Rural development

C

1981

AN EMPIRICAL EXAMPLE OF NONPARAMETRIC ANALYSIS IN RURAL DEVELOPMENT RESEARCH

by

Don Blayney, Research Associate

and

Gerald Marousek, Agricultural Economist

Dept. of Agric. Econ. and Applied Stat.

University of Idaho

Moscow

A. E. Research Series No. 232

Presented at the annual meeting of the Western Agricultural Economics Association, University of Nebraska-Lincoln, July 20, 1981.

An Empirical Example of Nonparametric Analysis in Rural Development Research

ABSTRACT

Agricultural economics research may pose problems of how to use limited, diverse information to draw inferences for larger populations. Nonparametric statistical procedures are useful when data cannot be analyzed by parametric means. A rural development problem illustrates nonparametric techniques for testing a distributional hypothesis and establishing service cost confidence bounds.

An Empirical Example of Nonparametric Analysis in Rural Development Research

INTRODUCTION

Many research situations exist where the commonly used parametric analysis is not applicable. One situation is the estimation of the impact of private sector growth on local governments. A goal in all research is to obtain data which is useful in both the local situation and in a more generalized setting. Obtaining the data to measure impacts requires detail not often available from centralized sources. Thus the case study approach is used. Case studies recognize the diversity of the community in several respects, such as population, data format, local service mix, and location, but they are relatively costly in terms of time and money. As a result, it is often not feasible to carry on a large number of such studies.

Information obtained as described is clearly useful in local situations but should also be generalized to a larger setting. The question is: How do we use very limited (small sample), diverse information to draw inferences for a larger set of situations? Nonparametric techniques provide an alternative for analyzing this type of data. The purpose of this paper is threefold: 1) to elaborate on the applicability of nonparametric analysis; 2) to report the results of such an application and 3) to discuss the inferences drawn.

WHY NONPARAMETRICS?

Most parametric techniques are grounded in the theory and assumptions of the normal probability distribution. The normal curve is a symmetric continuous distribution completely described by two parameters, the mean μ and variance σ^2. The familiar "bell-shaped" or normal curve at this point represents a population. When the analyst moves into sampling, and thus into estimating population parameters by $\hat{\mu}$ and $\hat{\sigma}^2$, another key factor comes into play. Large samples ($n \geq 25$) imply that normal conditions are met, thus implying the use of parametric approaches. The mechanics of and detailed theory for parametric analysis are described in most basic texts [Lentner; Snedecor & Cochran; Steel & Torrie]. The significant point to remember is that for inferential purposes, techniques such as analysis of variance (AOV), regression, and correlation analysis are appropriate only if normality assumptions are met.

The case for nonparametric techniques can be made point by point by recognizing the situations where normality assumptions are untenable. Many distributions are nonsymmetric and discrete, including χ^2, the F (the ratio of two χ^2) and the Poisson. Such distributions are only approximated by normal distributions and then only if the large sample assumption has been met. Other distributions may occur in nature that are too complex to easily specify at all. Sampling may not always occur in the methods required to obtain good estimates of population parameters and the sample may not be large enough to invoke the large sample properties of parametric procedures. Finally, the economist in particular has found that the assumption of constant variance is often violated (heteroskedasticity). Recognizing the pitfalls, both theoretical and experimental (empirical), statisticians developed procedures for making inferences which are not based on rigid assumptions. Information based on signs of differences, ranks of measures and counts of events or observations were found to meet this need [Mosteller & Rourke, p. 1].

By convention, the term nonparametrics defines two types of procedures: 1) the truly nonparametric techniques and 2) the distribution-free techniques [Daniel, p. 15]. The term "sturdy" has been suggested as an alternative descriptive term referring to the statistics' ability to stand up well under adverse conditions or the failure of assumptions [Mosteller & Rourke, p. 1].

To summarize, Daniel [Daniel, p. 16] lists the following situations when the nonparametric (sturdy) procedures are appropriate:

1. If the hypothesis to be tested does not involve a population parameter.
2. If the data is measured on a scale weaker than that required for a parametric test.
3. If assumptions to validate parametric procedures cannot be met.
4. If results are needed in a hurry and must be done by hand.

A major disadvantage is that the use of nonparametric techniques when parametric analysis is possible generally wastes information, i.e. they are less powerful. Also, although they have a reputation for requiring simple arithmetic computation, calculation of nonparametric statistics can be tedious [Daniel, p. 16].

The situations where nonparametrics are appropriate are usually easy to define. The following section describes a specific research project which was characterized by properties indicating nonparametric analysis was in order.

STUDY SETTING

The decade of the 1970's was marked by a dramatic shift in population to the Sunbelt and the West, the location of many of the nonmetropolitan areas of the country. Idaho is one of the most rural states in the nation and, according to 1980 preliminary census figures, one of the fastest growing with a 32.3% increase in population from 1970 to 1980. A study was undertaken to answer the question: What were the fiscal impacts of growth (development) on the municipal governments in small rural communities in Idaho? "Small" was defined as a population of 10,000 or less, the dominant size class as illustrated in Table 1.

Table 1. Distribution of Idaho Cities by Population Classes: Preliminary 1980 Census Data.

Population Classes	No. of Cities	Population by Class	Population Percentage by Class
greater than 50,000	1	102,125	10.8
25,000 to 50,000	5	164,777	17.5
10,000 to 25,000	5	75,676	8.0
less than 10,000	187	234,622	24.9
Subtotal	198	577,200	61.2
Unincorporated & open space		365,934	38.8
Total	198	943,134	100.0

Source: Division of Economic and Community Affairs, Governor's Office, State Capitol Bldg., Boise, Idaho.

Idaho's cities are diverse in many respects other than population. The various natural resource-based economies (agriculture, lumber, mining, recreation) make each community distinct. Location plays a large role in the communities' life, as do social institutions and attitudes. The format for community data is not uniform, a problem that makes comparisons tenuous at best. These factors suggest the case study as an appropriate methodology for analyzing communities. As stated in the intoduction, the case study is a method that allows the type of detailed study necessary to isolate the impacts of development, particularly the cost and revenue relationships.

The marginal additions to public service systems resulting from development and the outlays associated with them are important "cost" determinants. Property taxes and service user fees constitute the bulk of the revenue from the development. While the revenue side of the relationship is fairly well defined costs are less so, particularly those for operations and maintenance (O & M). Estimates of these costs for particular public services (water, sewer, streets and roads) were required in order to estimate the net fiscal impacts in Idaho communities. It is these estimates that are the primary focus of the following statistical analysis.

APPLYING NONPARAMETRIC TECHNIQUES

Two nonparametric techniques were applied using data obtained from 10 case study communities in Idaho. The relevant data are presented in Table 2. The premise that the array of Idaho cities, with respect to population, is not normally distributed is examined first. Inferences relating to the estimates of operations and maintenance (O & M) costs are the results of the second analysis.

There is a nonparametric technique, in the distribution-free sense, that can be used to test sample data to determine if it could have come from a specified theoretical distribution. This is the Kolmogorov-Smirnov goodness of fit test applied to one sample. The test is designed to make use of a theoretical distribution entirely specified a priori. Using the data in Table 1, a hypothetical normal distribution was specified with a mean of 2915 and a standard deviation of 9135. The sample data were tested using the SPSS NPAR procedures [Hull and Nie, pp. 66-98]. The computed test statistic was 1.348. The larger this value, the less likely that the observed and hypothetical distributions are the same. The rejection rule is that if the calculated test statistic is greater than the table value for a given α and n, then the null hypothesis that the data come from the hypothesized distribution is rejected. The table value for $n = 10$ and $\alpha = .10$ is .369 therefore the null hypothesis was rejected and it was concluded that the sample data did not come from the hypothetical normal distribution. The thesis that data collected from cities in Idaho is not normally distributed has gained a measure of empirical support.

The conclusion that the distribution of city population sizes is not normal doesn't imply that other data would not be normally distributed. The test just described could be used on other variables if the necessary information existed to specify a hypothesized distribution of data. If only the sample data were used to determine those parameters the test becomes very conservative, i.e., the null hypothesis is rejected less often. Only sample data were available for the other variables so they were not analyzed by this method. The small sample size and the lack of strong evidence regarding the distributions of the operations and maintenance (O & M) cost data support the use of the following nonparametric analysis.

The O & M costs were found to be of most significance to local leaders in analyzing the impacts of development in their cities. Estimating the marginal costs associated with service additons was thus a key part of the overall study. The data displayed in Table 2 under the labels water, sewer, and streets and roads O & M represent the average cost per foot for those services in each city. Averaging was deemed reasonable in that service systems and street and roadway systems include components of varying sizes which would be difficult to disaggregate. In the nonparametric approach, the median is a value of considerable importance. A hypothesized value for the median of each service, assumed to be the simple arithmetic mean (average) of the 10 observations was calculated. The values are $1.00/ft., $.66/ft. and $.81/ft. for water, sewer and streets and roads O & M costs, respectively.

The sign test is the nonparametric test that was applied to the O & M cost data. The number of signs (+ or -) is determined by subtracting the hypothesized median from the observation data. Under the null hypothesis that the population median is equal to a hypothesized median, one would expect about as many positives as negatives. If a sufficiently small number of either sign is present, the null hypothesis is rejected. The test statistic is the smaller count of signs. For example, consider the case of water O & M costs. Statistically, the hypothesis is written as

$$H_o: \text{ median }_{(\text{water})} = \$1.00/\text{ft.}$$

$$H_a: \text{ median }_{(\text{water})} \neq \$1.00/\text{ft.}$$

which is a two-tailed test. The significance level was chosen to be $\alpha = .10$. The test statistic is 4, the number of (+) differences. The decision rule is to reject H_0 if the probability of observing $K = 4$ or fewer (+) signs is less than or equal to $\alpha/2$ for sample size $n = 10$. The table value for this probability is $P(K \leq 4 \mid 10, 0.50) = .377$ which is much greater than .05, therefore the null hypothesis is not rejected. This implies that the hypothesized value is an acceptable estimate for use in the net impact calculations which were the ultimate objective of the study. The other two hypothesized medians were tested in the same way. The test for sewer O & M costs was exactly the same as for water since, $K = 4$. The test of streets and roads was based on $K = 3$, so that $P(K \leq 3 \mid 10, 0.50) = .172$ which is still greater than .05. All three estimates were considered acceptable, given the significance level selected, and used in the net impact analysis.

While the estimates were found to be acceptable they do not contribute any information to help define the limits of acceptable values. Confidence intervals were calculated to establish bounds on the estimates. The calculation of the confidence interval begins with an ordering of the sample from smallest to largest. The estimate of the median is assumed to be the sample median which is the middle value of the ordered array. Since n is even, the sample median is the average of the two middle values (observations 5 and 6 in the ordered array). The sample medians for the data are \$.86, \$.54 and \$.70, respectively for water, sewer, and streets and roads O & M costs. The upper and lower bounds are calculated by determining the largest number of signs that satisfies the probability statement $P(K \leq K' \mid n, 0.50) \leq \alpha/2$. K' is the number of interest and the signs may be either (+) or (-). For $n = 10$ and $\alpha/2 = .05$, $K' \leq 2$. This is the value of K' that results in a probability closest to .05 (.055). The lower bound is defined as the (K' + 1)th observation in the ordered array or the third observation. The upper bound is the (K' + 1)th observation from the top of the array or the eighth observation. The ordered arrays are shown in Table 3. The approximate 90% confidence intervals are as follows:

water O & M

$$(\$.70 \leq \$.86 \leq \$1.38)$$

sewer O & M

$$(\$.36 \leq \$.54 \leq \$.80)$$

streets/roads O & M

$$(\$.47 \leq \$.70 \leq \$1.27)$$

The confidence intervals serve as a guideline for determining the acceptability of estimated operations and maintenance costs for the selected services. For example, suppose that a particular development reqires a substantial addition to the water distribution system. Since the municipality will assume O & M responsibilities at some point, local leaders need to estimate the costs associated with the addition as part of the planning process. The results of this study indicate that estimates of the costs, given the confidence criterion of 90%, should be within the range of $.70 to $1.38 per foot. Other estimates should be viewed with a measure of caution.

Confidence intervals provide a range of statistically acceptable estimates. The one that a municipality finally decides to accept is subject to many factors, including political and fiscal attitudes. From a practical viewpoint, the goal is to get the most accurate estimate possible. Too low an estimate has potentially more serious consequences than an overestimate. If revenues are budgeted for an estimated cost that is too low, they may not cover the actual costs. This will result in a deficit. On the other hand; if revenues cover an estimated cost that is too high, a surplus results. Obviously, surpluses are easier to cope with than are deficits.

SUMMARY

The case study is often criticized as a research approach because it is very "localized," i.e. the case is very site specific. Because of this trait, generalized results are thought not obtainable. The methodology also usually results in a small sample, given the time and expense involved. Despite criticisms, the case study is a much-used approach, particularly in impact studies. Information obtained in such studies provides valuable insights into the development process.

The particular analysis described in this paper is offered as an example of procedures, not often used in economics, which optimize the value of information obtained under the less than "ideal" conditions one might encounter. While this particular situation involves case studies, any situation should be considered a candidate for nonparametric analysis if parametric theory or assumptions are untenable.

Table 2. Data From Idaho Sample Communities Utilized in the Nonparametric Analyses.

Community	Preliminary 1980 Population	O&M Costs for Water Systems $/ft.	O&M Costs for Sewer Systems $/ft.	O&M Costs for Streets/Roads $/ft.
Burley	8,680	.70	.73	.47
Gooding	2,953	.84	.58	1.31
Grace	1,217	.80	.80	.61
Mountain Home	7,522	1.38	.47	.77
Orofino	3,699	.87	.34	1.53
Rupert	5,460	1.93	.96	.67
Sandpoint	4,459	1.75	1.59	.40
Soda Springs	4,041	.38	.49	.73
Wendell	1,971	.30	.36	.30
Weiser	4,795	1.02	.28	1.27
Sample Average	4,480	1.00	.66	.81

Table 3. Ranking of the Sample O & M Cost Data for Obtaining the Approximate 90% Confidence Intervals Centered About the Sample Medians.

Ranking in Ascending Order	Water O & M Costs $/ft	Sewer O & M Costs $/ft	Streets/Roads O & M Costs $/ft
1	.30	.28	.30
2	.38	.34	.40
(Lower bound) 3	.70	.36	.47
4	.80	.47	.61
5	.84	.49	.67
6	.87	.58	.73
7	1.02	.73	.77
(Upper bound) 8	1.38	.80	1.27
9	1.75	.96	1.31
10	1.93	1.59	1.53
Sample median	.86	.54	.70

References

Daniel, Wayne W. Applied Nonparametric Statistics. Houghton Mifflin Company, Boston, Massachusetts. 1978.

Hull, C. Hadlai and Norman H. Nie. SPSS Update: New Procedures and Facilities for Releases 7 and 8. McGraw-Hill Book Company, New York, New York. 1979.

Lentner, Marvin. Elementary Applied Statistics. Bogden & Quigley, Inc., Publishers, Tarrytown-on-Hudson, New York. 1972.

Mosteller, Frederick and Robert E.K. Rourke. Sturdy Statistics - Nonparametrics and Order Statistics. Addison-Wesley Publishing Company, Reading, Massachusetts. 1973.

Snedecor, George W. and William G. Cochran. Statistical Methods, Sixth Edition. Iowa State University Press, Ames, Iowa. 1967.

Steel, Robert G.D. and James H. Torrie. Principles and Procedures of Statistics, A Biometrical Approach, Second Edition. McGraw-Hill Book Company, New York, New York. 1980.