NONPARAMETRICS:
STATISTICAL METHODS
BASED ON RANKS

NONPARAMETRICS

Statistical Methods Based on Ranks

E. L. LEHMANN
University of California, Berkeley

With the special assistance of
H. J. M. D'ABRERA
University of California, Berkeley

HOLDEN-DAY, INC.
Oakland, California

NONPARAMETRICS

4432 Telegraph Avenue, Oakland, California 94609

Library of Congress Catalog Card Number: 73-94384
ISBN: 0-8162-4994-6 (Holden-Day)

Printed in the United States of America

567890 MUMU 98765

PREFACE

1. History

Methods based on ranks form a substantial body of statistical techniques that provide alternatives to the classical parametric methods. Individual rank tests were proposed much earlier [the earliest use may be that of the sign test by Arbuthnot in 1710; for some additional history see, for example, Kruskal (1957 and 1958)*]; the modern development of the subject may be said to have begun with the papers by Hotelling and Pabst (1936), Friedman (1937), Kendall (1938), Smirnov (1939), and those of Wald and Wolfowitz in the early 1940s. An interesting survey of this work was given by Scheffé (1943).

A full-scale development of rank-based methods seems to have been sparked by the publication in 1945 of a paper by Wilcoxon in which he discussed the two tests, now bearing his name, for comparing two treatments, and by the book of Kendall (1948). Since then there has been a flood of publications that has not yet abated. A bibliography of nonparametric statistics (of which rank-based methods constitute the methodologically most important part) by Savage (1962) lists about 3,000 items. If brought up to date, it probably would contain twice that many entries.

* *See References following Preface.*

The feature of nonparametric methods mainly responsible for their great popularity (and to which they owe their name) is the weak set of assumptions required for their validity. Although it was believed at first that a heavy price in loss of efficiency would have to be paid for this robustness, it turned out, rather surprisingly, that the efficiency of the Wilcoxon tests and other nonparametric procedures holds up quite well under the classical assumption of normality and that these procedures may have considerable advantages in efficiency (as well as validity) when the assumption of normality is not satisfied. These facts were first brought out clearly by Pitman (1948) and were strengthened by results of Hodges and Lehmann (1956) and Chernoff and Savage (1958).

In the early stages, rank-based methods were essentially restricted to testing procedures. They thus did not provide a flexible array of methods, which would include not only tests but point and interval estimates as well as various simultaneous inference procedures. This difficulty is gradually being overcome, although rank-based methods do not yet have the flexibility and the wide applicability to complex linear models that make least squares and normal theory so attractive.

2. The present book

The purpose of this book is to provide an introduction to nonparametric methods for the analysis and planning of comparative studies. Only a relatively small number of basic techniques are presented in detail: These are mainly tests of the Wilcoxon type (which can be obtained from the corresponding classical tests by replacing the observations by their ranks) and the estimation and simultaneous inference procedures based upon these tests. These methods are simple, have good efficiency properties, and most are well tabled. They are treated rather fully here, with emphasis on the assumptions under which they are appropriate, the accuracy of the various approximations that are required, and the modifications needed for tied observations. For the simplest cases there is also a discussion of power or accuracy and the determination of the sample sizes required to achieve a given accuracy. The use of the methods is exemplified in the text, and numerous problems furnish opportunities for the student to try them for himself. In many cases the data for these illustrations are the results of actual studies reported in the literature.

An indication of some alternatives to and extensions of the above procedures, and of some additional properties, is provided by sections of further developments at the end of each chapter, which give an introduction to the literature on these subjects. Among the topics treated this way are the Normal Scores procedures, permutation tests, sequential methods, and optimum theory. Two topics that are not covered are multivariate techniques (because of lack of space) and goodness-of-fit tests (because both the problem and the data are quite different from those considered here). On the other hand, a discussion of some tests for two-way

contingency tables is included in the text because they can be viewed as special cases of rank tests with tied observations.

As mentioned above, an important advantage of nonparametric tests is the simplicity of the assumptions required for their validity. It is not necessary to postulate a population from which the subjects in a study have been obtained by random sampling, but only that the treatments being compared have been assigned to the subjects at random. All techniques are first discussed in terms of such a *randomization model.* This material (which constitutes Chapters 1 and 3 and some parts of Chapters 5 to 7) requires only the simplest mathematical background for its understanding. All that is needed is an elementary introduction to probability, such as that provided by mathematics courses in many high schools or the first lectures in an introductory course on probability or statistics.

It is possible within this framework to describe the tests, illustrate their use, and discuss the computation of significance or critical values. However, randomization models do not permit an evaluation of power that could be used to plan the size of a study. This is best discussed in terms of a population from which the subjects have been sampled. Unfortunately, an adequate treatment of *population models* (such as those underlying the normal distribution) requires some knowledge of the calculus.

The determination of sample size and the evaluation of the power of a test (which plays an analogous role as the variance of an estimate) seemed too important to omit. So as to include these somewhat more advanced topics, I have allowed the level of the book to vary with the requirements of the material. Despite the obvious disadvantages of such inconsistency, it is my hope that even the reader with little mathematical background will be able to follow the main ideas of the more advanced parts (Chapters 2 and 4 and portions of Chapters 5 to 7) and that the reader whose background is stronger will not be put off by the slow pace of the more elementary sections. I am encouraged in this hope by the fact that courses I have taught along these lines to students with very disparate backgrounds seem to have been reasonably successful.

The main text is followed by an appendix that provides the large-sample theory underlying the many approximations required where tables are not available and exact computations are too laborious. This material is at an intermediate level. It requires substantially more mathematical sophistication than the rest of the book, but it is much less advanced than the books by Hájek and Sidak (1967) and Puri and Sen (1971). A number of standard limit theorems from probability theory are stated and discussed but not proved, and on this basis the needed results are derived with relatively little effort.

3. Acknowledgments

There remains the pleasant task of acknowledging the many debts that I incurred during the writing of the book. I should like to thank Peter Bickel, Kjell Doksum, Gus Haggstrom, Joe Hodges, Bill Kruskal, Vida Lehmann, and Juliet Popper Shaffer who at various stages read portions of the manuscript and gave me their moral support and the benefit of their advice. Of the reviewers who examined the manuscript for the publishers, I am especially grateful to Ralph D'Agostino and Alfred Forsyth for their many valuable criticisms and ideas for improvements, and to Gottfried Noether for his crucial support and suggestions.

To two friends I owe a special debt. Ellen Sherman read critically, checked the computations, and prepared the tables for an early version of the first chapters. Above all, Howard D'Abrera performed the same task for later versions of the whole manuscript and prepared the answers for the problems. He corrected innumerable errors of style, thought, and arithmetic. Without his help I could not have brought this project to its completion.

The aspect of the book that caused me the most difficulty was to find suitable live examples and problems. Authors typically do not publish their data, and when a set of published data is potentially suitable, it usually turns out that the sample size is too large or small, there are too many or too few ties, the results are too obviously significant or too obviously not, or that the design or sampling procedure is not what is required to illustrate the particular point in question. I am grateful to the colleagues who put their unpublished data at my disposal as well as those who published data that I was able to use. To all I would like to extend an apology. When some minor modification of the data made them more suitable for my purpose, I have carried out such modifications (always with an acknowledgment). More seriously, I have used the data to illustrate the point I was trying to make, though this may have borne little relation to the purpose for which they were collected, and I have asked questions of the data that may seem foolish to someone more familiar with the actual situation. I hope that I will be forgiven for these violations of the authors' intentions, and I should like to ask readers who have available more suitable data to illustrate the techniques discussed here to let me know about them for a possible later revision.

Finally, it is a pleasure to thank the Office of Naval Research for their generous support of the research that has gone into this book.

E. L. Lehmann

Berkeley
1974

PREFACE REFERENCES

Arbuthnot, J. (1710): "An Argument for Divine Providence, Taken from the Constant Regularity Observed in the Births of Both Sexes," *Philos. Trans.* **27**:186–190.

Chernoff, Herman and Savage, I. Richard (1958): "Asymptotic Normality and Efficiency of Certain Nonparametric Test Statistics," *Ann. Math. Statist.* **29**:972–994.

Friedman, Milton (1937): "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *J. Amer. Statist. Assoc.* **32**:675–701.

Hájek, J. and Sidak, Z. (1967): *Theory of Rank Tests*, Academic Press.

Hodges, J. L., Jr. and Lehmann, E. L. (1956): "The Efficiency of Some Nonparametric Competitors of the *t*-Test," *Ann. Math. Statist.* **27**:324–335.

Hotelling, Harold and Pabst, Margaret Richards (1936): "Rank Correlation and Tests of Significance Involving No Assumptions of Normality," *Ann. Math. Statist.* **7**:29–43.

Kendall, Maurice G. (1938): "A New Measure of Rank Correlation," *Biometrika* **30**:81–93.

——— (1948): *Rank Correlation Methods*, 4th edition (1970), Griffin, London.

Kruskal, William H. (1957): "Historical Notes on the Wilcoxon Unpaired Two-sample Test," *J. Amer. Statist. Assoc.* **52**:356–360.

——— (1958): "Ordinal Measures of Association," *J. Amer. Statist. Assoc.* **53**:814–861.

Pitman, E. J. G. (1948): *Lecture Notes on Nonparametric Statistics*, Columbia Univ., New York.

Puri, M. L. and Sen, P. K. (1971): *Nonparametric Methods in Multivariate Analysis*, John Wiley.

Savage, I. Richard (1962): *Bibliography of Nonparametric Statistics*, Harvard Univ. Press.

Scheffé, Henry (1943): "Statistical Inference in the Nonparametric Case," *Ann. Math. Statist.* **14**:305–332.

Smirnov, N. V. (1939): "On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent Samples," *Bull. Math. Univ. Moscow 2*, No. 2, 3–14.

Wilcoxon, Frank (1945): "Individual Comparisons by Ranking Methods," *Biometrics* **1**:80–83.

CONTENTS

CHAPTER 1

RANK TESTS FOR COMPARING TWO TREATMENTS

1. RANKS IN THE COMPARISON OF TWO TREATMENTS

The problem of deciding whether a proposed innovation constitutes an improvement over some standard procedure arises in many different contexts. Does a new "cure" prolong the life of cancer patients? Is the harmful effect of cigarettes reduced by filtering? Does a new expensive gasoline additive increase the mileage? Does cloud-seeding lead to increased precipitation? Or conversely, is televised instruction less effective than live classroom teaching? The following example illustrates the kind of evidence that may be used to obtain at least tentative answers to such questions.

EXAMPLE 1. *A new drug.* A mental hospital wishes to test the effectiveness of a new drug that is claimed to have a beneficial effect on some mental or emotional disorder. There are five patients in the hospital suffering from this disorder to about the same degree. (Actually, this number typically would be too small to provide meaningful results.) Of these five, three are selected at random to receive the new drug, and the other two serve as controls: they are given a placebo, a harmless pill not containing any active ingredients. In this way the patients (and preferably also the staff) do not know which patients are receiving the new treatment. This eliminates the possibility of psychological effects that might result from such knowledge.

After some time, a visiting physician interviews the patients and ranks them according to the severity of their condition. The patient whose condition is judged to be most serious is assigned rank 1, the next most serious rank 2, and so on, up to rank 5. The claim made for the new treatment will be considered warranted if the three treated patients rank sufficiently high in this combined ranking of all five patients. A basis for evaluating the significance of the observed ranking is provided by the following consideration.

Suppose that the treatment has no effect, i.e., that a patient's health is in no way affected by whether or not he receives the new drug. We shall refer to this assumption as the *hypothesis H of no treatment effect*. Since under the assumption of this hypothesis (for short, *under H*) the rank of each patient is determined solely by his state of health, it is clear that the ranking of the patients does not depend on which of them receive the treatment and which serve as controls. We may thus think of each patient's rank as attached to him even before the assignments to treatment and control are made. The selection of three patients to receive the treatment then also selects three ranks: those attached to the selected patients. Each possible such selection divides the ranks into two groups: the ranks of the treated patients and of the controls. These divisions are displayed in (1.1) for all possible cases. Thus, for example, the first box in (1.1) corresponds to the possibility that the three patients who eventually are awarded the highest ranks (3, 4, 5) are those receiving the treatment.

(1.1)

Treated	(3,4,5)	(2,4,5)	(1,4,5)	(2,3,5)	(1,3,5)
Controls	(1,2)	(1,3)	(2,3)	(1,4)	(2,4)
Treated	(2,3,4)	(1,3,4)	(1,2,4)	(1,2,3)	(1,2,5)
Controls	(1,5)	(2,5)	(3,5)	(4,5)	(3,4)

As is seen from (1.1), the patients and hence their ranks can be divided into two groups in 10 different ways. The assumption that the three patients receiving the treatment are selected at random means that these 10 possible divisions are equally likely, i.e., that each of the 10 possibilities displayed in (1.1) has probability $\frac{1}{10}$. As will be discussed in the next section, this fact provides a basis for assessing the significance of the observed ranking.

At the beginning of the example it was assumed that the five patients are suffering from their disorder to about the same degree. Actually, no use was made of this assumption in the above derivation. The random assignment of the patients to treatment and control implies that the ranks are also assigned to these two groups at random (under the hypothesis of no treatment effect), regardless of the initial states of health of the patients. However, as we shall discuss later (Chap. 2, Sec. 2, and Chap. 3, Sec. 1), increased homogeneity of the patients with respect to initial state of health and to other relevant factors increases the power of the experiment, that is, the likelihood of detecting that the treatment has an effect when this is the case. This is intuitively quite plausible since with nearly identical responses for patients under the same regime any difference must be due to treatment and will stand out clearly, but it will be masked for patients with widely different inherent responses.

The considerations introduced in the context of the above example easily generalize. Suppose that N experimental subjects are available for a comparative

study and that n of these are selected at random to receive a new treatment with the remaining $m = N-n$ serving as controls.[1] Let us denote the number of possible choices of n out of N subjects by $\binom{N}{n}$. For $N = 5$ and $n = 3$ we have seen that $\binom{5}{3} = 10$. The number $\binom{N}{n}$ is variously known as *the number of combinations of N things taken n at a time, the number of samples of size n from a population of size N*, or as a *binomial coefficient*, and can be computed from the formula[2]

$$\binom{N}{n} = \frac{N(N-1)\cdot \cdots \cdot (N-n+1)}{1\cdot 2\cdot \cdots \cdot n} \tag{1.2}$$

A table of these coefficients for $N \leqslant 25$ and $n \leqslant 12$ is given as Table A at the end of the book.[3] The values for $N \leqslant 25$ and $n > 12$ can be obtained from the tabulated values by means of the formula (Prob. 70)

$$\binom{N}{n} = \binom{N}{N-n} \tag{1.3}$$

To find, for example, the value of $\binom{20}{16}$ one notes that by (1.3) it is equal to $\binom{20}{4}$; entering Table A in the row $N = 20$ and the column $n = 4$, it is seen that

$$\binom{20}{16} = \binom{20}{4} = 4{,}845$$

By assumption, the n subjects receiving the treatment are selected from the N available subjects *at random*; that is, all $\binom{N}{n}$ possible choices of these subjects are equally likely so that each has probability $1\Big/\binom{N}{n}$. At the termination of the study, the subjects are ranked (preferably by an impartial observer) with respect to the condition at which the treatment is aimed. As before, under the hypothesis H of no treatment effect, the ranking is not affected by which subjects received the treatment. The rank of each subject may be considered as determined (although unknown) before the assignment of subjects to treatment and control is performed, and hence to be assigned to treatment or control together with the subject. Thus,

[1] A convenient aid for making such a random selection is a table of random numbers such as the RAND Corporation's *A Million Random Digits with 100.000 Normal Deviates*. The Free Press, New York, 1955, or the book by Moses and Oakford (1963).

[2] For a proof of Eq. (1.2), see, for example, Goldberg (1960), Hodges and Lehmann (1970), or Mosteller, Rourke, and Thomas (1970). (Consult the References, Sec. 9, at the end of this chapter for detailed bibliographic data.)

[3] A more extensive table giving all values for $N \leqslant 200$ is the *Table of Binomial Coefficients*. Cambridge University Press, London, 1954.

under H, the $\binom{N}{n}$ possible assignments of n of the integers $1, \ldots, N$ as treatment ranks each have probability $1 \Big/ \binom{N}{n}$.

The above result is so fundamental that we shall now restate it more formally. Let the ranks of the treated subjects be denoted by $S_1, \ldots, S_n$, where we shall assume that they are numbered in increasing order so that $S_1 < S_2 < \cdots < S_n$, and let the ranks of the controls be $R_1 < R_2 < \cdots < R_m$. Between them, these $m+n$ ranks are just the integers 1, 2, ..., N. Since the R's are determined once the S's are known, the division of the ranks into the two groups can be specified by the n-tuple $(S_1, \ldots, S_n)$. The $\binom{N}{n}$ possible such n-tuples constitute the possible outcomes of the study.

The basic result derived above states that the probability, under H, of observing any particular n-tuple $(s_1, \ldots, s_n)$ is

$$P_H(S_1 = s_1, \ldots, S_n = s_n) = \frac{1}{\binom{N}{n}} \tag{1.4}$$

for each of the possible n-tuples $(s_1, \ldots, s_n)$.

Let us now illustrate the use of ranks in comparative studies with another example.

EXAMPLE 2. *Effect of discouragement.* To test whether discouragement adversely affects performance in an intelligence test, 10 subjects were divided at random into a control and treatment group of 5 each. Both were given Form L of the revised Stanford-Binet test under the conditions prescribed for this test. Two weeks later they were given Form F, the controls under the prescribed conditions, the treated subjects under conditions of discouragement (you are doing terribly, etc.). The following were the differences in their scores: later value − original value.[1]

Controls: 5 0 16 2 9 Treated: 6 −5 −6 1 4

If the subjects are ranked, with rank 1 going to the subject with the smallest difference, rank 2 to the next smallest, etc., the ranks of the treatment subjects are 1, 2, 4, 6, and 8, and those of the controls 3, 5, 7, 9, and 10. Both the differences and their ranks suggest that, on the whole, the five subjects receiving the discouragement did less well than the other five. An assessment of the significance of this result will be taken up in the next section and will be based on the following consideration. Let H denote the hypothesis that the treatment has no effect, i.e., that the discouragement will have no influence on the score obtained by a subject. If H is

[1] Part of data of Gordon and Durea, "The Effect of Discouragement on the Revised Stanford-Binet Scale," *J. Genetic Psychol.* **73**:201–207 (1948). The original experiment involved 20 subjects in each group. Of these, five were selected at random for the present example. For the original data, see Prob. 48.

true, the difference in the scores of each of the 10 subjects, and hence his rank, is unaffected by the method to which he is assigned. The $\binom{10}{5} = 252$ possible sets of values of the five treatment ranks are therefore equally likely, each having probability $1/252$.

The structure of the two examples is basically the same, but they differ in one important respect. Although in the first example only a ranking of the subjects (patients) was available, the data in Example 2 consisted of a measurement for each subject (the difference in the scores) and the subjects were ranked according to the values of these measurements. The reader may feel that in the second case a test of the hypothesis of no treatment effect should be based on the original measurements rather than on the ranks derived from them. It turns out, however, that tests based on ranks have certain advantages, which will be discussed in Chap. 2.

Throughout this section, we have assumed that the subjects which are available for observation are not chosen but are given and that they are assigned at random, n to treatment and $N-n$ to control. We shall call this model, in which chance enters only through the assignment of the subjects to treatment and control, the *randomization model.* This is to distinguish it from another model, to be considered in Chap. 2, according to which the N subjects are not fixed but are drawn in some specified manner from a population of such subjects. In this case, chance is involved also (in a way that can be taken into account) in the selection of the subjects.

2. THE WILCOXON RANK-SUM TEST

For comparing a new treatment or procedure with the standard method, N subjects (patients, students, etc.) are divided at random into a group of n who will receive a new treatment and a control group of m who will be treated by the standard method. At the termination of the study, the subjects are ranked either directly or according to some response that measures the success of the treatment such as a test score in an educational or psychological investigation, the amount of rainfall in a weather experiment, or the time needed for recuperation in a medical problem. The hypothesis H of no treatment effect is rejected, and the superiority of the new treatment acknowledged, if in this ranking the n treated subjects rank sufficiently high. (Here it is assumed that the success of the treatment is indicated by an increased response; if instead the aim is to decrease the response, H is rejected when the n treated subjects rank sufficiently low.)

To complete the specification of the procedure, it is necessary to decide just when the treatment ranks $(S_1, \ldots, S_n)$ are sufficiently large. Such an assessment is typically made in terms of some test statistic, large values of which correspond to the treatment ranks being large. A simple and effective such statistic is the sum of