit. "A Randomized Response Approach to Dichotomous Choice Contingent Valuation." *Am. J. of Agr. Econ.* 79(1 1997): 252-66.

Blumenschein, K., M. Johannesson, G. C. Blomquist, B. Liljas, and R. M. O'Connor, "Experimental Results on Expressed Uncertainty and Hypothetical Bias in Contingent Valuation." *Southern Econ. J.*, 65(1 1998): 169-177.

Champ, P. A. and Bishop, R. C. "Donation Payment Mechanisms and Contingent Valuation: An Empirical Study of Hypothetical Bias." *Env. and Res. Econ.* 19(4 2001): 383-402.

Champ, P. A., R. C. Bishop, T. C. Brown and D. W. McCollum. "Using Donation Mechanisms to Value Nonuse Benefits from Public Goods." *J. of Environ. Econ. and Manag.* 33(2 1997): 151-62.

Chilton, S. M. and Hutchinson, W. G. "Some Further Implications of Incorporating the Warm Glow of Giving into Welfare Measure: A Comment on the Use of Donation Mechanisms by Champ *et al.*" *J. of Environ. Econ. and Manag.* 37(2 1999): 202-209.

Diamond, P. "Testing the Internal Consistency of Contingent Valuation Surveys." *J. of Environ. Econ. and Manag.* 30(3 1996): 337-347.

Efron, B. "The Bootstrap and Modern Statistics." *J. Am. Stat. Soc.* 95(452 2000): 1293-1296.

Efron, B. and R. J. Tibshirani, *An Introduction to the Bootstrap.*(New York, Chapman and Hall, 1993).

Goldberger, A. S., *A Course in Econometrics*, (Cambridge, Harvard Univ. Press, 1991).

Giraud, K., J. Loomis and R. Johnson. "Internal and External Scope in Willingness to Pay Estimates for Threatened and Endangered Wildlife." *J. Env. Manag.* 56(1999): 221-29.

Green, R., W. Hahn, and D. Rocke. "Standard Errors for Elasticities: A Comparison of Bootstrap and Asymptotic Standard Errors." *J. Bus. Econ. Stats.* 5(1987): 145-149.

Hannemann, W. "Welfare Evaluations in Contingent Valuation Experiments With Discrete Responses." *Am. J. Agr. Econ*. 66(1984): 332-341.

Hannemann, W. "Welfare Evaluations in Contingent Valuation Experiments With Discrete Response Data: Reply." *Am. J. Agr. Econ.* 71(1989): 1057-1061.

Hutchinson, W. G., R. Scarpa, S. M. Chilton, and T. McCallion, "Parametric and Non-Parametric Estimates of Willingness to Pay for Forest Recreation in Northern Ireland: A Discrete Choice Contingent Valuation Study with Follow-Ups." *J. of Agr. Econ*. 52(1 2001): 104-122.

Johannesson, M., K. Blumenschein, P.-O. Johansson, B. Liljas, and R. M. O'Conor, "Calibrating Hypothetical Willingness to Pay Responses." *J. of Risk and Uncert.* 8(1999): 21-32.

Johannesson, M., B. Liljas, and P.-O. Johansson. "An Experimental Comparison of Dichotomous Choice Contingent Valuation Questions and Real Purchase Decisions." *Applied Econ*. 30(1998): 643-647.

Johnson, N. L. and S. Kotz, *Continuous Univariate Distributions -1*. (New York, John Wiley and Sons, 1970).

Kling, C. "Estimating the Precision of Welfare Measures." *J. Env. Econ. Manag.* 21(3 1991): 244-259

Kotchen, M. J., and S. D. Reiling, "Do Reminders of Substitutes and Budget Constraints Influence Contingent Valuation Estimates? Another Comment." *Land Econ.* 75(3 1999): 478-482.

Krinsky, I. and A. L. Robb "On Approximating the Statistical Properties of Elasticities." *Rev. Econ. and Stat*. 68(1986): 715-719.

Krinsky, I. and A. L. Robb, "Three Methods for Calculating the Statistical Properties of Elasticities: A Comparison." *Empirical Econ.* 16(1991): 199-209.

Li, H., and G. S. Maddala. "Bootstrap Variance Estimation of Nonlinear Functions of Parameters: An Application to Long-Run Elasticities of Energy Demand." *Rev. of Econ. and Stats.* 81(4 1999): 728-733.

Loomis, J., M. Creel, and T. Park, "Comparing Benefit Estimates from Travel Cost and Continent Valuation Using Confidence Intervals for Hicksian Welfare Measures." *Applied Econ.* 23(1991): 1725-1731.

Mood, A. M., F. A. Graybill, and D. C. Boes, *Introduction to the Theory of Statistics (Third Edition)*. (New York, McGraw-Hill Book Company, 1974).

Park, T., J. Loomis, and M. Creel. "Confidence Intervals for Evaluating Benefits from Dichotomous Choice Contingent Valuation Studies." *Land Econ.* 67(1 1991): 64-73.

Poe, G. L., J. E. Clark, D. Rondeau, and W. D. Schulze. **"**Provision Point Mechanisms and Field Validity Tests of Contingent Valuation." *Envir. and Res. Econ.* (2001): forth.

Poe, G., E. Severance-Lossin, and M. Welsh. "Measuring the Difference (X-Y) of Simulated Distributions: A Convolutions Approach." *Am. J. of Agr. Econ.* 76(4 1994): 904-15.

Poe, G., M. Welsh and P. Champ. "Measuring the Difference in Mean Willingness to Pay When Dichotomous Choice Contingent Valuation Responses Are Not Independent." *Land Econ.* 73(2 1997):255-67.

Ready, R., J. Buzby, D. Hu. "Differences Between Continuous and Discrete Contingent Value Estimates." *Land Econ.* 72(3 1996):397-411.

Reaves, D. R., Kramer, and T. Holmes, "Does Question Format Matter? Valuing an Endangered Species." *Env. and Res. Econ.* 14(1999): 365-383.

Rollins, K. and A. Lyke. "The Case for Diminishing Marginal Existence Values." *J. Env. Econ. Manag.* 36(3 1998):324-44.

Weninger, Q. "An analysis of the Efficient Production Frontier in the Fishery: Implications for Enhanced Fisheries Management." *Applied Econ.* 33(2001): 71-79.

Appendix: Simple Do-Loop Approaches for Estimating Differences in Distributions (in Gauss)

Independent Distributions: Let vecthigh and vectlow denote the two independent vectors, respectively, with vecthigh having high values relative to vectlow. Both vectors have n elements. The objective is to calculate the difference vecthigh – vectlow.

The following provides a Gauss code for estimating the complete combinatorial. The notes between @ provide a verbal description of the corresponding command.

```
@compute all possible combinations of the difference: 1,000, 000 by 1 vector for two 1000 by 1 vectors @
Let i=1;                          @set the beginning value for the do loop@
  do while i le n;                @set the maximum value for the do loop@
vectdiff= vecthigh[i] - vectlow;  @calculate the difference between the ith element of vecthigh and the
                                    entire distribution of vectlow@
    if i lt 2;
       vdvect = vectdiff;         @establishes the first difference as a vector@
      else;
       vdvect=vdvect|vectdiff;    @adds additional calculations to the existing vector when i is greater
                                  than 1@
     endif;
   i=i+1;
endo;
```

The vector can then be sorted and simple indexing programs can be used to identify cumulative distributions values associated with a difference of zero, which provides the one-sided significance level of the difference.

Correlated distributions: Using the same notation as in the independent samples case, the do loop is replaced by the following:

vectdiff = vecthigh – vectlow;

Table 1: One-Sided Significance Levels of the Difference of Weibull Distributions, Computing Time, and Memory Used for Alternative Methods

Weibull Distributions[a]		Method	Significance % (")	Time (secs)	Memory (bytes)
Dist. 1	Dist. 2				
C=3.6[b]	c=3.6[b]	Complete Combination[c]	4.95	56.03	~8,000k
		Convolution			
		) = 0.001[d]	4.97	0.55	~357k
		) = 0.0001[d]	4.95	159.72	~3,570k
		Mean of 100 Samples[e] [Range]	5.01 [4.00-6.50]	<0.01	~24k
		Normal[f]	4.96	<0.01	~24k
		Overlapping Confidence Intervals[g]	12.70	<0.01	~16k
C=2.0[b]	c=4.0[b]	Complete Combination[c]	6.18	63.17	~8,000k
		Convolution			
		) = 0.001[d]	6.19	0.82	~409k
		) = 0.0001[d]	6.18	220.03	~3,775k
		Mean of 100 Samples[e] [Range]	6.10 [4.80,-7.30]	<0.01	~24k
		Normal[f]	4.93	<0.01	~24k
		Overlapping Confidence Intervals[g]	12.30	<0.01	~16k

Notes:

a. Weibull distribution in which standard cumulative distribution is specified as $F(x) = 1 - e^{-x^c}$ (Johnson and Kotz, 1970).
b. For c=3.6 the standard Weibull distribution is distributed approximately normal, with mean, skewness parameter, and standard error of 0.9011, 0.00, and 0.2780, respectively. For c=2 the distribution is positively skewed with a mean, skewness parameter, and standard error of 0.8862, 0.63, 0.4663, respectively. For c=4 the distribution is slightly negatively skewed with a mean, skewness parameter, and standard error of 0.9064, -0.09, 0.2543, respectively.
c. The significance level using this method is taken to be the actual difference, and, hence, serves as a reference for evaluating the accuracy of alternative approximations.
d.) indicates the increment size for the convolution approximation.
e. Each sample taken from a random reordering of each distribution Time and memory pertain to one sample.
f. Assumes that each distribution is normally distributed.
g. Identifies the lowest α% significance levels at which the (1-α%) confidence intervals that do not overlap.

Table 2: Descriptive Statistics

Variable Name	Variable Description	Mexican Spotted Owl (MSO) then 62 Species, Certainty =7 and Higher (Bottom-Up) Mean (standard deviation)	62 Species then Mexican Spotted Owl (MSO) Certainty = 7 and Higher (Top-Down) Mean (standard deviation)
TEKnow	0 to 3 scale of respondent knowledge of threatened and endangered species. 3 = high.	1.89 (0.99)	1.92 (1.03)
ProJob	-10 to -1 scale from two Likert-scale questions relating to endangered species protection and jobs. -10 = high concern for jobs.	-6.70 (2.02)	-6.54 (2.18)
Protect	-10 to -1 scale from four Likert-scale opinion questions about endangered species protection. -10 = high concern for species protection.	-4.85 (2.17)	-5.02 (2.24)
Bid	Dichotomous choice bid value.	78.81 (98.93)	73.25 (95.44)
Certainty Level MSO	1 to 10 scale for follow-up certainty question. 10 = very certain.	7.95 (1.91)	7.94 (2.12)
Proportion of Yes Responses, MSO (Certainty = 7 or higher)	Proportion of respondents who responded yes to the MSO dichotomous choice question and indicated a certainty level of 7 or higher	0.37 (0.48)	0.38 (0.49)
Certainty Level 62 Species	1 to 10 scale for follow-up certainty question. 10 = very certain.	8.06 (1.83)	8.11 (1.99)
Proportion of Yes Responses, 62 Species (Certainty = 7 or higher)	Proportion of respondents who responded yes to the 62 species dichotomous choice question and indicated a certainty level of 7 or higher.	0.38 (0.49)	0.45 (0.50)

Table 3: Mexican Spotted Owl (MSO) and 62 Species – Independent Probit and Joint Bivariate Probit Models

	MSO then 62 Species, Certainty =7 and Higher (Bottom-Up)		62 Species then MSO Certainty = 7 and Higher (Top-Down)	
	Independent	Joint	Independent	Joint
<u>62 Species</u>				
Constant	-0.3292	-0.3199	0.3353	0.3017
	(0.4842)	(0.4970)	(0.4299)	(0.4322)
TEKnow	0.2349	0.2363	0.0159	-0.0009
	(0.0848)***	(0.0961)**	(0.0781)	(0.0796)
Protect	0.2867	0.2913	0.3031	0.3027
	(0.0460)***	(0.0405)***	(0.0433)***	(0.0457)***
Projob	-0.1746	-0.1757	-0.1953	-0.2031
	(0.0485)***	(0.0531)***	(0.0444)***	(0.0517)***
Bid	-0.00437	-0.00435	-0.00455	-0.00453
	(0.0009)***	(0.0007)***	(0.0010)***	(0.0009)***
<u>MSO</u>				
Constant	-0.2236	-0.2394	-0.184	-0.1483
	(-0.4835)	(-0.4900)	(0.4284)	(0.3786)
TEKnow	0.2616	0.2642	0.0224	0.0034
	(0.0847)***	(0.0951)***	(0.0786)	(0.0772)
Protect	0.2717	0.2682	0.2485	0.2534
	(0.0885)***	(0.0391)***	(0.0424)***	(0.0423)***
Projob	-0.1367	-0.1352	-0.2016	-0.2024
	(0.0477)***	(0.0505)***	(0.0438)***	(0.0430)***
Bid	-0.00437	-0.00451	-0.00553	-0.00535
	(0.0009)***	(0.0008)***	(0.0011)***	(0.0010)***
D		0.989		0.943
		(0.008)***		(0.024)***
Likelihood Ratio P^2_1	125.40***		136.94***	
Likelihood Ratio P^2_2	112.57***		120.08***	
- 2*Log Likelihood[a]	444.63+438.46***	422.72***	448.79+432.10***	479.62***
N	334	334	326	326

Note: Numbers in ( ) are asymptotic standard errors. *, **, and *** indicate significance levels of 0.10, 0.05 and 0.01, respectively.

[a] $-2\ (LL_i - LL_{j,u}) = 401.27$ for the bottom-up format and 460.37 for the top-down format. $P^2_{1,\ 0.10} = 2.71$.

Table 4: Estimated Mean WTP Distributions

	Data 1000, MSO then 62 Species, Certainty =7 and Higher (Bottom-Up)		Data 2000, 62 Species then MSO Certainty = 7 and Higher (Top-Down)	
	Independent	Joint	Independent	Joint
62 Species	79.16 [60.80,108.90]	79.30 [63.76, 104.70]	101.15 [78.49, 143.51]	99.82 [78.98, 136.84]
MSO	72.35 [56.63, 102.53]	73.80 [58.79, 97.37]	65.45 [51.22, 91.52]	65.95 [52.90, 86.79]

Note: Numbers in [ ] are 90 percent confidence intervals.

Table 5: One-Sided Significance Levels of External Test, Computing Time, and Memory Used for Alternative Methods

Distributions		Method	Significance % (")	Time (secs)	Memory (bytes)
MSO	62 Species	Complete Combination[a]	9.56	65.09	~8,000k
		Convolution			
		) = 1[b]	9.92	0.06	~100k
		) = 0.1[b]	9.59	0.99	~543k
		) = 0.01[b]	9.56	287.04	~5,440k
		Mean of 100 Samples[c] [Range]	9.64 [8.20-11.10]	<0.01	~24k
		Normal[c]	14.97	<0.01	~24k
		Overlapping Confidence Intervals[e]	16.50	<0.01	~16k

Notes:

a. The significance level using this method is taken to be the actual difference, and, hence, serves as a reference for evaluating the accuracy of alternative approximations.
b.) indicates the increment size for the convolution approximation.
c. Each sample taken from a random reordering of each distribution Time and memory pertain to one sample.
d. Assumes that each distribution is normally distributed.
e. Identifies the lowest x% significance levels at which the (1-x%) confidence intervals that do not overlap.

Table 6: Significant Levels of Internal Scope Tests

Comparison	$\hat{\alpha}_{\text{Independent}}$	$\hat{\alpha}_{\text{Joint}}$
Bottom-Up: MSO and 62 Species	36.8	28.7
Top-Down: MSO and 62 Species	4.9	0.9

Figure 1: Text of Dichotomous Choice Contingent Valuation Questions

Mexican Spotted Owl (MSO) Question[a]

If the Mexican Spotted Owl Recovery Federal Trust Fund was the only issue on the next ballot and would cost your household $___ every year, would you vote in favor of it? (Please circle one)

YES NO

62 Threatened and Endangered Species (T&E) Question[a]

If the Four Corners Region Threatened and Endangered Species Trust Fund was the only issue on the next ballot and it would cost your household $___ every year, would you vote in favor of it? (Please circle one)

YES NO

[a] The titles for each question are used for identification here, and were not included in the questionnaire.