NATIONAL BUREAU OF STANDARDS REPORT

NBS Project	NBS Report
1103-11-1107	1699

A NOTE ON NON-PARAMETRIC METHODS

By

I. Richard Savage

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

FORWORD

This report is a synopsis of a talk presented by Mr. Savage entitled "Some univariate aspects of multivariate non-parametric statistics" presented at the "Statistical Engineering Seminar", National Bureau of Standards, 16 November 1951.

Work on this research project is carried out at the National Bureau of Standards under the general supervision of Dr. Churchill Eisenhart, chief of the Statistical Engineering Laboratory. The Statistical Engineering Laboratory is Section 11.3 of the National Applied Mathematics Laboratories (Division 11, National Bureau of Standards) and is concerned with the development and application of modern statistical methods in the physical sciences and engineering.

J. H. Curtiss
Chief, National Applied
Mathematics Laboratories

A. V. Astin
Acting Director
National Bureau of Standards

A NOTE ON NON-PARAMETRIC METHODS

By

I. Richard Savage
National Bureau of Standards

SUMMARY

The main tools of non-parametric statistics are described. It is pointed out that some of these tools do not have natural extensions to multivariate problems. The possibility of finding tests of randomness that are independent of ordinary tests of hypotheses is explained.

I INTRODUCTION

The object of non-parametric methods is to make probability statements about random variables when "little" is known of the distribution of these random variables. It is hard to give the precise meaning of "little" in the above statement. It refers here to the fact that all we know about the random variables of interest can be phrased in terms of such general function theory concepts as continuity of the distribution functions, the existence of a density function, the possibility of factoring the distribution functions that are involved, etc.

It is interesting to note that most of the work that has been done in a non-parametric manner has been concerned with real-valued (scalar) random variables. Exceptions to

this are, the work done on multidimensional tolerance limits [2] and rank correlation [5]. We shall point out that this is a natural limitation of the common tools that have been used for most of this work, i.e. the use of the probability integral transform, of topological invariance, of randomization, and of dichotomization.

II PROBABILITY INTEGRAL TRANSFORM

The probability integral transform technique says that if $F(x)$ is the cumulative distribution function (continuous) of the random variable X, then the new random variable $Y = F(X)$ has a uniform distribution. This fact has been most useful in achieving by non-parametric methods such results as finding confidence intervals for the median [15], finding confidence intervals for a distribution by the Wald-Wolfowitz method [11], developing the Neyman smooth test [9], etc.

Simpson [10] pointed out that the corresponding transformation for two-dimensional random variables, and consequently for higher dimensions does not have the desired properties. However, in facing a multivariate non-parametric problem it is sometimes possible to consider a real-valued random variable that is a function of the several original random variables, and in this manner the problem can be reduced to one that can be handled. This is essentially what Simpson did in [10].

M. Rosenblatt in his paper ("Limit theorems associated with variants of the von Mises statistic" presented at the

March 21, 1952 contributed paper session of the 51st meeting of the Institute of Mathematical Statistics), has given an analogue to the probability integral transformation which should prove useful in future multivariate, non-parametric studies.

III TOPOLOGICAL INVARIANCE

If points are placed on a line, and then a topological transformation of the line into itself is made, the order of the points on the line either remains the same or is reversed. Many non-parametric procedures depend on this fact, i.e. the Wald-Wolfowitz test of goodness-of-fit using the total number of runs [12], the Mann test of randomness [7], the Mann-Whitney two-sample test [8], etc.

Wolfowitz [16] pointed out that for higher dimensions there was no analogue to this procedure. This is most unfortunate since it has turned out that this is one of the most useful of the non-parametric methods.

It is interesting to note that rank order statistics and symmetric statistics are independent, given a sample of real-valued continuous random variables that are identically and independently distributed [4]. By a rank order statistic is meant a statistic which depends only on the order in which the ranked variables were observed, and thus using the topological invariance, its distribution under random sampling

is non-parametric. Examples of such are the longest run above the median [11], and the statistics used by Mann [7] to test for trend. By a symmetric statistic is meant a symmetric function of the sample observations. The median, the mean, etc., of the sample etc, are symmetric statistics. The proof of the independence theorem for rank order statistics and symmetric statistics and its extension to several samples is simple. The independence of rank order statistics and symmetric statistics is useful in situations where we wish to test the hypothesis that we have a random sample from a population with some specific property; for example, the hypothesis that we have a random sample from a population with median equal to zero. Ordinarily in this type of testing problem all of the emphasis is put on testing whether the median is equal to zero; but in many situations it would also be important to test whether the sample was random. This method allows us to test both parts of the hypothesis with tests which are independent of each other when the null hypothesis is true. Thus it is possible to get known significance levels [14].

IV RANDOMIZATION

The method of randomization is based on the fact that under the assumption of random sampling we can find the exact conditional distribution for any statistic once we have made the set of observations [12].

In this method there is no problem in extending the technique to the multivariate case. Here the main problem consists in the excessive computation required. Thus we would have to make a separate table giving the distribution for each sample drawn. It does not seem possible to give a simple technique for making exact probability statements using this method, although the asymptotic theory has been worked out in detail [3], for the univariate case.

V DICHTOMIZATION

Many problems where non-parametric methods are called for can be so transformed that all of the variables involved have binomial distributions. Typical of such problems is the testing of whether the median of a continuous distribution is equal to zero. Here one classifies the observations as positive or negative; then one proceeds by using the fact that the number of positive and the number of negative observations have a binomial distribution.

By changing one's problems so that they only involve binomial variates one quickly has the required distribution theory. This method has been used by Lehmann [6] for attacking multivariate problems. Perhaps the biggest drawback in using this method is that for some problems it appears to be very inefficient [6].

1. Bateman, G. On the power function of the longest run as a test for randomness in a sequence of alternatives. Biometrika, 35 (1948).

2. Fraser, D. A. S. Sequentially determined statistically equivalent blocks. American Mathematical Society, 22 (1951), p. 372.

3. Hoeffding, W. A combinatorial central limit theorem. Annals of Mathematical Statistics, 22 (1951) p. 558, No. 4.

4. Hotelling, H. and Pabst, M. R. Rank correlation and tests of significance involving no assumption of normality. Annals of Mathematical Statistics, 7 (1936) p. 29.

5. Kendall, M. G. Rank Correlation Methods. Charles Griffin Company Limited, London, 1948.

6. Lehmann, E. Consistency and unbiasedness of certain non-parametric test. Annals of Mathematical Statistics, 22 (1951) p. 165.

7. Mann, H. B. On a test for randomness based on signs of difference. Annals of Mathematical Statistics, 16 (1945), p. 193.

8. Mann, H. B. and Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 18 (1947) p. 50.

9. Neyman, J. "Smooth Test" for goodness of fit. Skandinavisk Aktuarietidskrit, (1937) p. 149.

10. Simpson, P. Note on the estimation of a bivariate distribution function. Annals of Mathematical Statistics, 22 (1951) p. 476.

11. Wald, A. and Wolfowitz J. Confidence limits for continuous distribution functions. Annals of Mathematical Statistics, 10 (1939) p. 105.

12. Wald, A. and Wolfowitz, J. On a test whether two samples are from the same population. Annals of Mathematical Statistics, 11 (1940) p. 147.

13. Wald, A. and Wolfowitz, J. Statistical tests based on permutations of the observations. Annals of Mathematical Statistics, 15 (1944) No. 4.

14. Wallis, W. A. Compounding probabilities from independent significance tests. Econometrica, 10 (1942) No. 3.

15. Wilks, S. S. Order statistics. Bulletin of the American Mathematical Society, 54 (1948) No. 1, Part I, p. 6.

16. Wolfowitz, J. Additive partition functions and a class of statistical hypotheses. Annals of Mathematical Statistics, 13 (1942) p. 247.

June 3, 1952