Wiley Series in Probability and Statistics

THIRD EDITION

Nonparametric Statistical Methods

MYLES HOLLANDER
DOUGLAS A. WOLFE
ERIC CHICKEN

WILEY

Nonparametric Statistical Methods

Third Edition

Myles Hollander

Department of Statistics
Florida State University
Tallahassee, Florida

Douglas A. Wolfe

Department of Statistics
Ohio State University
Columbus, Ohio

Eric Chicken

Department of Statistics
Florida State University
Tallahassee, Florida

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Hollander, Myles.
Nonparametric statistical methods. –Third edition / Myles Hollander, Department of Statistics, Florida State University, Tallahassee, Florida, Douglas A. Wolfe, Department of Statistics, Ohio State University, Columbus, Ohio, Eric Chicken, Department of Statistics, Florida State University, Tallahassee, Florida.
pages cm
Includes bibliographical references and indexes.
ISBN 978-0-470-38737-5 (hardback)
1. Nonparametric statistics. I. Wolfe, Douglas A. II. Chicken, Eric, 1963–III. Title.
QA278.8.H65 2013
519.5–dc23

2013007061

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Nonparametric Statistical Methods

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

To our wives,
Glee, Marilyn, and Rebecca.

Preface

The nonparametric approach is the preferred methodology for statisticians and other scientists. We list some of its advantages in Section 1.1. In the third edition, we retain our emphasis on applications to real-world situations. We want our readers to learn how to apply nonparametric techniques in a variety of settings and to understand the assumptions underlying the methods.

In this third edition, we have improved the 11 chapters of the second edition and added five new chapters. The new chapters cover topics of recent and current interest, namely, density estimation, wavelets, smoothing, ranked set sampling, and Bayesian nonparametrics. R programs are now used to perform calculations. See Section 1.5 for a description of R.

The second edition was used primarily for a one-semester senior undergraduate/first-year graduate course for students having had a prior course in statistics. With the added coverage here, there is ample material for a two-semester course. Nevertheless, we expect most teachers will still opt for a one-semester course and choose specific chapters in accordance with their interests and those of their students.

Many friends and colleagues have helped us with this project.

Grant Schneider, a graduate student at the Ohio State University, provided invaluable support in the conversion from complete reliance on null distribution tables in our second edition to the exclusive use of R programs to obtain appropriate critical values and compute associated P-values in this third edition. He wrote new R programs for all of the statistical procedures in Chapter 15 and for a majority of the many procedures in Chapters 5–7, and modified existing programs for the other procedures in those three chapters, leading to significantly improved computational speed in most cases. He also organized all of the R programs used in this third edition into a documented collection that is formally registered as an R package specifically linked to this third edition. We owe Grant a special thanks for his leadership role in this important aspect of our new edition.

Rachel Becvarik wrote new R programs for Chapters 11 and 16 and provided a spark.

Jelani Wiltshire and Michael Rosenthal assisted with LaTeX typesetting.

James Stricherz provided computing support and Pamela McGhee and Marylou Tatis provided office support.

Our editors Steve Quigley, Susanne Steitz-Filler, and Sari Friedman were dedicated from the inception to the completion. Our production manager Melissa Yanuzzi carefully guided the manuscript through the production process.

To all these helpmates, we are very grateful.

Tallahassee, Florida
Columbus, Ohio
Tallahassee, Florida
August 2013

MYLES HOLLANDER
DOUGLAS A. WOLFE
ERIC CHICKEN

Contents

Chapter 1

Introduction

1.1 ADVANTAGES OF NONPARAMETRIC METHODS

Roughly speaking, a nonparametric procedure is a statistical procedure that has certain desirable properties that hold under relatively mild assumptions regarding the underlying populations from which the data are obtained. The rapid and continuous development of nonparametric statistical procedures over the past $7\frac{1}{2}$ decades is due to the following advantages enjoyed by nonparametric techniques:

1. Nonparametric methods require few assumptions about the underlying populations from which the data are obtained. In particular, nonparametric procedures forgo the traditional assumption that the underlying populations are normal.
2. Nonparametric procedures enable the user to obtain exact P-values for tests, exact coverage probabilities for confidence intervals, exact experimentwise error rates for multiple comparison procedures, and exact coverage probabilities for confidence bands without relying on assumptions that the underlying populations are normal.
3. Nonparametric techniques are often (although not always) easier to apply than their normal theory counterparts.
4. Nonparametric procedures are often quite easy to understand.
5. Although at first glance most nonparametric procedures seem to sacrifice too much of the basic information in the samples, theoretical efficiency investigations have shown that this is not the case. Usually, the nonparametric procedures are only slightly less efficient than their normal theory competitors when the underlying populations are normal (the home court of normal theory methods), and they can be mildly or wildly more efficient than these competitors when the underlying populations are not normal.
6. Nonparametric methods are relatively insensitive to outlying observations.
7. Nonparametric procedures are applicable in many situations where normal theory procedures cannot be utilized. Many nonparametric procedures require just the ranks of the observations, rather than the actual magnitude of the observations, whereas the parametric procedures require the magnitudes.
8. The Quenouille–Tukey jackknife (Quenouille (1949), Tukey (1958, 1962)) and Efron's computer-intensive (1979) bootstrap enable nonparametric approaches to be used in many complicated situations where the distribution theory

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

needed to support parametric methods is intractable. See Efron and Tibshirani (1994).

9. Ferguson's Dirichlet process (1973) paved the way to combine the advantages of nonparametric methods and the use of prior information to form a Bayesian nonparametric approach that does not require distributional assumptions.
10. The development of computer software has facilitated fast computation of exact and approximate P-values for conditional nonparametric tests.

1.2 THE DISTRIBUTION-FREE PROPERTY

The term *nonparametric*, introduced in Section 1.1, is imprecise. The related term *distribution-free* has a precise meaning. The distribution-free property is a key aspect of many nonparametric procedures. In this section, we informally introduce the concept of a distribution-free test statistic. The related notions of a distribution-free confidence interval, distribution-free multiple comparison procedure, distribution-free confidence band, asymptotically distribution-free test statistic, asymptotically distribution-free multiple comparison procedure, and asymptotically distribution-free confidence band are introduced at appropriate points in the text.

Distribution-Free Test Statistic

We introduce the concept of a distribution-free test statistic by referring to the two-sample Wilcoxon rank sum statistic, which you will encounter in Section 4.1.

The data consist of a random sample of m observations from a population with continuous probability distribution F_1 and an independent random sample of n observations from a second population with continuous probability distribution F_2. The null hypothesis to be tested is

$$H_0 : F_1 = F_2 = F, \; F \text{ unspecified.}$$

The null hypothesis asserts that the two random samples can be viewed as a single sample of size $N = m + n$ from a common population with unknown distribution F. The Wilcoxon (1945) statistic W is obtained by ranking the combined sample of N observations jointly from least to greatest. The test statistic is W, the sum of the ranks obtained by the Y's in the joint ranking.

When H_0 is true, the distribution of W does not depend on F; that is, when H_0 is true, for all a-values, the probability that $W \leq a$, denoted by $P_0(W \leq a)$, does not depend on F.

$$P_0(W \leq a) \text{ does not depend on } F. \tag{1.1}$$

The distribution-free property given by (1.1) enables one to obtain the distribution of W under H_0 without specifying the underlying F. It further enables one to exactly specify the type I error probability (the probability of rejecting H_0 when H_0 is true) without making distributional assumptions, such as the assumption that F is a normal distribution; this assumption is required by the parametric t-test.

The details concerning how to perform the Wilcoxon test are given in Section 4.1.