PERMUTATION TESTS FOR COMPLEX DATA

Theory, Applications and Software

Fortunato Pesarin • Luigi Salmaso
University of Padua, Italy

A John Wiley and Sons, Ltd., Publication

PERMUTATION TESTS FOR COMPLEX DATA

WILEY SERIES IN PROBABILITY AND STATISTICS

PERMUTATION TESTS FOR COMPLEX DATA

Theory, Applications and Software

Fortunato Pesarin • Luigi Salmaso
University of Padua, Italy

A John Wiley and Sons, Ltd., Publication

This edition first published 2010

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloguing-in-Publication Data
Pesarin, Fortunato.
Permutation tests for complex data : theory, applications, and software / Fortunato Pesarin, Luigi Salmaso.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-51641-6 (cloth)
1. Statistical hypothesis testing. 2. Permutations. 3. Multivariate analysis. I. Salmaso, Luigi. II. Title.
QA277.P374 2010
519.5′6 – dc22

2010000776

A catalogue record for this book is available from the British Library.

ISBN: 978-0-470-51641-6

Typeset in 9/11 Times by Laserwords Private Limited, Chennai, India.
Printed and bound in the United Kingdom by Antony Rowe Ltd, Chippenham, Wiltshire.

To:
Annamaria, Albertina, Annalisa and Alessandro Pesarin
and
Davide, Emanuele, Rosa, Paolina and Serio Salmaso

The obvious is that which is never seen
until someone exposes it simply.
Kahlil Gibran

Contents

Preface

This book deals with the combination-based approach to permutation hypothesis testing in several complex problems frequently encountered in practice. It also deals with a wide range of difficult applications in easy-to-check conditions. The key underlying idea, on which the large majority of testing solutions in multidimensional settings are based, is the nonparametric combination (NPC) of a set of dependent partial tests. This methodology assumes that a testing problem is properly broken down into a set of simpler sub-problems, each provided with a proper permutation solution, and that these sub-problems can be jointly analysed in order to maintain underlying unknown dependence relations.

The first four chapters are devoted to the theory of univariate and multivariate permutation tests, which has been updated. The remaining chapters present real case studies (mainly observational studies) along with recent developments in permutation solutions. Observational studies have enjoyed increasing popularity in recent years for several reasons, including low costs and availability of large data sets, but they differ from experiments because there is no control of the assignment of treatments to subjects. In observational studies the experimenter's main concern is usually to discover an association among variables of interest, possibly indicating one or more causal effects. The robustness of the nonparametric methodology against departures from normality and random sampling are much more relevant in observational studies than in controlled clinical trials. Hence, in this context, the NPC method is particularly suitable. Moreover, given that the NPC method is conditional on a set of sufficient statistics, it shows good general power behaviour, and the Fisher, Liptak or direct combining functions often have power functions which are quite close to the best parametric counterparts, when the latter are applicable, even for moderate sample sizes. Thus NPC tests are relatively efficient and much less demanding in terms of underlying assumptions with respect to parametric competitors and to traditional distribution-free methods based on ranks, which are generally not conditional on sufficient statistics and so rarely present better unconditional power behaviour. One major feature of the NPC with dependent tests, provided that the permutation principle applies, is that we must pay attention to a set of partial tests, each appropriate to the related sub-hypotheses, because the underlying dependence relation structure is nonparametrically and implicitly captured by the combining procedure. In particular, the researcher is not explicitly required to specify the dependence structure on response variables. This aspect is of great importance particularly for non-normal and categorical variables in which dependence relations are generally too difficult to define, and, even when well defined, are hard to cope with. Furthermore, in the presence of a stratification variable, NPC through a multi-phase procedure allows for quite flexible solutions. For instance, we can firstly combine partial tests with respect to variables within each stratum and then combine the combined tests with respect to strata. Alternatively, we can first combine partial tests related to each variable with respect to strata and then combine the combined tests with respect to variables. Moreover, once a global inference is found significant, while controlling for multiplicity it is possible to recover which partial inferences are mostly responsible of that result.

Although dealing with essentially the same methodology as contained in Pesarin (2001), almost all the material included in this book is new, specifically with reference to underlying theory and case studies.

Chapter 1 contains an introduction to general aspects and principles concerning the permutation approach. The main emphasis is on the principles of conditionality, sufficiency and similarity, relationships between conditional and unconditional inferences, why and when conditioning may be necessary, why the permutation approach results from both conditioning with respect to the data set and exchangeability of data in the null hypothesis, etc. Moreover, permutation techniques are discussed along with computational aspects. Basic notation is then introduced. Through a heuristic discussion of simple examples on univariate problems with paired data, two-sample and multi-sample (one-way ANOVA) designs, the practice of permutation testing is introduced. Moreover, discussions on conditional Monte Carlo (CMC) methods for estimating the distribution of a test statistic and some comparisons with parametric and nonparametric counterparts are also presented.

Chapters 2 and 3 formally present: the theory of permutation tests for one-sample and multi-sample problems; proof and related properties of conditional and unconditional unbiasedness; the definition and derivation of conditional and unconditional power functions; confidence intervals for treatment effect δ; the extension of conditional inferences to unconditional counterparts; and a brief discussion on optimal permutation tests and of the permutation central limit theorem.

Chapter 4 presents multivariate permutation testing with the NPC methodology. It includes a discussion on assumptions, properties, sufficient conditions for a complete theory of the NPC of dependent tests, and practical suggestions for making a reasonable selection of the combining function to be used when dealing with practical problems. Also discussed are: the concept of finite-sample consistency, especially useful when the number of observed variables in each subject exceeds that of subjects in the study; the multi-aspect approach; separate testing for two-sided alternatives; testing for multi-sided alternatives; the Behrens–Fisher problem, etc.

Chapter 5 deals with multiple comparisons and multiple testing issues. A brief overview of multiple comparison procedures (MCPs) is presented. The main focus is on closed testing procedures for multiple comparisons and multiple testing. Some hints are also given with reference to weighted methods for controlling family-wise error (FWE) and false discovery rate (FDR), adjustment of stepwise p-values, and optimal subset procedures.

Chapter 6 concerns multivariate permutation approaches for categorical data. A natural multivariate extension of McNemar's test is presented along with the multivariate goodness-of-fit test for ordered variables, the multivariate analysis of variance (MANOVA) test with nominal categorical data, and the issue of stochastic ordering in the presence of multivariate categorical ordinal variables. A permutation approach to test allelic association and genotype-specific effects in the genetic study of a disease is also discussed. An application concerning how to establish whether the distribution of a categorical variable is more heterogeneous (less homogeneous) in one population than in another is presented as well.

Chapter 7 discusses some quite particular problems with repeated measurements and/or missing data. Carry-over effects in repeated measures designs, modelling and inferential issues are treated extensively. Moreover, testing hypothesis problems for repeated measurements and missing data are examined. The rest of the chapter is devoted to permutation testing solutions with missing data.

Chapter 8 refers to permutation approaches for hypothesis testing when a multivariate monotonic stochastic ordering is present (with continuous and/or categorical variables). Umbrella testing problems are also presented. Moreover, two applications are discussed: one concerning the comparison of cancer growth patterns in laboratory animals and the other referring to a functional observational battery study designed to measure the neurotoxicity of perchloroethylene, a solvent used in dry cleaning (Moser, 1989, 1992).

Chapter 9 is concerned with permutation methods for problems of hypothesis testing in the framework of survival analysis.

Chapter 10 deals with statistical shape analysis. Most of the inferential methods known in the shape analysis literature are parametric in nature. They are based on quite stringent assumptions, such as the equality of covariance matrices, the independency of variation within and among landmarks or the multinormality of the model describing landmarks. But, as is well known, the assumption of equal covariance matrices may be unreasonable in certain applications, the multinormal model in the tangent space may be doubted and sometimes there are fewer individuals than landmarks, implying over-dimensioned spaces and loss of power. On the strength of these considerations, an extension of NPC methodology to shape analysis is suggested. Focusing on the case of two independent samples, through an exhaustive comparative simulation study, the behaviour of traditional tests along with nonparametric permutation tests using multi-aspect procedures and domain combinations is evaluated. The case of heterogeneous and dependent variation at each landmark is also analysed, along with the effect of superimposition on the power of NPC tests.

Chapter 11 presents two interesting real case studies in ophthalmology, concerning complex repeated measures problems. For each data set, different analyses have been proposed in order to highlight particular aspects of the data structure itself. In this way we enable the reader to choose the most appropriate analysis for his/her research purposes. The autofluorescence case study concerns a clinical trial in which patients with bilateral age-related macular degeneration were evaluated. In particular, their eyes were observed at several different and fixed positions. Hence, repeated measures issues arise. Five outcome variables were recorded and analysed. The confocal case study concerns a clinical trial with a five-year follow-up period, aiming to evaluate the long-term side-effects of a drug. Fourteen variables and four domains in total were analysed.

Chapter 12 deals with case studies in the field of survival analysis and epidemiology. NPC Test R10 software, SAS, MATLAB® and R codes have been used to perform the analyses. A comparison between logistic regression and NPC methodology in exploratory studies is then provided.

One of the main features of this book is the provision of several different software programs for performing permutation analysis. Various programs have been specifically developed. In particular:

- NPC Lib MATLAB library has been developed by Livio Finos, with consulting team Rosa Arboretti, Francesco Bertoluzzo, Stefano Bonnini, Chiara Brombin, Livio Corain, Fortunato Pesarin, Luigi Salmaso and Aldo Solari. For updates on the NPC Lib MATLAB library we refer to `http://homes.stat.unipd.it/livio`.
- NPC Test Release 10 (R10) standalone software (which is an extended version of the former NPC Test 2.0© produced by Methodologica S.r.l. and designed by Luigi Salmaso) has been updated by Luigi Salmaso, Andrey Pepelyshev, Livio Finos and Livio Corain, with consulting team Rosa Arboretti, Stefano Bonnini, Federico Campigotto and Fortunato Pesarin. For further updates to the NPC Test software we refer to `http://www.gest.unipd.it/~salmaso`.
- R code developed by Dario Basso, with consulting team Stefano Bonnini, Chiara Brombin, Fortunato Pesarin and Luigi Salmaso.
- SAS macros developed by Rosa Arboretti and Luigi Salmaso, with consulting team Stefano Bonnini, Federico Campigotto, Livio Corain and Fortunato Pesarin.

The above software is available from the book's website, `http://www.wiley.com/go/npc`. Raw data for all examples presented in the book, along with corresponding software routines, are also available from the website. Any errata, corrigenda or updates related to theory and software will be posted at `http://www.gest.unipd.it/~salmaso`.

We would like to express our thanks to the members of the Nonparametric Research Group at the University of Padua for their research collaboration on different topics included in this book: Rosa Arboretti, Dario Basso, Stefano Bonnini, Francesco Bertoluzzo, Chiara Brombin, Federico Campigotto, Livio Corain, Francesco Dalla Valle, Livio Finos, Patrizia Furlan, Susanna Ragazzi, Monjed Samuh, Aldo Solari and Francesca Solmi. We also wish to thank Susan Barclay, Richard

Davies, Beth Dufour, Heather Kay, Prachi Sinha-Sahay and the John Wiley & Sons Group in Chichester for their valuable publishing suggestions. Moreover, we owe a debt of thanks to our colleagues in different scientific fields: Ermanno Ancona, PierFrancesco Bassi, Fabio Bellotto, Patrizio Bianchi, Mario Bolzan, Carlo Castoro, Bruno Cozzi, Giovanni Fava, Roberto Filippini, Annarosa Floreani, Alessandro Frigiola, Luca Guarda-Nardini, Franco Grego, Lorenzo Menicanti, Edoardo Midena, Bruno Mozzanega, Virginia Moser, Andrea Peserico, Stefano Piaserico, Alberto Ruol, Luigi Sedea, Luca Spadoni, Tiziano Tempesta, Catherine Tveit, Carla Villanova, and several others with whom we have had stimulating discussions related to complex case studies; some of them are included on the book's website.

We wish to acknowledge Chiara Brombin for her valuable help through all stages of the production of this book and Professors N. Balakrishnan, O. Cucconi, P. Good, S. Kounias, V. Seshadri and J. Stoyanov as well as several colleagues for stimulating us in various ways to do research on multivariate permutation topics and to write the book. We also thank Stefania Galimberti, Ludwig Hothorn and Maria Grazia Valsecchi for revising some chapters.

In addition, we would like to acknowledge the University of Padua (CPDA088513/08) and the Italian Ministry of University and Research (PRIN 2008_2008WKHJPK_002) for providing the financial support for the necessary research and the development of part of the software.

Both authors share full responsibility for any errors or ambiguities, as well as for the ideas expressed throughout the book. A large part of the material presented in the book has been compiled from several publications and real case studies have been fully developed with the proposed different software codes. Although we have tried to detect and correct errors and eliminate ambiguities, there may well be others that have escaped our scrutiny. We take responsibility for and would warmly welcome notification of any that remain.

Finally, the second author (LS) wishes to acknowledge the first author (FP) as an inspiration for his open-mindedness and deep passion for innovative research in permutation methods during the course of our long-lasting collaboration and throughout the writing of this book.

We welcome any suggestions to the improvement of the book and would be very pleased if the book provides users with new insights into the analysis of their data.

Fortunato Pesarin
Department of Statistical Sciences
University of Padua

Luigi Salmaso
Department of Management and Engineering
University of Padua

Padova, January 2010

Notation and Abbreviations

A: an event belonging to the collection $\mathcal{A}$ of events

$\mathcal{A}$: a collection (algebra) of events

$\mathcal{A}_{/A} = \mathcal{A} \bigcap A$: a collection of events conditional on A

ANCOVA: analysis of covariance

ANOVA: analysis of variance

AUC: area under the curve

B: the number of conditional Monte Carlo iterations

$\mathcal{B}n(n, \theta)$: binomial distribution with n trials and probability θ of success in one trial

CDF: cumulative distribution function

CLT: central limit theorem

CMC: conditional Monte Carlo

$\mathbb{C}ov(X, Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X) \cdot \mathbb{E}(Y)$: the covariance operator on (X, Y)

CSP: constrained synchronized permutations

$\mathcal{C}y(\eta, \sigma)$: Cauchy distribution with location η and scale σ

d.f.: degrees of freedom

$\delta = \int_{\mathcal{X}} \delta(x) \cdot dF_X(x)$: the fixed treatment effect (same as δ-functional or pseudo-parameter), $\delta \in \Omega$

Δ: stochastic treatment effect

EDF: empirical distribution function: $\hat{F}_{\mathbf{X}}(t) = \hat{F}(t|\mathcal{X}_{/\mathbf{X}}) = \sum_i \mathbb{I}(X_i \leq t)/n,\ t \in \mathcal{R}^1$

EPM: empirical probability measure: $\hat{P}_{\mathbf{X}}(A) = \hat{P}(A|\mathcal{X}_{/\mathbf{X}}) = \sum_i \mathbb{I}(X_i \in A)/n,\ A \in \mathcal{A}$

ESF: empirical survival function (same as significance level): $\hat{L}_{\mathbf{X}}(t) = \hat{L}(t|\mathcal{X}_{/\mathbf{X}}) = \sum_i \mathbb{I}(X_i \geq t)/n,\ t \in \mathcal{R}^1$

$\mathbb{E}(X) = \int_{\mathcal{X}} x \cdot dF_X(x)$: the expectation operator (mean value) of X

$\mathbb{E}_A[X)] = \mathbb{E}[X|A] = \int_A x \cdot dF_X(x|A)$: the conditional expectation of X given A

$\overset{d}{=}$: equality in distribution: $X \overset{d}{=} Y \leftrightarrow F_X(z) = F_Y(z), \forall z \in \mathcal{R}^1$

$\overset{d}{>}$: stochastic dominance: $X \overset{d}{>} Y \leftrightarrow F_X(z) \leq F_Y(z), \forall z$ and $\exists A : F_X(z) < F_Y(z),\ z \in A$, with $\Pr(A) > 0$

$<\neq>$: means '$<$', or '$\neq$', or '$>$'

$\sim$: distributed as: e.g. $X \sim \mathcal{N}0, 1)$ means X follows ths standard normal distribution

$\approx$: permutationally equivalent to

FDR: false discovery rate

FWER: family-wise error rate

$f_P(z) = f(z)$: the density of a variable X, with respect to a dominating measure ξ and related to the probability P

$F_X(z) = F(z) = \Pr\{X \leq z\}$: the CDF of X
$F_{X|A}(z) = \Pr\{X \leq z|A\}$: the conditional CDF of $(X|X \in A)$
$F_T^*(z) = F^*(z) = \Pr\{T^* \leq z|\mathcal{X}_{/\mathbf{X}}\}$: the permutation CDF of T given $\mathbf{X}$
$\mathcal{H}_G(N, \theta, n)$: hypergeometric distribution with N the number of units, $\theta \cdot N$ the number of units of interest, n the sample size
i.i.d.: independent and identically distributed
$\mathbb{I}(A)$: the indicator function, i.e. $\mathbb{I}(A) = 1$ if A is true, and 0 otherwise
ITT: intention-to-treat principle
$\lambda = \Pr\left\{T \geq T^o|\mathcal{X}_{/\mathbf{X}}\right\}$: the attained p-value of test T on data set $\mathbf{X}$
$L_X(t) = L(t) = \Pr\{X \geq t\}$: the significance level function (same as the survival function)
$\boldsymbol{\mu} = \mathbb{E}(\mathbf{X})$: the mean value of vector $\mathbf{X}$
MAD: median of absolute deviations from the median
MANOVA: multivariate analysis of variance
MC: number of Monte Carlo iterations
MCP: multiple comparison procedure
MCAR: missing completely at random
$\mathbb{M}d(X) = \tilde{\mu}$: the median operator on variable X such that $\Pr\{X < \tilde{\mu}\} = \Pr\{X > \tilde{\mu}\}$
$\#(X \in A) = \sum_i \mathbb{I}(X_i \in A)$: number of points X_i belonging to A
n: the (finite) sample size
MNAR: missing not at random
MTP: multiple testing problem
$\mathcal{N}(\mu, \sigma^2)$: Gaussian or normal variable with mean μ and variance σ^2
$\mathcal{N}_V(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: V-dimensional normal variable ($V \geq 1$) with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$O(d_n) = c_n$: given two sequences $\{c_n\}$ and $\{d_n\}$, $O(d_n) = c_n$ if c_n/d_n is bounded as $n \to \infty$
$o(d_n) = c_n$: given two sequences $\{c_n\}$ and $\{d_n\}$, $o(d_n) = c_n$ if $c_n/d_n \to 0$ as $n \to \infty$
Ω: the set of possible values for δ
$\pi(\delta)$: the prior distribution of $\delta \in \Omega$
PCLT: permutation central limit theorem
P: a probability distribution on $(\mathcal{X}, \mathcal{A})$
$\mathcal{P}$: a family of probability distributions
$P(A) = \int_A dP(z)$: the probability of event $A \in \mathcal{A}$ with respect to P
p-FWE: adjusted p-value from a closed testing procedure
$\Pr\{A\}$: a probability statement relative to $A \in \mathcal{A}$
$\mathcal{R}^n$: the set of n-dimensional real numbers
$\mathbb{R}$: the rank operator
$R_i = \mathbb{R}(X_i) = \sum_{1 \leq j \leq n} \mathbb{I}(X_j \leq X_i)$: the rank of X_i within $\{X_1, \ldots, X_n\}$
SLF: sgnificance level function
UMP: uniformly most powerful
UMPS: uniformly most powerful similar
UMPU: uniformly most powerful unbiased
USP: unconstrained synchronized permutations
$\mathbb{V}(X) = \mathbb{E}(X-\mu)^2 = \sigma^2$: the variance operator on variable X
WORE: without replacement random experiment

WRE: with replacement random experiment
X: a univariate or multivariate random variable
$\mathbf{X}$: a sample of n units, $\mathbf{X} = \{X_i,\ i = 1, \ldots, n\}$
$\mathbf{X}^*$: a permutation of $\mathbf{X}$
$|\mathbf{X}| = \{|X_i|,\ i = 1, \ldots, n\}$: a vector of absolute values
XOR: exclusive or relationship: (A XOR B) means one but not both
$\mathcal{X}$: the sample space (or support) of variable X
$(\mathcal{X}, \mathcal{A})$: a measurable space
$(\mathcal{X}, \mathcal{A}, P)$: a probability space
$\mathcal{X}_{/\mathbf{X}}$: the orbit or permutation sample space given $\mathbf{X}$
$(\mathcal{X}_{/\mathbf{X}}, \mathcal{A}_{/\mathbf{X}})$: a permutation measurable space
$T \div \mathcal{X} \to \mathcal{R}^1$: a statistic
$T^o = T(\mathbf{X})$: the observed value of test statistic T evaluated on $\mathbf{X}$
$\mathbf{U}^{*\top}$: the transpose of $\mathbf{U}^*$
$\mathcal{U}(a, b)$: uniform distribution in the interval (a, b)
$\lfloor X \rfloor$: the integer part of X
$\biguplus$: the operator for pooling (concatenation) of two data sets: $\mathbf{X} = \mathbf{X}_1 \biguplus \mathbf{X}_2$
Z: the unobservable random deviates or errors: $X = \mu + Z$

1

Introduction

1.1 On Permutation Analysis

This book deals with the permutation approach to a variety of univariate and multivariate problems of hypothesis testing in a typical nonparametric framework. A large number of univariate problems may be usefully and effectively solved using traditional parametric or rank-based nonparametric methods as well, although under relatively mild conditions their permutation counterparts are generally asymptotically as good as the best parametric ones (Lehmann, 2009). It should also be noted that permutation methods are essentially of a nonparametric exact nature in a conditional context (see Proposition 2, 3.1.1 and Remarks 1, 2.2.4 and 1, 2.7). In addition, there are a number of parametric tests the distributional behaviour of which is only known asymptotically. Thus, for most sample sizes of practical interest, the relative lack of efficiency of permutation solutions may sometimes be compensated by the lack of approximation of parametric asymptotic counterparts. Even when responses follow the multivariate normal distribution and there are too many nuisance parameters to estimate and remove, due to the fact that each estimate implies a reduction of the degrees of freedom in the overall analysis, it is possible for the permutation solution to be more efficient than its parametric counterpart (note that 'responses', 'variables', 'outcomes', and 'end points' are often used as synonyms). In addition, assumptions regarding the validity of most parametric methods (such as homoscedasticity, normality, regular exponential family, random sampling, etc.) rarely occur in real contexts; so that consequent inferences, when not improper, are necessarily approximated and their approximations are often difficult to assess.

In practice parametric methods reflect a modelling approach and generally require the introduction of a set of stringent assumptions, which are often quite unrealistic, unclear, and difficult to justify. Sometimes these assumptions are merely set on an *ad hoc* basis for specific inferential analyses. Thus they appear to be mostly related to the availability of the methods one wishes to apply rather than to well-discussed necessities obtained from a rational analysis of reality, in accordance with the idea of modifying a problem so that a known method is applicable rather than that of modifying methods in order to properly deal with the problem. For instance, too often and without any justification researchers assume multivariate normality, random sampling from a given population, homoscedasticity of responses also in the alternative, etc., so that it becomes possible to write down a likelihood function and to estimate a variance–covariance matrix and so consequent inferences are without real credibility. In contrast, nonparametric approaches try to keep assumptions at a lower workable level, avoiding those that are difficult to justify or interpret, and possibly without excessive loss of inferential efficiency. Thus, they are based on more realistic foundations for statistical inference. And so they are intrinsically robust and resulting inferences are credible.

Permutation Tests for Complex Data: Theory, Applications and Software Fortunato Pesarin, Luigi Salmaso

However, there are many complex multivariate problems (commonly in areas such as agriculture, biology, business statistics, clinical trials, engineering, the environment, experimental data, finance data, genetics, industrial statistics, marketing, pharmacology, psychology, quality control, social sciences, zoology, etc.) that are difficult to solve outside the conditional framework and, in particular, outside the method of nonparametric combination (NPC) of dependent permutation tests (solutions to several complex problems are discussed in Chapter 4 and beyond).

Moreover, within parametric approaches it is sometimes difficult, if not impossible, to obtain proper solutions even under the assumption of normal errors. Some examples are:

1. problems with paired observations when scale coefficients depend on units;
2. two-sample designs when treatment is effective only on some of the treated subjects, as may occur with drugs having genetic interaction;
3. two-way ANOVA;
4. separate testing in cross-over designs;
5. multivariate tests when the number of observed variables is larger than the sample size;
6. jointly testing for location and scale coefficients in some two-sample experimental problems with positive responses;
7. exact testing for multivariate paired observations when some data are missing, even when not at random;
8. unconditional testing procedures when subjects are randomly assigned to treatments but are obtained by selection-bias sampling from the target population;
9. exact inference in some post-stratification designs;
10. two-sample testing when data are curves or surfaces, i.e. testing with countably many variables.

As regards problem 1, within a parametric framework it is impossible to obtain standard deviation estimates for observed differences on each unit with more than zero degrees of freedom, whereas exact and effective permutation solutions do exist (see Sections 1.9 and 2.6). A similar impossibility also occurs with Wilcoxon's signed rank test. In problem 2, since effects, either random or fixed, behave as if they depend on some unobserved attitudes of the subjects, traditional parametric approaches are not appropriate. Hints as to proper permutation solutions will be provided in Chapters 2, 3 and 4. In problem 3 it is impossible to obtain independent or even uncorrelated separate inferences for main factors and interactions because all related statistics are compared to the same estimate of the error variance (see Remark 8, 2.7). In addition, it is impossible to obtain general parametric solutions in unbalanced designs. We shall see in Example 8, 2.7 and Chapter 11 that, within the permutation approach, it is at least possible to obtain exact, unbiased and uncorrelated separate inferences in both balanced and unbalanced cases. Regarding problem 4, we will see in Remark 5, 2.1.2 that in a typical cross-over problem with paired data ($[A, B]$ in the first group and $[B, A]$ in the second group) two separate hypotheses on treatment effect ($X_B \stackrel{d}{=} X_A$) and on interaction due to treatment administration ($X_{AB} \stackrel{d}{=} X_{BA}$) are tested separately and independently. In problem 5 it is impossible to find estimates of the covariance matrix with more than zero degrees of freedom, whereas the NPC method discussed in Chapter 4 allows for proper solutions which, in addition, are often asymptotically efficient. In problem 6, due to its close analogy with the Behrens–Fisher problem, exact parametric solutions do not exist, whereas, based on concurrent multi-aspect testing, an exact permutation solution does exist, provided that positive data are assumed to be exchangeable in the null hypothesis and the two cumulative distribution functions (CDFs) do not cross in the alternative (see Example 8, 4.6). In problem 7 general exact parametric solutions are impossible unless missing data are missing completely at random and data vectors with at least one missing datum are deleted. In Section 7.10, within the NPC methodology, we will see a general approximate solution and one exact solution even when some of the data are missing not completely at random. In problem 8 any selection-biased mechanism usually produces

quite severe modifications to the target population distribution, hence unless the selection mechanism is well defined and the consequent modified distribution is known, no proper parametric inference to the target population is possible; instead, within the permutation approach we may properly extend conditional inferences to unconditional counterparts. Moreover, in cases where the minimal sufficient statistic in the null hypothesis is the whole set of observed data, although the likelihood model would depend on a finite set of parameters, univariate statistics capable of summarizing the necessary information do not exist, so that no parametric method can be claimed to be uniformly better than others; indeed, conditioning on the pooled data set, i.e. considering the permutation counterpart, improves the power behaviour of any test statistics (see Cox and Hinkley, 1974; Lehmann, 1986). However, in order to attenuate the loss of information associated with using one overall statistic, we will find solutions within the so-called multi-aspect methodology based on the NPC of several dependent permutation test statistics, each capable of summarizing information on a specific aspect, so that it takes account of several complementary viewpoints (see Example 3, 4.6) and improves interpretability of results. In problem 9, as far as we know, the exact parametric inference for post-stratification analysis is based on the combination for independent partial tests (one test per stratum), provided that their null continuous distributions are known exactly. In problem 10, as far as can be seen from the literature (see Ramsay and Silverman, 2002; Ferraty and Vieu, 2006), only some regression estimate and predictive problems are solved when data are curves; instead, within the NPC strategy, several testing problems with countably many variables (the coefficients of suitable curve representations) can be efficiently solved.

Although authoritative, we agree only partially with opinions such as that expressed by Kempthorne (1955): 'When one considers the whole problem of experimental inference, that is of tests of significance, estimation of treatment differences and estimation of the errors of estimated differences, there seems little point in the present state of knowledge in using a method of inference other than randomization analysis.'

We agree with the part that stresses the importance for statisticians of referring to conditional procedures of inference and, in particular, to randomization (i.e. permutation) methods. Indeed, there is a wide range of inferential problems which are correctly and effectively solved within a permutation framework; however, there are others which are difficult or perhaps impossible to solve outside it.

We partially disagree, however, because there are very important families of inferential problems, especially connected to unconditional parametric estimation and testing, or to nonparametric prediction, classification, kernel estimation, or more generally within the statistical decision approach, which cannot be dealt with and/or solved in a permutation framework. These are often connected to violations of the so-called exchangeability condition (see Chapter 2) or are related to analysis of too few observed subjects. Moreover, all procedures of exploratory data analysis and all testing methods for which we cannot assume exchangeability of the data with respect to groups (i.e. samples) in the null hypothesis, generally lie outside the permutation approach. In addition, the traditional Bayesian inference (see Remark 4, 3.4, for suggestions on a *permutation Bayesian* approach) also lies outside the permutation approach.

Thus, although we think that permutation methods should be in the tool-kit of every statistician interested in applications, methodology or theory, we disagree because we do not believe that *all* inferential problems of interest for analysis of real problems fall within the permutation approach. In order to apply permutation methods properly, a set of initial conditions must be assumed, and if these conditions are not satisfied, their use may become erroneous.

However, and following remarks made by Berger (2000), these arguments support our decision to develop methods in the area of permutation testing, especially for multivariate complex problems. In this sense, this book attempts to go deeper into the main aspects of conditional methods of inference based on the permutation approach and to systematically study proper solutions to a set of important problems of practical interest. Section 1.4 lists a brief set of circumstances in which conditional testing procedures may be effective or even unavoidable.

1.2 The Permutation Testing Principle

For most problems of hypothesis testing, the observed data set $\mathbf{x} = \{x_1, \ldots, x_n\}$ is usually obtained by a *symbolic* experiment performed n times on a population variable X, and taking values in the sample space $\mathcal{X}$. We sometimes add the word 'symbolic' to names such as experiments, treatments, treatment effects, etc., in order to refer to experimental, pseudo-experimental and observational contexts. For the purposes of analysis, the data set $\mathbf{x}$ is generally partitioned into *groups* or *samples*, according to the so-called *treatment levels* of the symbolic experiment. In the context of the discussion up to and including Section 1.6, we use capital letters for random variables and lowercase for the observed data set. From Section 1.7 onwards, we shall dispense with this distinction, in that only capital letters will be used because the context is always sufficiently clear. Of course, when a data set is observed at its $\mathbf{x}$ value, it is presumed that a sampling experiment on a given underlying population has been performed, so that the resulting sample distribution is related to that of the parent population P. This is, of course, common to any statistical problem, and not peculiar to the permutation framework.

For any general testing problem in the null hypothesis, denoted by H_0, which typically assumes that data come from *only one* (with respect to groups) unknown population distribution P, $H_0 : \{X \sim P \in \mathcal{P}\}$, say, the whole set of observed data $\mathbf{x}$ is considered to be a random sample, taking values in the sample space $\mathcal{X}^n$, where $\mathbf{x}$ is one observation of the n-dimensional sample variable $X^{(n)}$ and where this random sample does not necessarily possess independent and identically distributed (i.i.d.) components (see Chapters 2 and 3 for more details).

We note that the observed data set $\mathbf{x}$ is always a set of sufficient statistics in H_0 for whatever underlying distribution. In order to see this in a simple way, let us assume that H_0 is true and all members of a nonparametric family $\mathcal{P}$ of non-degenerate and distinct distributions are dominated by one *dominating* measure $\xi_{\mathcal{P}}$; moreover, let f_P denote the density of P with respect to $\xi_{\mathcal{P}}$, and $f_P^{(n)}(\mathbf{x})$ denote the density of the sample variable $X^{(n)}$. As the identity $f_P^{(n)}(\mathbf{x}) = f_P^{(n)}(\mathbf{x}) \cdot 1$ is true for all $\mathbf{x} \in \mathcal{X}^n$, except for points such that $f_P^{(n)}(\mathbf{x}) = 0$, due to the well-known factorization theorem, any data set $\mathbf{x}$ is therefore a sufficient set of statistics for whatever member P of the nonparametric family $\mathcal{P}$.

1.2.1 Nonparametric Family of Distributions

Let us consider the following definition.

Definition 1. *A family of distributions $\mathcal{P}$ is said to behave nonparametrically when it is not possible to find a finite-dimensional space Θ such that there is a one-to-one relationship between Θ and $\mathcal{P}$, in the sense that each member P of $\mathcal{P}$ cannot be identified by only one member θ of Θ, and vice versa*.

If of course such a one-to-one relationship exists, θ is called a parameter, Θ is the parameter space, and $\mathcal{P}$ the corresponding parametric family. Families of distributions which are either unspecified or specified except for an infinite number of unknown parameters do satisfy the definition and so are nonparametric. Definition 1 also includes all those situations where the sample size n is smaller than the number of parameters, even though this is finite. All nonparametric families $\mathcal{P}$ which are of interest in permutation analysis are assumed to be sufficiently *rich* in such a way that if x and x' are any two distinct points of $\mathcal{X}$, then $x \neq x'$ implies $f_P(x) \neq f_P(x')$ for at least one $P \in \mathcal{P}$, except for points with null density for P. Also note that the characterization of a family $\mathcal{P}$ as being nonparametric essentially depends on the knowledge we assume about it. When we assume that the underlying family $\mathcal{P}$ contains all continuous distributions, then the data set $\mathbf{x}$ is *complete minimal sufficient*.

Permutation tests are known to be conditional procedures of inference, where conditioning is done with respect to a set of sufficient statistics in the null hypothesis. Thus consequent inferences at least concern the sample data $\mathbf{x}$ actually observed and the related observed subjects. The act of conditioning on a set of sufficient statistics in H_0, and the assumption of exchangeability *with respect to groups* (samples) for observed data, make permutation tests independent of the underlying likelihood model related to P (see Section 2.1.3). As a consequence, P may be unknown or unspecified, either in some or all of its parameters, or even in its analytic form. We specify this concept in the permutation testing principle.

1.2.2 The Permutation Testing Principle

Let us consider the following definition.

Definition 2. *If two experiments, taking values on the same sample space $\mathcal{X}$ and respectively with underlying distributions P_1 and P_2, both members of $\mathcal{P}$, give the same data set $\mathbf{x}$, then the two inferences conditional on $\mathbf{x}$ and obtained using the same test statistic must be the same, provided that the exchangeability of data with respect to groups is satisfied in the null hypothesis. Consequently, if two experiments, with underlying distributions P_1 and P_2, give respectively $\mathbf{x}_1$ and $\mathbf{x}_2$, and $\mathbf{x}_1 \neq \mathbf{x}_2$, then the two conditional inferences may be different.*

One of the most important features of the permutation testing principle is that in theory and under a set of mild conditions conditional inferences can be extended unconditionally to all distributions P of $\mathcal{P}$ for which the density with respect to a suitable dominating measure ξ is positive, i.e. $dP(\mathbf{x})/d\xi^n > 0$ (see Section 3.5). It should be emphasized, however, that this feature derives from the sufficiency and conditionality principles of inference (see Cox and Hinkley, 1974; Lehmann, 1986; Berger and Wolpert, 1988), by which inferences are related to all populations sharing the same value of conditioning statistics, particularly those which are sufficient for underlying nuisance entities. For instance, Student's t extends inference to all normal populations which assign positive density to the variance estimate $\hat{\sigma}^2$ and so its inference is for a family of distributions. Therefore, such unconditional extensions should be carried out carefully. Another important feature occurs in multivariate problems, when solved through NPC methods. For these kinds of problems, especially when they are complex and in very mild and easy-to-check conditions (see Chapter 4), it is not necessary to specify or to model the structure of dependence relations for the variables in the underlying population distribution. In this way analysis becomes feasible and results are easy to interpret. For instance, it is known that, for multivariate categorical variables, it is extremely difficult to properly model dependence relations among variables (see Joe, 1997). In practice, therefore, except for very particular cases, only univariate (or component variable by component variable) problems are considered in the literature. From Chapter 4 onwards we will see that, within the permutation testing principle and the NPC of dependent partial tests, a number of rather difficult problems can be effectively and easily solved, provided that partial tests are *marginally unbiased and consistent* (see Section 4.2.1). Also of interest is an application of this principle in the context of the Bayesian permutation approach (see Remark 4, 3.4).

However, the conditioning on sufficient statistics provides permutation tests with good general properties. Among the most important of these, when exchangeability is satisfied in the null hypothesis, is that permutation tests are always exact procedures (see Remark 1, 2.2.4 and Proposition 2, 3.1.1). Another property is that their conditional rejection regions are similar, as intended by Scheffé (1943a, 1943b). The former means that, at least in principle, the null rejection probability can be calculated exactly in all circumstances. The latter means that, if data comes from continuous distributions (where the probability of finding ties in the data set is zero), the null rejection probability is invariant with respect to observed data set $\mathbf{x}$, for almost all $\mathbf{x} \in \mathcal{X}^n$, and with respect to

the underlying population distribution P (see Chapter 2). As a consequence, conditional rejection regions are similar to the unconditional region. When data comes from non-continuous distributions, unless referring to randomized tests (see Section 2.2), the similarity property is only asymptotically valid. Moreover, if the stochastic dominance condition is satisfied in H_1, permutation tests based on divergence of suitable statistics are *conditionally unbiased* procedures, since the rejection probability of any test T, for all data sets $\mathbf{x} \in \mathcal{X}^n$, satisfies the relation $\Pr\{\lambda(\mathbf{x}(\delta)) \leq \alpha | \mathbf{x}\} = W(\delta, \alpha, T|\mathbf{x}) \geq \alpha$, where $\lambda(\mathbf{x}(\delta))$ indicates the p-value and $W(\delta, \alpha, T|\mathbf{x})$ indicates the *conditional power* of T given $\mathbf{x}$ with fixed treatment effect δ and significance level α (see Section 3.2).

It is worth noting that when exchangeability may be assumed in H_0, the similarity and unbiasedness properties allow for a kind of weak extension of conditional to unconditional inferences, irrespective of the underlying population distribution and the way sample data are collected. Therefore, this weak extension may be made for any sample data, even if they are not collected by well-designed sampling procedures, in which each unit is truly selected at random from a given population and subject to an experiment. Conversely, parametric solutions permit proper extensions only when data comes from well-designed sampling procedures on well-specified parent populations. Specifically, a general situation for unconditional extensions in parametric contexts occurs when likelihood functions are known except for nuisance parameters, and these are removed by invariant statistics or by conditioning on boundedly complete estimates (see Section 3.5 and Remark 2 therein).

For this reason, permutation inferences are proper with most observational data (sometimes called non-experimental), with experimental data, with selection-biased sampling procedures, and with well-designed sampling procedures. However, we must note that well-designed sampling procedures are quite rare even in most experimental problems (see Ludbrook and Dudley, 1998). For instance, if we want to investigate the effects of a drug on rats, the units to be treated are usually not randomly chosen from the population of all rats, but are selected in some way among those available in a laboratory and are *randomly assigned* to the established treatment levels. The same occurs in most clinical trials, where some patients, present in a hospital and that comply with the experiment, are randomly assigned to one of the pre-established treatment levels.

In one sense, the concept of random sampling is rarely achieved in real applications because, for various reasons, real samples are quite often obtained by selection-bias procedures. This implies that most of the forms of unconditional inferences usually associated with parametric tests, being based on the concept of random sampling, are rarely applicable in real situations. In addition, due to the similarity and unbiasedness properties, permutation solutions allow for relaxation of most of the common assumptions needed by parametric counterparts, such as the existence of mean values and variances, and the homoscedasticity of responses in the alternative hypothesis (see also Section 1.4). This is why permutation inferences are so important for both theoretical and application purposes, not only for their potential exactness.

Many authors have emphasized these aspects. A review of the relevant arguments is given in Edgington (1995), Edgington and Onghena (2007), and in Good (2000, 2005). One of these relates to the fact that reference null distributions of ordinary parametric tests are explicitly based on the concept of infinitely repeated and well-designed random sampling from a given well-specified population, the existence of which is often merely virtual. Another argument relates to the fact that, as occurs in many experimental problems, it is often too unrealistic to assume that treatment does not also influence scale coefficients or other distributional aspects, so that traditional parametric solutions may become improper.

Conversely, when exchangeability may be assumed in H_0, reference null distributions of permutation tests always exist because, at least in principle, they are obtained by enumerating all permutations of available data (see Chapter 2). In addition, permutation comparisons of means or of other functionals do not require homoscedasticity in the alternative, provided that underlying CDFs are ordered so that they do not cross each other (see Section 2.1.1). For these reasons, on the