Applied Nonparametric Statistical Methods

Peter Sprent

CHAPMAN AND HALL
World Publishing Corp

Applied Nonparametric Statistical Methods

Peter Sprent

Emeritus Professor of Statistics
University of Dundee, Scotland, UK

CHAPMAN AND HALL
World Publishing Corp

UK	Chapman and Hall, 11 New Fetter Lane, London EC4P 4EE
USA	Chapman and Hall, 29 West 35th Street, New York NY 10001
JAPAN	Chapman and Hall Japan, Thomson Publishing Japan, Hirakawacho Nemoto Building, 7F, 1-7-11 Hirakawa-cho, Chiyoda-ku, Tokyo 102
AUSTRALIA	Chapman and Hall Australia, Thomas Nelson Australia, 480 La Trobe Street, PO Box 4725, Melbourne 3000
INDIA	Chapman and Hall India, R. Sheshadri, 32 Second Main Road, CIT East, Madras 600 035

First edition 1989
Reprinted 1990

Typeset in 10/12 Times by
Thomson Press (India) Ltd, New Delhi

ISBN 0 412 30600 X (hardback)
0 412 30610 7 (paperback)

British Library Cataloguing in Publication Data

Sprent, Peter
Applied nonparametric statistical methods.
1. Nonparametric statistical mathematics. Applications
I. Title
519.5

ISBN 0-412-30600-X. ISBN 0-412-30610-7 (pbk.)

Library of Congress Cataloging in Publication Data

Sprent, Peter
Applied nonparametric statistical methods/Peter Sprent.
p. cm.
Bibliography: p.
Includes index.
ISBN 0-412-30600-X. ISBN 0-412-30610-7 (pbk.)
1. Nonparametric statistics. I. Title.
QA278.8.S74 1989
519.5--dc19

This edition first published in the United Kingdom by Chapman & Hall, 1989
Reprinted by World Publishing Corporation, Beijing, 1991
Not for sale or distribution outside The People's Republic of China
ISBN 7-5062-1205-6

Applied Nonparametric Statistical Methods

OTHER STATISTICS TEXTS FROM
CHAPMAN AND HALL

The Analysis of Time Series
C. Chatfield

Statistics for Technology
C. Chatfield

Introduction to Multivariate Analysis
C. Chatfield and A. J. Collins

Applied Statistics
D. R. Cox and E. J. Snell

An Introduction to Statistical Modelling
A. J. Dobson

Introduction to Optimization Methods and their Application in Statistics
B. S. Everitt

Multivariate Statistics – A Practical Approach
B. Flury and H. Riedwyl

Multivariate Analysis of Variance and Repeated Measures
D. J. Hand and C. C. Taylor

Multivariate Statistical Methods – a primer
Bryan F. Manley

Statistical Methods in Agriculture and Experimental Biology
R. Mead and R. N. Curnow

Elements of Simulation
B. J. T. Morgan

Probability Methods and Measurement
Anthony O'Hagan

ial Statistics
Rees

Decision Analysis: A Bayesian Approach
J. Q. Smith

Applied Statistics: A Handbook of BMDP Analyses
E. J. Snell

Elementary Applications of Probability Theory
H. C. Tuckwell

Intermediate Statistical Methods
G. B. Wetherill

Further information on the complete range of Chapman and Hall *statistics books is available from the publishers.*

Preface

This book is a practical introduction to statistical techniques called **nonparametric methods**. Using examples, we explain assumptions and demonstrate procedures; theory is kept to a minimum. We show how basic problems are tackled and try to clear up common misapprehensions so as to help both students of statistics meeting the methods for the first time and workers in other fields faced with data needing simple but informative analysis.

An analogy between experimenters and car drivers describes our aim. Statistical analyses may be done by following a set of rules without understanding their logical basis, but this has dangers. It is like driving a car with no inkling of how the internal combustion engine, the gears, the ignition system, the brakes actually work. Understanding the rudiments helps one get better performance and makes driving safer; appropriate gear changes become a way to reduce engine stress, prolong engine life, improve fuel economy, minimize wear on brake linings. Knowing how to change the engine oil or replace worn sparking plugs is not essential for a driver, but it will reduce costs. Learning such basics will not make one a fully fledged mechanic, even less an automotive engineer; but it all contributes to more economical and safer driving, alerting one to the dangers of bald tyres, a leaking exhaust, worn brake linings.

Many research workers, industrialists and businessmen carry out their own basic statistical analyses. Professional statisticians may deplore this (as skilled mechanics grumble about do-it-yourself car servicing). These professional attitudes are a mixture of self-interest and genuine concern that serious mistakes may be made by the amateur.

This book is not meant to turn those meeting data in their daily work into professional statisticians, any more than a guide to do-it-yourself car servicing will turn one into a trained mechanic.

Relatively straightforward nonparametric counterparts of old established statistical tools are dealt with in Chapters 1 to 8 with only occasional references to sophisticated material (e.g. log-linear models in Section 8.3).

In Chapter 9 we look at some recent developments, while Chapter 10 outlines more advanced techniques; use of these will generally require guidance from a professional statistician, but it is handy for data-producers to know what is on offer. Using our motoring analogy, few do-it-yourself car servicers have the skill or tools to replace the gearbox or install a new engine; but it helps to know a little about available alternatives which may include

removal and repair, replacement by a new unit or by a reconditioned one; all have pros and cons; so too with advanced statistical analyses.

We use real (or at least realistic) data in examples and exercises; some specially obtained for this book, some extracted from larger published sets with sources indicated. Reference to the source will often show that the complete data sets have been analysed with different objectives using more advanced techniques (parametric or nonparametric). These advanced analyses are akin to specialist mechanical maintenance of a car.

The book is a detailed and modernized development from an earlier work of mine, *Quick Statistics* (Penguin Books, 1981). Emphasis there was on the 'quick'; here it is on the 'statistics'. Some common ground is covered but the change in emphasis is a logical development in the light of new attitudes to statistical methods stimulated by availability of ever-increasing computer power.

To keep the book at a reasonable length for an introductory text without making discussion of each topic too terse, I have given references to accounts of some topics that were strong candidates for inclusion where these are well covered at this level by other writers.

P. Sprent
December 1987

A new computing development worth mentioning is a statistical package called STATXACT from Cytel Software Corporation, 137 Erie St., Cambridge, Mass., USA. It uses efficient algorithms to compute exact permutation probabilities for some tests described in this book, comparing these with asymptotic results. It is available for IBM-PC compatibles and is particularly useful in situations where small sample sizes or other factors may result in questionable validity for asymptotic tests.

P. Sprent
June 1990

Contents

1

Introducing nonparametric methods

1.1 BASIC STATISTICS

In most of this book we assume only a rudimentary knowledge of statistics like that provided by a service or introductory course of 10 to 20 lectures, or by reading a simple text like *Statistics Without Tears* (Rowntree, 1981).

Readers knowing no statistics may still follow this book by reading a text like Rowntree's in parallel; those happy with elementary mathematical notations may prefer the more sophisticated but basic approach in Chapters 1 to 8 of *Statistics for Technology* (Chatfield, 1983), or one of the many other introductory texts by both British and American authors.

In the Appendix we summarize some general statistical concepts that are especially relevant to nonparametric methods. An 'A' before a section number implies it is in the Appendix and references to the Appendix are given in the form 'see Section A6', etc. Section A8 gives tables for nonparametric procedures. The headings of these tables have an 'A' before their number, e.g. Table A6 is the sixth table in Section A8.

Basic statistics courses do not always include practical applications of nonparametric methods. In this chapter we survey some fundamentals and give one or two illustrations. Specific techniques are discussed in later chapters.

1.1.1 Parametric methods

Before explaining the nonparametric kind, a word about parametric inference.

Early in statistics courses one meets random variables belonging to the family of **normal distributions**. Members of that family are distinguished by different means and/or variances, often denoted by the Greek letters μ, σ^2 respectively, and called **parameters.**

Another well known family is that of **binomial distributions**, characterized by two parameters, n, the number of observations and p, the probability of one of two possible outcomes at each observation (often called success and failure). The number of successes in a sequence of n independent observations (trials) when there is probability p of success at each has a binomial distribution.

Given one or more sets of independent observations (called random

samples) assumed to come from some distribution belonging to a named family, we often want to infer something about the unknown parameters. The sample mean provides an estimate of the distribution (or population) mean. With a sample from a normal distribution the t-test (Section A4) may be used to decide if the sample is consistent with an *a priori* hypothesized population mean μ_0. The related t-distribution lets us establish a **confidence interval**: an interval (see Section 1.3.1) in which we are reasonably confident the true but unknown mean μ lies.

For a binomial distribution, if there are r successes in n independent observations we call r/n a **point estimate** of p and we may test whether that estimate supports an *a priori* hypothesized value p_0, say, or obtain a confidence interval for the true value of p, the probability of success at each independent observation.

In practice an assumption that observations come from a particular family of distributions such as the normal or the binomial may be quite reasonable. Experience, backed to some extent by theory, suggests that, for many measurements, inferences based on an assumption that observations are random samples from a normal distribution, known apart from one or both parameters, may not be seriously astray even if the normality assumption is incorrect. But this is not always true.

We sometimes want to make inferences that have nothing to do with parameters, or we may have data in a form that makes, say, normal theory tests inappropriate; we may not have precise measurement data, but only the rank order of observations. For example, although it is often reasonable to assume examination marks are approximately normally distributed, these marks may not be published. We may only know the numbers of candidates in banded and ordered grades designated Grade A, Grade B, Grade C, ..., or Level I, Level II, etc. In Example 1.1 we consider a situation where we know only total numbers of items and the proportions with a certain characteristic.

Even when we have precise measurements it may be obvious that we cannot assume a normal distribution. We may be able to say little except perhaps that the distribution is skew, or symmetric, or has some other characteristic.

Appropriate methods of inference in these situations are described as **nonparametric**, or sometimes more aptly, as **distribution-free**.

Many writers regard 'distribution-free' and 'nonparametric' as synonyms, a view that ignores subtle distinctions that need not worry us here. Some tests that are generally classed as nonparametric or distribution-free do indeed involve parameters and distributions and the 'distribution-free' or 'nonparametric' tag simply means they can be applied to samples that come from populations having any one of a wide class of distributions. In general, these methods are applicable to estimation or hypothesis-testing problems when the population distributions need only be specified in broad terms, e.g. as being

continuous, symmetric, identical, differing only in median or mean; they need not belong to specific families such as the normal, uniform, exponential, etc. Logically, the term distribution-free may then be more appropriate than nonparametric, but the latter term is well established in popular usage.

1.1.2 Why do we need nonparametric methods?

A parametric test may depend crucially for its validity on an assumption that we are sampling randomly from a distribution belonging to a particular family. If there is doubt a nonparametric test that is valid under weaker assumptions is preferable. Nonparametric methods are invaluable – indeed they are usually the only methods available when we have data that specify just order or ranks or proportions and not precise observational values.

It must be stressed that weaker assumptions do not mean (as research workers sometimes misguidedly think) that nonparametric methods are assumption-free. What can be deduced depends on what assumptions can validly be made; an example demonstrates this.

Example 1.1

A manufacturer mass-produces an item that has a nominal weight of 1000 g and gives a guarantee that in large batches not more than 2.5% will weigh less than 990 g. The plant is highly automated and to check that the two machines being used are producing goods of acceptable quality the manufacturer takes samples of 500 at regular intervals from the production run for each machine. These are put through a quick-operating electronically controlled checker that rejects all items from the 500 that weigh less than 990 g. This provides the only check that the requirement is being met that not more than 2.5% are below 990 g. This is a typical observation for a quality control programme.

To give reasonable protection the machines may be adjusted to produce not more than 2.25% underweight items (i.e. below 990 g) when operating properly. If underweight items are produced at random and the target of 2.25% is maintained for a large batch, then the number underweight in samples of 500 will have a binomial distribution with $n = 500$ and $p = 0.0225$. Standard quality control methods use such information to indicate if a batch is reasonably likely to meet the guarantee criterion; such test procedures are parametric, based on the binomial distribution.

But the manufacturer may ask if other deductions can be made from the test information. For example, do the numbers underweight throw light on the underlying distribution of weights? For example, can we use the observed numbers underweight in samples of 500 from each of the two machines to test whether the mean weights of items produced by each machine are equal, or have a specific value? We cannot do this without further assumptions about the distributions of weights for each machine. This is immediately apparent

because the proportion weighing less than 990 g will be 2.25% for an infinity of possible distributions. For example, if the weights are normally distributed with a mean of 1000 g and a standard deviation of 5 g then the long-run proportion below 990 g can be shown to be 2.25% (more exactly 2.28%, but for simplicity we ignore this rounding difference). Also, if the weights were distributed normally with a mean of 1005 g and standard deviation 7.5 g, the proportion below 990 g is again 2.25%, as it also would be if the weights had a uniform (rectangular) distribution over the interval (989.55, 1009.55). We could find not only an infinity of other normal or uniform distributions, but also gamma and other distributions which gave the same proportion below 990 g. For all of these the binomial distribution with $n = 500$ and $p = 0.0225$ is relevant to the distribution of underweight items in our samples. Tests based on the binomial distribution cannot therefore tell us on their own if the two machines are producing items with the same mean weight. However, if we now make an additional assumption that the output from the two machines have **identical distributions apart perhaps from a shift in median if something has gone wrong with one**, we may use our binomial-type information to test the hypothesis that the medians are identical, against the alternative that the two population medians differ. This is a nonparametric test, but it would not be very efficient. If, from past experience, the manufacturer could say that the weights always has a normal distribution with known standard deviation, say 6, but the mean was liable to shift if things went wrong, then knowing the number of items with weight below 990 g in a sample of 500 enables one to test a hypothesis that the mean is 1002 (the target value to give 2.25% defectives with standard deviation 6) against the alternative that the mean had some other value. This would be a parametric test about a parameter μ, the mean of a normal distribution with standard deviation 6. Again, it would not be a very efficient test, but the best we could do without more detailed information about weights of the items sampled. We say more about hypothesis testing in Section 1.2.

Means and medians (see Section A1.2) are common measures of location. The most common problems with measurement data concern **location**. Is it reasonable to suppose a sample comes from a population with a certain specified mean or median? Can we reasonably assert that two samples come from populations whose means differ by at least 10 units? Given a sample, what is an appropriate estimate of the population mean or median? How good an estimate is it?

Some nonparametric methods require only minimal information. To test whether we may reasonably assert that a sample might be drawn from a distribution with pre-specified median θ, say, we need only know how many sample values are greater than θ and how many are less than θ. If it were difficult or expensive to get exact observations, but easy to determine

numbers above or below θ, this nonparametric approach may be very cost-effective.

Example 1.2

I have on my bookshelf 114 books on statistics. I take a random sample of 12 of these (i.e. the 12 are selected by a method that gives each of the 114 books an equal chance of selection – see Section A3). I want to test the hypothesis that the median number of pages per volume is 220.

I check in each of the 12 volumes selected whether or not the number of pages exceeds 220. In 9 it does, in the remaining 3 it does not. I record this as 9 pluses (+), or excesses over the hypothetical value, and 3 minuses (−).

A population median of 220 would imply half of the 114 books on my shelf have 220 or less pages, half that or more. This follows from the definition of a median (Section A1.2). Thus, if the median were 220, when we select a random sample it is (for practical purposes) equally likely that each book selected will have more than 220 pages (scored as +) or less (scored as −). A complication we discuss in Section 2.1.1 occurs if a sampled book has exactly 220 pages; this did not happen in my sample.

By associating a plus with a 'head' and a minus with a 'tail' we have a physical analogue with coin tossing; if the median is really 220, then the result 9 plus and 3 minus signs is physically equivalent to 9 heads and 3 tails when a fair coin is tossed 12 times. We show in Section 1.2 that this evidence does not justify rejection of the hypothesis that the population median is 220.

Non-rejection of a hypothesis in this sense does not **prove** it is true; it only means that currently we have insufficient evidence to reject it. We do not reject, because 9 heads and 3 tails is in a set of reasonably likely results when we toss a true coin. Had we got 12 plus and no minus signs, or vice versa (equivalent to 12 heads or 12 tails) we could reasonably reject the hypothesis that the median is 220. Indeed, the probability of getting one or other of these extremes in a sample of 12 is only 1/2048, so such a result in just one experiment means either we have observed an event with odds heavily stacked against it, or our hypothesis of a fair coin (or that the median is 220) is not correct. The latter seems more plausible. For those who are not already familiar with these ideas we formalize them in Section 1.2. The relevant test is called the **sign test**.

1.1.3 An historical note

It is fashionable to claim that nonparametric methods were first used when J. Arbuthnot (1710) found that in each year from 1629 to 1710 the number of males christened in London exceeded the number of females. He regarded this as strong evidence that the probabilities of any birth being male or female were not exactly equal, a discrepancy Arbuthnot attributed to 'divine providence'. A sign test is appropriate for his data.

Spearman (1904) proposed a rank correlation coefficient that bears his name, but a serious study of nonparametric methods for statistical inference began some fifty years ago in an era when applied statistical methods were dominated by grossly oversimplified mathematical models of real-world situations chosen partly because they led to not too demanding computational procedures: the inaptly named normal distribution was the key to analytical methods for continuous data; the binomial and Poisson distributions to methods for discrete data. These distributions still are – and always will be – important, but they are not all-embracing.

Research into nonparametric and distribution-free methods was stimulated firstly by attempts to show that even if assumptions of normality often stretched credulity, then at least in some cases making those assumptions would not greatly alter valid inferences. This was the stimulus of work by R. A. Fisher, E. J. G. Pitman and B. L. Welch on randomization or permutation tests; tests which at that time (the 1930s) were computationally too demanding for general use.

At about the same time there was a growing realization that observational data that were not numerically precise but consisted of preferences or rankings could be used to make inferences in a way that required little computation effort.

A few years later F. Wilcoxon and others realized that even if we had precise numerical data we sometimes lost little useful information by replacing them by their rank order and basing our analysis on computationally simple procedures using these ranks. Indeed, if an assumption of normality were not justified, analyses based on ranks were sometimes the most efficient available. This heralded an era when nonparametric methods developed as practical tools for use either when data were by nature simply ordinal (ranks or preferences) as distinct from precise measurements (interval or scalar); or as a reasonably efficient method that reduced computation even when full numerical data were available, but could easily by replaced by ranks. Used in this way there were still many limitations: simple hypothesis testing was usually easy; interval estimation much more difficult.

Ever-increasing calculating power of modern computers has revolutionized our approach to data analysis and statistical inference (see e.g. Durbin, 1987). Pious hope that data fit a restricted mathematical model with few parameters and emphasis on simplifying concepts such as linearity have been replaced by the use of robust methods and by exploratory data analysis in which we investigate different potential models; areas where nonparametric methods have a central role.

They may also be applied to counts, these often recorded as numbers of items in various categories; e.g. numbers of examination candidates obtaining Grade A, Grade B, Grade C passes. Here the categories are **ordinal**; Grade A is better than Grade B; Grade B is better than Grade C; and so on. Categories

that cannot be ordered by the inequalities greater than or less than are called **nominal**; e.g. people may be classified as single, married, widowed or divorced. For data in these forms nearly all analyses are by nature nonparametric.

A disadvantage of nonparametric methods in the pre-computer era was that simplicity only applied to basic procedures and nonparametric methods lacked the flexibility of much linear model and least squares theory that are cornerstones of normal distribution parametric inference. The advent of computers has revolutionized this aspect of using nonparametric methods, for many advanced and flexible methods are tedious only in that they require repeated application of simple calculations – a task for which computers are admirably suited and easily programmed.

The dramatic post-war development of nonparametric methods is described by Noether (1984). Some idea of the volume of literature is given in the nonparametric bibliography compiled by Singer (1979). It considers work relevant to applications in just one subject – psychology. Work continues at an increasing pace.

1.2 HYPOTHESIS TESTS

We assume a basic familiarity with hypothesis testing like that implicit in the use of t-tests, but we summarize a few fundamentals and illustrate application to a simple nonparametric situation.

In Example 1.2 we wanted to test acceptability of a hypothesis that the median number of pages in all 114 books was 220. This implies the median is some other number if that hypothesis is not true, so our hypothesis that it equals 220 is something of a cockshy. It may have been suggested by past experience in assessing book lengths, or have been asserted with confidence by somebody else. We call this a **null hypothesis**, writing it H_0. If θ denotes the population median we often use the shorthand notation

$$H_0: \theta = 220$$

Our alternative hypotheses, collectively labelled H_1, are written

$$H_1: \theta \neq 220$$

We speak of testing H_0 against the **two-sided** (greater or less than) alternatives H_1.

The sign test in Example 1.2 involved a known distribution of signs if the null hypothesis H_0 were true, namely the binomial with $n = 12$ and $p = \frac{1}{2}$. The probabilities of each number of successes (here represented by plus signs) is tabulated; see, e.g. Neave (1981, p. 6). They are given (to three decimal places) in Table 1.1.

From Table 1.1 we see that if H_0 is true, 6 plus (hence also 6 minus) has maximum probability, and the probabilities fall off