1

Introduction

1.1 On Permutation Analysis

This book deals with the permutation approach to a variety of univariate and multivariate problems of hypotheses testing in a typical nonparametric framework. A large number of univariate problems are usefully and effectively solved by standard parametric or rank-based nonparametric methods, although in 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. In addition, there is a number of parametric tests the distributional behaviour of which are only known asymptotically. Thus, for most sample sizes of practical interest, the slight lack of efficiency of permutation solutions, with respect to the best, if any, may sometimes be compensated by the lack of approximation of parametric asymptotic counterparts. In this respect, especially when the rate of convergence of some parametric tests to their asymptotic distributions is unknown or is too slow, there are tests whose probability to reject H_0 when H_1 is true is smaller than α, $\Pr\{\textit{Reject } H_0|H_1\} < \alpha$ say, thereby violating a fundamental requisite of significance tests. An important case, connected with the testing for equivalence, is discussed in Example 4.11. Moreover, when responses are homoscedastic multivariate normal, there are too many nuisance parameters to estimate and remove since each estimate implies a reduction of one degree of freedom in the overall analysis, e.g. as with Hotelling's T^2, it is possible for some permutation solutions (Remark 4.35) to be even infinitely more efficient than their parametric counterparts. Moreover, assumptions regarding the validity of most traditional parametric methods, such as homoscedasticity, normal distribution, regular exponential family, random sampling from a given population, and treatment effect only on the parameter of interest, rarely occur in real contexts. Hence, consequent inferences, when not improper, at most are necessarily approximated and their approximations are often difficult to assess.

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

Companion website: www.wiley.com/go/permutationtests2e

In practice, parametric methods reflect a modelling approach and generally require the introduction of a number of stringent assumptions, which are sometimes unrealistic, unclear and difficult to justify. Often these assumptions are merely set on an *ad hoc* basis so as to facilitate specific analyses by well-established methods. Thus, they appear as mostly related to the availability of methods one wishes to apply than with well-discussed necessities derived from a rational analysis of reality, according to the idea of modifying a problem so that a known method becomes applicable than to modify known methods in order to properly dealing with the problem. For instance, too often and without any justification, researchers assume multivariate normality even when regressions are not linear; random sample from a given population even when a selection-bias process is used; homoscedasticity of responses even when treatment(s) have non-null effect(s) also on dispersion(s); and so on. So as it is possible to write down a likelihood to work with and to estimate all its nuisance parameters, consequent inferences might have no real credibility. On the contrary, nonparametric approaches try to keep assumptions at a lower workable level, avoiding those that are difficult to justify or interpret, and possibly without important loss of inferential efficiency, if any. Indeed, they are based on more realistic foundations and are intrinsically robust, therefore, consequent inferences are credible.

In addition, parametric tests are used to remove nuisance parameters by conditioning on their estimates, and when these are based on sufficient statistics, their conclusions are essentially about the same as those based on their permutation – conditional – counterparts (Section 2.6). However, there are many complex multivariate studies – quite common in: agriculture, biology, business, chemometrics, clinical trials, engineering, the environment, experimental and/or observational data, finance, genetics, industry, marketing, pharmacology, psychology, quality control, social sciences, zoology, etc. – that are difficult to solve outside the conditional approach and, in particular, outside the method of Nonparametric Combination (NPC) of dependent permutation tests. Solutions to several of such problems are discussed in Chapter 4 and onwards. Moreover, within parametric approaches, it is difficult, if not impossible, to have 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 layouts when treatment is effective only on some of the treated subjects, as it may occur with some drugs having genetic interaction;
3) the two-way ANalysis Of VAriance (ANOVA);
4) separate testing in cross-over designs;
5) multivariate tests when the number V of observed variables is – even infinitely – larger than sample size n: $n \ll V$;
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 not at random;
8) unconditional testing procedures when subjects are randomly assigned to treatments but are obtained from the target population by selection-bias samples;
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;
11) two-sample testing when the null is an interval of equivalent points;
12) two-sample testing for ordered categorical repeated measurements with time-dependent transition matrices.

Regarding Problem 1, within the parametric approach it is impossible to obtain estimates of standard deviation for observed differences on each unit with more than zero degree of freedom, whereas exact and effective permutation solutions do exist (Sections 1.9 and 3.1). A similar impossibility also occurs with Wilcoxon's signed rank test. In Problem 2, since either random or fixed effects behave as if they depend on some unobserved attitudes of the subjects, traditional parametric approaches are not appropriate. Hints for proper permutation solutions are provided in Chapters 2 and 3.

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 (Remark 3.12). In addition, it is impossible to obtain general parametric solutions in unbalanced designs. We shall see in Example 3.17 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 4, we will see in Remark 2.6 that in a typical cross-over problem with paired data, $[A, B]$ in the first and $[B, A]$ in the second sample, two separate hypotheses on treatment effect ($X_B \overset{d}{=} X_A$) and on the interaction due to treatment administration ($X_{AB} \overset{d}{=} X_{BA}$) are independently tested. In 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 permits proper solutions that, in addition, are often asymptotically efficient. In 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 under the alternative (Example 4.13). In 7, general 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 7.9, within the NPC methodology, we will see an example of permutation solution even when missingness is not completely at random.

In 8, every selection-biased mechanism may generate quite severe modifications to the distribution of selected population; hence, unless the selection mechanism is well defined, the related modified distribution is known except for estimable nuisance entities and an invariance property for nuisance parameters takes place, no proper parametric inference to the target population is possible; instead, within the permutation approach, we may properly extend – extrapolate, generalize – conditional to unconditional – population – inferences under very general and easy to justify conditions (Sections 1.3.2 and 2.9). Moreover, in cases where the minimal sufficient statistic is the whole set of observations, and so it is n-dimensional, although the likelihood model would depend on a finite set of parameters, univariate statistics fitted to summarize the necessary information do not exist. Thus, no test statistic can be claimed to be uniformly better than others. To attenuate the loss of information associated with using only one overall statistic, it is possible to find solutions within the so-called multi-aspect methodology based on a list of different statistics, each suitable to summarize information on a specific aspect of interest. Actually, these different statistics, if combined with the NPC method, may take account of several complementary points of view (Example 4.8) and may improve power and interpretability of results. In 9, as far as we know, the exact parametric inference for post-stratification analysis is based on the combination of independent partial tests – one test per stratum, provided that their null continuous distributions are known exactly. In 10, as far as it can be seen from the literature (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, some testing problems with countably many variables, as with the coefficients of curve representations, can be efficiently solved. Problem 11 finds proper parametric solutions only with distributions lying within the regular exponential family with only one nuisance parameter provided that the invariance property works. Within the permutation approach (Example 4.11), we will find proper workable solutions when the null hypothesis is an interval of equivalent points – like $H_0 : -\varepsilon_L \leq \delta \leq \varepsilon_U$ – and also when the null contains all non-equivalent points – like $H_0 : (\delta < -\varepsilon) \bigcup (\delta > \varepsilon)$, as is common in bioequivalence. In Problem 12, the number of parameters is often much larger than sample size: $n \ll D_{im}(\Theta)$, so no proper estimates are possible.

We substantially agree with the authoritative opinion expressed by R. A. Fisher (1935), the father of the permutation testing idea, who said '*…statisticians do not carry out*– by hand calculations – *this very simple and very tedious process, but their conclusions have no justification beyond the fact that they agree with those which could have arrived at by this elementary method …*' (Section 2.6). The idea of permutation test has been first introduced by J. Splawa-Neyman in a Polish paper (1923, translated into English by Dabrowska & Speed, Splawa-Neyman, 1990).

Except for essentially two very simple problems, both related to contingency tables with binary data – Fisher's exact probability test on two independent samples, and McNemar's test on one sample with paired observations – in practice, the general great computational complexity implied by permutation methods suggested him to base the statistical testing on the concepts of *likelihood* and of *hypothetical repeated sampling from the parent population* by examining the whole sample space of the experiment. So, one might argue that Fisher seems considering the role of likelihood-based tests as that of approximating in most circumstances their permutation distributions, especially in the presence of nuisance parameters, however, at the price of some lack of generality, if any. In the last decades, mostly due to the availability of relatively cheap and powerful computers and efficient software, permutation tests have increased in a number of applications and in facing and solving quite complex multidimensional problems, including some that are impossible by likelihood methods (Sen, 2007).

We partially agree with the opinion 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 emphasizes the importance of statisticians to refer to conditional procedures of inference and, in particular, to randomization methods. Indeed, there is a wide range of inferential problems that are correctly and effectively solved within a permutation/randomization frame, there are others that are difficult or even impossible to solve outside of it. We partially disagree because there are very important families of inferential problems, especially connected to unconditional parametric estimation and testing, or to nonparametric estimation and prediction, or more generally within the statistical decision approach, that cannot be dealt with and/or solved within a permutation approach. These are often connected to violations of the so-called exchangeability condition (Chapter 2; Remark 2.18). In addition, all procedures of exploratory data analysis and all testing methods for which we cannot assume exchangeability of the data in the point null hypothesis generally lie outside the permutation approach. Moreover, the traditional Bayesian inference (Remark 2.44, for suggestions on a *permutation Bayesian* approach) also lies outside the permutation frame.

Thus, we think that permutation methods should be in the toolkit of every statistician interested in applications, methodology and theory. We do not believe, however, that *all* inferential problems of interest for analyzing real problems fall within the permutation approach. Indeed, if conditions for their proper application were not satisfied, their use might become erroneous. Section 1.4 lists a set of circumstances in which permutation testing procedures may be effective or even unavoidable (Pesarin and Salmaso, 2010).

1.2 Basic Notation

Generally, theorems from the literature are reported without proof, whereas the most important properties of permutation tests, regarding their conditional and unconditional exactness, similarity, unbiasedness, consistency, power properties, etc., are explicitly established and proved in respective sections. The symbol ■ concludes a proof. Simple proofs of some specific properties as well as extensions of some results are often proposed to the reader as exercises. Several exercises and problems are proposed at the end of some subsections.

Onwards, unless it is necessary to make reference to specific countable sequences (as in Sections 2.5.3 and 4.5), we suppress the superscript $(^n)$ when referring to $\mathcal{X}^n$, $\mathbf{X}^{(n)}$, $\mathcal{X}^n_{/\mathbf{X}}$, $\Pi(\mathbf{X}^{(n)})$, etc. and shall simply write: $\mathcal{X}$, $\mathbf{X}$, $\mathcal{X}_{/\mathbf{X}}$, $\Pi(\mathbf{x})$, etc. Therefore, we do not distinguish among the dimensionalities of response variables, sample spaces of responses, etc., the context generally suffices avoiding misunderstandings.

The n-dimensional sample data $\mathbf{X}$ are assumed to be related to a response variable X. Response variables are usually indicated by italic capitals, such as X and Y, if they are univariate and by bold capitals, such as $\mathbf{X}$ and $\mathbf{Y}$, if multivariate. Responses are assumed to be observed on statistical units – individuals or subjects. Sample units are generally elicited through – symbolic – experiments carried out on a given population P. Similar to multivariate responses, sample data are indicated by bold capitals: $\mathbf{X}$, $\mathbf{Y}$, etc. The context is generally sufficient to avoid ambiguities. Lower-case letters are generally used to indicate integer numbers, real variables or constants: i, j, h, k,n, r, t, $\mathbf{x}$, z, etc. The most important exceptions are: A, which represents an event or an experimental factor; B, an experimental factor; RR, the number of re-randomization runs; R, the number of runs of a Conditional Monte Carlo (CMC) procedure; MC, the number of ordinary Monte Carlo runs in simulation studies; N_i, the cumulative frequencies in contingency tables.

As a rule, we do not use the notational conventions of linear algebra, so we do not distinguish between column and row vectors, etc. This is because, on the one hand, the context is always sufficiently clear; on the other, it is impractical in linear algebra notation to represent responses that are partly quantitative and partly categorical on which arithmetic operators do not work. The only adopted notational convention is that, when necessary, the transpose of a matrix $\mathbf{B}$ is denoted by $\mathbf{B}^\intercal$. Sometimes, $|\mathbf{X}|$ is used to denote the vector of absolute values as in $|\mathbf{X}| = \{|X_i|, i = 1, \dots, n\}$.

To indicate sample data, we often need to use the so-called *unit-by-unit representation*. For instance, in a two-sample univariate layout sized n_1 and n_2, with $n = n_1 + n_2$, we denote the whole pooled data by $\mathbf{X} = \{X(i), i = 1, \dots, n; n_1, n_2\}$, where $X(i) = X_i$ is the response related to the ith unit. This notation means that the first n_1 elements in the list belong to the first, the other n_2 to the second

sample. This representation is useful to express permutations for both categorical and quantitative responses, especially in multivariate contexts. The symbol $\biguplus$ is used for pooling – concatenating – two files of data into the pooled file, e.g. as $\mathbf{X} = \mathbf{X}_1 \biguplus \mathbf{X}_2$.

In general, we implicitly refer to the population and to the experiment of interest by means of a statistical model such as $(X, \mathcal{X}, \mathcal{A}, P \in \mathcal{P})$, where X represents the response variable, $\mathcal{X}$ the related sample space, $\mathcal{A}$ a suitable collection – an σ-algebra – of subsets of $\mathcal{X}$, and P the underlying parent distribution belonging to family $\mathcal{P}$. Thus, we assume that response variables take their values on the measurable space $(\mathcal{X}, \mathcal{A})$. Unless necessary, we do not distinguish between random variables and their explicit observations in sampling experiments.

We often refer to a sub-space such as $\mathcal{X}_{/A}$, the orbit or coset associated with A, i.e. the set of points $\mathbf{X} \in \mathcal{X}$ sharing condition A, where A is some event belonging to $\mathcal{A}$. The main conditioning set referred to in the context of permutation testing is the pooled set of observed data $\mathbf{X}$. Thus, $\mathbf{X}^*$ represents one of its permutations and $\mathcal{X}_{/\mathbf{X}}$ represents the related *permutation sample space* – often, especially for simple settings, $\mathcal{X}_{/\mathbf{X}}$ coincides with the plain data permutations $\Pi(\mathbf{X})$. Moreover, we assume that all statistics of interest are measurable with respect to $(\mathcal{X}, \mathcal{A})$ and, of course, with respect to the conditional or restricted algebra $\mathcal{A}_{/\mathbf{X}} = \mathcal{A} \bigcap \mathcal{X}_{/\mathbf{X}}$, so that conditional probability distributions associated with any $P \in \mathcal{P}$ are well defined on the measurable permutation space $(\mathcal{X}_{/\mathbf{X}},\ \mathcal{A}_{/\mathbf{X}})$ induced by conditioning on $\mathbf{X}$. Note that the measurability assumption with respect to the conditional algebra $\mathcal{A}_{/\mathbf{X}}$ of test statistics T is generally self-evident because all statistics of interest are transformations of the data $\mathbf{X}$ that, by assumption, are required to induce a probability distribution over $(\mathcal{X}_{/\mathbf{X}},\ \mathcal{A}_{/\mathbf{X}})$. Thus, any associated conditional inference has a clear interpretation. In general, unless necessary, we do not indicate the dimensionalities of variables and the cardinalities of sets and spaces, since these are clear from the context. The conditional expectation of X given A is denoted by $\mathbb{E}_A[X]$ or $\mathbb{E}[X|A]$.

We sometimes need partitioning the permutation sample spaces $\mathcal{X}_{/\mathbf{X}}$ into sub-orbits (e.g. Section 7.9.1 and Remarks 2.5 and 2.6) in order to take into consideration restrictions of invariance properties induced by post-stratification arrangements or by some statistics of interest when related to specific problems, so that solutions may become easier to construct.

A test statistic, $T : \mathcal{X} \to \mathcal{T} \in \mathcal{R}^1$, is a measurable real function taking values on a suitable space $\mathcal{T} = T(\mathcal{X}) \subseteq \mathcal{R}^1$ and is usually represented by symbols like $T = T(\mathbf{X})$ or $T^* = T(\mathbf{X}^*)$, etc., where the latter emphasizes the role of $\mathbf{X}^* \in \mathcal{X}_{/\mathbf{X}}$ as a permutation of the data $\mathbf{X}$. Conditional statements like $\Pr\{T^* \geq T^o|\mathbf{X}\}$ and $\Pr\{T^* \geq T^o|\mathcal{X}_{/\mathbf{X}}\}$ having essentially the same meaning may be used indifferently, though the latter is generally preferred. The set $\mathcal{T}$ is also named the support of T, and so the set $\mathcal{T}(\mathbf{X}) = \{T(\mathbf{X}^*),\ \mathbf{X}^* \in \mathcal{X}_{/\mathbf{X}}\}$ indicates the

permutation – conditional – support of T associated with the data $\mathbf{X}$. In order to emphasize the role of actual data, we use the notation $T^o = T(\mathbf{X})$ to indicate the *observed value*of T on the data $\mathbf{X}$. In general, the superscript *, as in $\mathbf{X}^*$, F^*, T^*, is used to indicate a variable, a distribution or a statistic related to permutation entities. We also use a hat, as in $\hat{F}$, $\hat{\lambda}$, $\hat{\sigma}$, $\hat{\lambda}$, to indicate an estimate, when referring to sample or Monte Carlo estimates.

We use symbols like $X(\Delta) = X + \Delta$ to denote responses when a treatment effect Δ – or δ – is to be emphasized; so, $X(0)$ implies $\Delta \stackrel{d}{=} 0$. For most problems of hypotheses testing, the observed data vector $\mathbf{x} = \{x_1, \ldots, x_n\}$ is usually obtained by a *symbolic* experiment performed n times on a population variable X taking values in sample space $\mathcal{X}$ with distribution P. We sometimes add the adjective 'symbolic' to names such as experiments, treatments and treatment effects, in order to refer to experimental, pseudo-experimental and observational contexts. For purposes of analysis, the data $\mathbf{x}$ are generally partitioned into *groups* or *samples*, according to the so-called *treatment levels* of the symbolic experiment. Up to Section 1.6, we use capital letters for random variables and lower case for observed data. From Section 1.7 onwards, we shall dispense with this distinction, so only capital letters will be used because the context is always sufficiently clear. Of course, when a dataset is observed at its $\mathbf{x}$ value, it is presumed that a sampling experiment on a given population P has been performed, so that the resulting sampling distribution is related to that of P. This is, of course, common to any statistical problem, not peculiar to the permutation approach.

1.3 The Permutation – Conditional – Testing Principle

In typical testing problems, we denote the null hypothesis by H_0, which assumes that data are from *one* unknown distribution P, $H_0 : X \sim P$ say, with P belonging to family $\mathcal{P}$. The observed data $\mathbf{x}$ are considered to be a random sample, taking values on sample space $\mathcal{X}^n$, so it is one observation from the n-dimensional variable $\mathbf{X}^{(n)}$, where this random sample does not necessarily have independent and identically distributed (i.i.d.) components (Chapters 2–4 for more details).

The observed data $\mathbf{x}$ are always an n-dimensional set of sufficient statistics for whatever parent distribution $P \in \mathcal{P}$. To see this statement in a simple way, let us assume that all members of $\mathcal{P}$ are dominated by a common *dominating* measure ξ and let f_P denote the density of P with respect to ξ, and $f_P^{(n)}(\mathbf{x})$ the density of $\mathbf{X}^{(n)}$. A practical implication of the common dominating measure requirement is that the data must be observed by using the same observational protocol criteria (see Remark 1.1). Since 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 $f_P^{(n)}(\cdot) = 0$, due to the well-known factorization theorem, any sample point $\mathbf{x}$ is therefore a sufficient set of statistics for whatever $P \in \mathcal{P}$.

Sometimes $\mathbf{x}$ is also *minimal sufficient* (Lehmann and Romano, 2005). In this vein, we know that if in the V-dimensional case, it is $n < V$ as is common with high-dimensional data, then $\mathbf{x}$ is always minimal sufficient. Also we know that when the n-dimensional $\mathbf{x}$ is minimal sufficient and there are nuisance parameters to remove from analysis, as is usual in most practical contexts, no uniformly best tests exist. If $\mathcal{P}$ is the set of all continuous distributions, vector $\mathbf{x}$ is *complete minimal sufficient*.

It is worth noting that $\mathbf{x}$ is a set of sufficient statistics under both H_0 and H_1. This fact implies that the statistician has to find suitable test statistics the conditional distributions of which are sensitive to both hypotheses by possibly providing stochastic dominance of the distribution under H_0 over that under H_1. This task is usually related to the fact that only under – point or sharp – H_0 data are *likelihood invariant*, i.e. exchangeable among samples; whereas, under H_1, the exchangeability property fails; hence, distributions of sample statistics are dependent on treatment effects, i.e. are no longer distributional invariant. Based on this notion, it is possible to characterize test statistics useful for correct inferences (Chapter 2).

Remark 1.1 When data are observed by means of different non-commensurable protocol criteria, as for instance data in centre 1 by protocol 1, those in centre 2 by protocol 2, and so on, – e.g. COVID-19 were observed in the 27 European countries (EU) with about the same protocol, those in Asian countries by quite different protocols, etc. – inferential comparisons on most aspects of interest, for instance by Multivariate ANalysis of VAriance (MANOVA) layouts, are possible with data within each protocol but are too hard or even impossible to justify in case of different protocols unless related measurements are commensurable. In non-commensurable cases, it seems that only descriptive analyses are possible.

1.3.1 Nonparametric Family of Distributions

Let us consider the following.

Definition 1.1 *A family $\mathcal{P}$ of distributions is called nonparametric when it is impossible to find a finite-dimensional space Θ, i.e. $\mathcal{D}_{im}(\Theta) < \infty$, such that there is a one-to-one relationship between Θ and $\mathcal{P}$, so that each member P of $\mathcal{P}$ cannot be identified by only one member θ of Θ, and vice versa.*

When such a one-to-one relationship exists, θ is called a parameter, Θ the parameter space, and $\mathcal{P}$ the corresponding parametric family. Families of distributions that are either unspecified or specified except for an infinite number of parameters satisfy the definition and so they are nonparametric. Def 1 also includes all those

situations where sample size n is smaller than the number of parameters, even when this is finite. All nonparametric families $\mathcal{P}$ that are of interest for permutation analysis are assumed to be sufficiently *rich* so that if x and x' are any two points of $\mathcal{X}$, then $x \neq x'$ implies $f_P(x) \neq f_P(x')$ for at least one $P \in \mathcal{P}$ in a set of points with positive P-probability. Thus, the qualification of $\mathcal{P}$ as nonparametric essentially depends on the knowledge we assume about it.

Permutation tests essentially are conditional procedures of inference, where the conditioning is on a set of sufficient statistics for P. So, consequent inferences at least concern the actually observed data $\mathbf{x}$ and their associated units (Section 2.9). The act of conditioning on a set of sufficient statistics engenders permutation tests independent of the underlying likelihood model related to P (Property 2.1). As a consequence, P can be unknown or unspecified; actually, only its existence is assumed. We specify this notion in the permutation testing principle.

1.3.2 The Permutation – Conditional – Testing Principle

Definition 1.2 *If two experiments, taking values on the same sample space $\mathcal{X}$ with distributions P_1 and P_2, both members of $\mathcal{P}$, give the same data $\mathbf{x}$, then the two inferences conditional on $\mathbf{x}$ and obtained using the same test statistic must be the same, provided that the data exchangeability is satisfied under H_0. 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 might differ.*

One of the most important features of the permutation testing principle is that, in theory and under a set of mild conditions, related conditional inferences can be unconditionally generalized – extended, extrapolated – to all distributions $P \in \mathcal{P}$ for which the density is positive, i.e. $\mathrm{d}P(\mathbf{x})/\mathrm{d}\xi^n > 0$ (Sections 2.1.1 and 2.9). It should be emphasized, however, that this feature derives from the sufficiency and conditionality principles of inference (Cox and Hinkley, 1974; Berger and Wolpert, 1988; Lehmann and Romano, 2005; Pesarin, 2015). Actually, resulting inferences are proper to all populations sharing the same conditioning statistics, particularly those that are sufficient for the underlying nuisance entities. For instance, Student's t test extends inference, more than to only one, to all normal populations that assign positive density to the variance estimate $\hat{\sigma}^2$ and so its inference is for a whole (sub)family of distributions. Therefore, such unconditional extensions should be carried out with care (Section 2.9). Another important feature occurs in multivariate problems, if solved through the NPC method. For these problems, especially when are complex and in very mild and easy-to-check conditions (Property 2.4 and Section 4.2.1), it is not necessary to specify and to model the structure of dependence relations for the concerned population variables. So, the analysis becomes feasible and results are easy to

interpret. For example, it is known that for multivariate categorical variables, it could be extremely difficult to properly model dependence relations among variables (Joe, 1997). Therefore, except for very particular cases, only some univariate problems – separately by each marginal component – are considered in the literature (Agresti, 2002). 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 – separately – unbiased and at least one of them is consistent* (Section 4.2.2). Also of interest is the application of this principle in the context of a Bayesian permutation approach (Remarks 2.43 and 2.44).

However, both the data exchangeability under H_0 and the conditioning on a set of sufficient statistics provide permutation tests with nice and important general properties useful for correct inferences (Pitman, 1937b; Box and Andersen, 1955; Watson, 1957). One is that they are always exact procedures (Property 2.2, Remark 2.18). One more is that their conditional rejection regions are *uniformly similar* to the unconditional counterparts (Scheffé, 1943a,b). This means that the null rejection rate, at least in principle, can be exactly determined. Moreover, if data comes from continuous distributions – where the probability of ties is zero, the resulting null rejection probability is invariant on the observed data $\mathbf{x}$, for almost all $\mathbf{x} \in \mathcal{X}^n$ and all distributions $P \in \mathcal{P}$ (Property 2.5). When data come from non-continuous distributions, unless referring to randomized tests, the exact similarity property is asymptotically attained (Property 2.6). Further, if the stochastic dominance condition with respect to δ is satisfied under H_1, permutation tests based on divergence of suitable statistics are *uniformly unbiased*, since the rejection rate of any test T, for all sample data $\mathbf{x} \in \mathcal{X}^n$, satisfies the relation $\Pr\{\lambda_T(\mathbf{x}(\delta)) \leq \alpha | \mathcal{X}_{/\mathbf{x}}\} = W(\delta, \alpha, T | \mathcal{X}_{/\mathbf{x}}) \geq \alpha$, where $\lambda_T(\mathbf{x}(\delta))$ indicates the T-related p-value-like statistic and $W(\delta, \alpha, T | \mathcal{X}_{/\mathbf{x}})$ the *conditional power* of T given $\mathbf{x}$, at effect δ and significance level α (Sections 2.2.4 and 2.5.2). One more generalizes the latter to the *uniform stochastic monotonicity* of conditional rejection probability with respect to δ (Corollary 2.1), that is:

$$W(\delta', \alpha, T | \mathcal{X}_{/\mathbf{x}}) \leq W(0, \alpha, T | \mathcal{X}_{/\mathbf{x}}) = \alpha \leq W(\delta, \alpha, T | \mathcal{X}_{/\mathbf{x}}), \forall \delta' < 0 < \delta.$$

It is noticeable that when the exchangeability can be assumed under H_0, the similarity and unbiasedness properties allow us to *weakly generalize– extend, extrapolate –conditional to unconditional inferences*, irrespective of the underlying population, distribution P and the way sampling data are obtained from $\mathcal{X}$ (Remark 2.31). Therefore, this weak generalization may be made with any sample data, even if they are obtained by selection-bias, not well-designed, procedures from the target population. Conversely, parametric solutions permit proper correct generalizations only in some cases when data are obtained by means of well-designed sampling procedures on well-specified parent populations P.

Specifically, a general situation for unconditional extensions in parametric contexts occurs when likelihood functions are known except for nuisance parameters, provided that these are removed by invariant statistics.

For these reasons, permutation inferences are proper with most observational data, sometimes called non-experimental, with experimental data, with selection-biased as well as with well-defined sampling procedures. However, we must note that well-defined sampling procedures are quite rare even in most experimental problems (Ludbrook and Dudley, 1998). For instance, if one wants to investigate the effects of a drug on caws, the units to be treated are usually not randomly selected from the target population of all caws to which that inference is designed but are selected in some way among those available in a few, compliant, farms and are *randomly assigned* to the treatments. 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 designed treatments.

In one sense, the concept of random sampling is rarely attained in real applications because, for various reasons, the large majority, if not all, real samples are obtained by selection-bias procedures. This implies that most unconditional inferences usually associated with parametric tests, being based on the concept of random sampling from the target, are rarely applicable. From a general point of view, the inferential generalization to the target population may be seen as an important characterization of permutation, conditional, testing approach. In addition, due to similarity and uniform unbiasedness, permutation solutions allow for relaxation of most of the assumptions needed by parametric counterparts, such as existence of variances and homoscedasticity of responses under the alternative (Sections 1.4 and 2.9, Remark 2.18).

A further argument relates to the fact that as it takes place in many experimental problems, the assumption that a treatment does not influence scale coefficients or other distributional aspects is often too unrealistic. Indeed, in accordance with a kind of generalized Behrens–Fisher model (Section 2.4, Corollary 2.2; Example 4.13; Pesarin, 2001, Chapter 10, for details), the assumption that a treatment only affects the parameter of interest, leaving unaffected all the nuisance, can rarely be justified in practice. Thus, on the one hand, traditional parametric solutions may become improper; on the other, permutation inferences are much important for both theoretical and applicational purposes, not only for their potential exactness. Many authors have emphasized these features. A review of some related arguments is in: Edgington and Onghena (2007), Good (2000, 2005), Mielke and Berry (2007), Pesarin and Salmaso (2010a), and Berry et al. (2014).

When the exchangeability can be assumed under H_0, reference null distributions of permutation tests always exist because, at least in principle, they are obtained by examining all permutations of available data (Chapter 2). In addition, permutation comparisons of means or of other indicators do not require

homoscedasticity under the alternative, when treatment may also modify dispersion or other nuisance entities, provided that there is dominance of distributions, i.e. the underlying CDFs do not cross each other (Section 2.1.1).

It is also to be emphasized, in accordance with the notion of *pure significance tests* (Cox and Hinkley, 1974), that the treatment effects for which nonparametric testing inferences are of interest, more than on parameters, are properly concerning the so-called *population functionals or pseudo-parameters*, such as characteristic functions, cumulants, frequencies, interpoint distances (Bartoszynski et al., 1997), pair-wise (Buyse, 2010) and simplicial (Gillard et al., 2022b), means, moments and moment-generating functions. Specifically, the most common two-sample test is based, with clear meaning of the symbols, on a comparison of two sample averages, $\overline{X}_1 - \overline{X}_2$, that are linear estimates of two population means: $\mu_j = \int_{\chi} x \, \mathrm{d}P_j(x;\boldsymbol{\theta})$, $j = 1, 2$, where $\boldsymbol{\theta}$ is the set of parameters and the μ_j are the concerned functionals. In general, a φ-functional for $X \sim P$ is just the mean value of $\varphi(X)$ defined as $\varphi(X) = \int_{\chi} \varphi(x) \mathrm{d}P(x;\boldsymbol{\theta})$. If φ is monotonic, it is usual, although not necessary, to express these functionals in the same unit of measurements of X, i.e. as $X_{\varphi} = \varphi^{-1}[\varphi(X)]$. Also of interest in many applications are the so-called π-quantile functionals defined as $x_{\pi} : \{\inf {}_x \int_{-\infty}^{x} \mathrm{d}P(t;\boldsymbol{\theta}) \geq \pi + \sup {}_x \int_{-\infty}^{x} \mathrm{d}P(t;\boldsymbol{\theta}) \leq \pi\}/2$. Of course, these φ-means and π-quantiles are functions of all known and unknown parameters $\boldsymbol{\theta}$ characterizing P_j. For instance, if the population distribution of a variable X is a mixture of two normal variates with parameters $a \in [0, 1]$, $\mu_j \in \mathcal{R}^1$, and $\sigma_j > 0$, $j = 1, 2$, i.e. $X \sim a \cdot \mathcal{N}(\mu_1, \sigma_1) + (1 - a) \cdot \mathcal{N}(\mu_2, \sigma_2)$, its mean value $\mu = a\mu_1 + (1 - a)\mu_2$ is a function of three parameters, to any point of which there are infinitely many vectors $\boldsymbol{\theta} = (a, \mu_1, \mu_2; \sigma_1, \sigma_2)$ sharing the same μ: formally, $\boldsymbol{\theta}_{\mu} \in \Theta_{\mu} \subset \Theta$. This emphasizes a very important difference between parametric and nonparametric tests. Of course, a parameter can be a functional for P, like μ in the ordinary normal case, but this is not true in general. Moreover, to use a parametric test based on comparison of sample functionals, it is necessary to assume that only one component θ of $\boldsymbol{\theta}$ is affected by the treatment, whereas all others, the nuisance, are left unaffected. One implication of this notion is that the parametric and the nonparametric testing methodologies are not always reasonably comparable. In practice, the use in an analysis of a specific functional is justified if it has an appropriate physical meaning.

1.4 Permutation Approaches

In the literature, three leading approaches for constructing permutation, conditional, tests are considered: one is essentially heuristic and two others are more formal. The heuristic, typically based on intuitive reasoning, is the most commonly

adopted for simple problems, especially when common sense may suffice; it is essentially operational – how to compute it, but without providing rational motivations for – (among the many: Pitman, 1937a,b, 1938; Lunneborg, 1999; Edgington and Onghena, 2007). But when problems are not simple, this approach may become inadequate. The two formal approaches are much more elegant, effective, precise and, for most unidimensional simple problems, essentially equivalent. One is based on the concept of group-invariance of the null distribution under the action of a finite group of transformations, essentially an *algebraic* approach. Similar to the heuristic, this approach is guided by the task of finding exact solutions under point, sharp, null hypotheses; very little is said and known about its distributional behaviour under the alternative (Scheffé, 1943a,b; Lehmann and Stein, 1949; Hoeffding, 1952a; Romano, 1990; Runger and Eaton, 1992; Nogales et al., 2000; Langsrud, 2005; Hemerik and Goeman, 2021; Dobriban, 2022). The other, which is as elegant and precise as the group-invariance, is based on the concept of conditioning on a set of sufficient statistics for the underlying distribution P (Fisher, 1925; Watson, 1957; Lehmann, 1986; Pesarin, 2001, 2015; Pesarin and Salmaso, 2010a), which also defines a particular case of group-invariance (Section 2.1.2).

At least in one-dimensional cases, since for null hypotheses they often, but not always, provide the same solutions, two formal approaches appear as equivalent (Watson, 1957; Odén and Wedel, 1975; Nogales et al., 2000). However, we prefer the conditional approach since it is based on a clear theory of conditional inference and so it is easier to understand and is more constructive and more natural to use; moreover, test statistics are generally simpler to justify; main testing properties are easier to prove; inferential conclusions are easier to interpret and explain; its power characterisation under the alternative is easier to establish; its extension to composite null and composite alternatives is straightforward (Sections 2.2 and 2.5; Remark 2.18). In few words, even if it appears somewhat less general when the study is restricted to point null hypotheses, generally it looks more realistic and more inclusive. Further, the extension to multidimensional problems of the group-invariance approach presents some difficulties, especially when some null and/or alternatives are composite and/or are subject to geometrical restrictions and/or variables are mixed – partly numerical and partly categorical. In addition, the group-invariance property seems to be assumed without deriving it from clearly stated conditions related to the problems under study and so it appears as a superimposition, so that the interpretation of the related inferences is not completely clear. Instead, the conditioning on a set of sufficient statistics deductively provides permutation procedures with the group-invariance property (Properties 2.1 and 2.2, Section 2.1.3), which then appears as one of its cases.

Willing to emphasize the role of conditioning, in place of permutation testing, a more comprehensive name would be *nonparametric conditional testing*. A notion

that looks slightly more general than the strictly intended permutation testing. Actually, it also includes approaches where, due to specific side assumptions allowing some symmetries on the component variables or on the experimental design leading to restraints on the conditional orbits, the related analyses can be computationally simplified, like, e.g. with the so-called rotation tests (Langsrud, 2005). But, to stay in line with tradition, we still prefer using the name of *permutation testing* since it is the most commonly used both by researchers and practitioners.

In this book, we will see some of the only apparently different ways of considering permutations: a) one is related to paired and repeated measurement layouts, where the permutations are the product of within unit permutations (an example is in Section 1.9.1); b) one is related to two-sample and multi-sample layouts, where the permutations are obtained by permuting data among samples, providing the explicit justification of the name of permutation testing (examples are in Sections 1.10.1 and 1.11.1); c) another one considers the so-called synchronized permutations useful for separately testing main effects and interactions on some factorial layouts (Example 3.17); d) one more is related to stratification and post-stratification layouts where the permutation sample space becomes the Cartesian product of concerned partial spaces (Remark 2.6); e) a further one occurs with some cross-over designs for separately testing on main treatment effect and on interaction (Remark 2.7); etc.

For very simple problems, we often adopt the heuristic approach, especially if the related solutions are clear and no ambiguities arise. We use it in the present chapter. From Chapter 2 onwards, the conditional approach is preferred to the group-invariance. The reason for this preference is that conditioning appears to be a more constructive way, easier to work with and easier to understand. Additionally, it is much easier to establish required properties of statistical tests, in particular with regard to their behaviour under the alternative (Sections 2.2.4, 2.6, 2.8). Besides, the group-invariance is quite more difficult to apply properly in multivariate situations, especially when that invariance notion is not exactly uniform on all component variables as when some are numeric, others nominal or ordered categorical (Chapter 4 and beyond).

1.5 When and Why Conditioning Is Appropriate

We know that parametric testing methods may be available, appropriate and effective when:

1) sample data are obtained by well-defined random experiment on well-specified parent populations;

2) population distributions for responses, i.e. the likelihood models, are well defined and treatment effect works only on the parameter of interest, typically the location vector in multivariate cases, while all others nuisance entities are not affected;
3) with respect to all nuisance entities, well-defined likelihood models are provided with either boundedly complete estimates under H_0 or at least with invariant statistics;
4) at least asymptotically, null sampling distributions of test statistics do not depend on any unknown entity.

Therefore, as there are circumstances in which parametric tests may be appropriate from the point of view of interpretation of related inferential results or for their efficiency, in other circumstances, they may be inappropriate or even impossible to use correctly. Conversely, there are circumstances in which conditional testing procedures may become appropriate and at times unavoidable. A brief, incomplete, list of such circumstances is:

- Distributional models for responses are nonparametric.
- Distributional models are not well-specified.
- Distributional models, although well-specified, depend on too many nuisance entities.
- Treatment effects, even on well-specified models, act on more than one parameter or on other aspects of a distribution, as when heteroscedasticities occur only in the alternative.
- With respect to some nuisance entities, well-specified distributional models do not possess invariant statistics or boundedly complete estimates under H_0.
- Treatment effects are presumed to act possibly on more than one aspect of interest for analysis, leading to multi-aspect testing methods (Example 4.8).
- Data are randomly obtained from finite populations, so independent observations cannot be assumed, where only data exchangeability can take place (Remark 2.4, Point **i**).
- Ancillary statistics in well-specified parametric models have a strong influence on inferential results, leading to post-stratification analyses.
- Ancillary statistics in well-specified models are confounded with other nuisance entities.
- Asymptotic null sampling distributions depend on unknown entities.
- Problems in which the number of nuisance entities is larger than, or may increase with, sample sizes.
- The number of response variables to be analyzed is larger than sample sizes, even infinitely larger.
- Problems in which the observational precision depends on the value to be observed.

- In multivariate problems, some variables are categorical, nominal and/or ordinal, and others quantitative, including unobserved, non-detected and zero-inflated data.
- Multivariate alternatives are subject to order restrictions.
- In multivariate problems and in view of particular inferences, there are variables in the analysis, which have different degrees of importance (point (f) in Section 4.2.4).
- Data contain non-ignorable/informative missing values (Sections 7.6–7.14).
- Data are obtained by ill-specified selection-bias procedures (Section 2.9).
- Treatment effects may depend on unknown entities, e.g. when are confounded with some nuisance parameters (Example 4.13)
- The null hypothesis is an interval or a union of half lines, as with the so-called *testing for equivalence* (Example 4.11).
- Testing problems with multivariate nominal categorical variables (Sections 6.4–6.10).
- Testing problems for repeated measurements with univariate or multivariate ordered categorical data (Section 6.3).

In addition, we may decide to adopt conditional testing inferences not only when parametric counterparts are not possible but also when we want to lay more importance on the actually observed units and related data $\mathbf{x}$, than on the population P from which data are obtained.

Conditional inferences are also of interest when, for whatever reason, we wish to limit ourselves to conditional methods by explicitly restricting to the actual data $\mathbf{x}$ Section 2.9 (Greenberg, 1951; Kempthorne, 1966, 1977, 1979, 1982; Basu, 1978, 1980; Thornett, 1982; Greenland, 1991; Celant and Pesarin, 2000, 2001; Pesarin, 2001, 2002, 2015; Lehmann, 2009). For example, this situation agrees with the idea that, when assessing the reliability of cars, the owner may be mostly interested in his own car, or fleet of cars if he has more than one because he is responsible for all related reliability maintenance costs, thus giving rise to a *conditional assessment*. Of course, the point of view of the car manufacturer, whose reputation and warranty costs are related to the whole population of similar cars, may be mostly centred on a sort of *average behaviour*, giving rise to a form of *unconditional assessment*.

Thus, both conditional and unconditional points of view are important and useful in real problems because there are situations, such as that of the owner, in which we may be interested in conditional inferences, and others, such as that of the manufacturer, in which we may be interested in unconditional inferences; hence, both are of interest. However, within conditional testing approaches, provided that exchangeability of data is satisfied under H_0, permutation methods play a central role (Section 2.9).

1.6 Randomization and Permutation

In most experiments, units are randomly assigned to treatments, so under H_0, the observed data appear as if they were *randomly assigned* to samples. Based on this notion, several authors prefer the term *randomization tests* (Pitman, 1937a,b; Kempthorne, 1977; Good, 2005; Edgington and Onghena, 2007) or even *re-randomization tests* (Gabriel and Hall, 1983; Lunneborg, 1999; Rosenberger and Lachin, 2016) in place of *permutation tests* In Section 1.4 we have shown that a more inclusive term would be *nonparametric conditional tests*. In Section 2.7, related to the so-called conditional and post-hoc empirical, power functions, a more precise notion of re-randomization will be used. Although of no great importance, being a mere question of words, we prefer the term 'permutation' because at least for two- and multi-sample layouts it is closer to the true state of things and because to some extent, it has a wider meaning than the others. Indeed, a sufficient condition for properly applying permutation tests is that H_0 implies the exchangeability of observed data. In Remark 2.18, we will see an important extension of this assumption leading to testing for composite hypotheses, a situation in which the notion of group-invariance of the null distribution under the action of a finite group of transformations at least becomes difficult to apply, especially if considered as unrelated to the conditioning on a set of sufficient statistics. For instance, in a symbolic experiment where a variable is observed on male and female groups of some animals, the notion of randomization is difficult to apply exactly because in no way gender can be randomly assigned to units. Instead, the permutation idea is rather more natural because in the null hypothesis of no distributional difference due to gender we are led to assume that observed data may be indifferently assigned to either males or females, a notion that justifies exchangeability of data and the permutation of unit-labels, but not the randomization of units.

The greater emphasis on the notion of randomization through random assignment of units to treatments resides in that it is generally easier and rather natural to justify the assumption of exchangeability under H_0 for experimental than for observational data. However, when the exchangeability property cannot be assumed under H_0, where even the group-invariance cannot be applied, both parametric and permutation inferences are generally not exact (Example 4.13; for hints on some approximated permutation solutions see Brunner et al., 2019).

In these cases, especially when even approximate solutions are difficult to obtain, it might be wise considering bootstrap techniques, which are less demanding in terms of assumptions and might be effective for exploratory or for asymptotic purposes, in spite of the fact that related inferences for finite sample sizes are neither conditional nor unconditional (Remarks 2.51, 2.52).

The conditional property of permutation tests leads to two rather different concepts of inferences and power functions. One concern is the actually observed

data $\mathbf{x}$ and is related to the *conditional inference* and the *conditional post-hoc*, *power function*, respectively; the other concerns are the parent population P and are the *unconditional inference* and the *unconditional power function*. Their definition and determination for some simple problems are discussed in Section 2.7, an algorithm for evaluating the post-hoc power function is presented in Section 2.7.1. In any case, it should be stressed that, except for some cases of asymptotic or approximate calculations, both power functions cannot generally be expressed in closed forms. Of course, being nonparametric, the post-hoc conditional power does not require knowledge of the population distribution P and is the most important for conditional inferences. Instead, the unconditional power implies knowledge of P, including all its parameters. One important feature of both conditional and unconditional powers of permutation tests based on divergence of suitable statistics is that they are monotonically related to the non-centrality functional, i.e. the treatment effect δ, sometimes also called the *effect-size*, and that this is true independently of the underlying population distribution P (Section 2.7).

1.7 Computational Aspects

One of the major problems associated with permutation tests is that their null distributions, except for very particular cases, Fisher's exact probability test and McNemar's test, are generally impossible to express in closed forms. In fact, they depend on specific data $\mathbf{x}$, and thus they vary as data vary in the sample space $\mathcal{X}^n$. In addition, when sample sizes are not small, direct computations are practically impossible because of the very large cardinality of associated permutation sample spaces, here denoted by $\mathcal{X}^n_{/\mathbf{x}}$ or, occasionally, by $\Pi(\mathbf{x})$. Besides, the approximation of such distributions by means of asymptotic arguments is not always appropriate, unless sample sizes are very large, because their dependence on actual data $\mathbf{x}$ makes it difficult to express and to check the conditions needed to evaluate the rate of approximation in practice (Sections 2.11 and 2.12).

Some algorithms, not based on the complete examination of the whole permutation space $\mathcal{X}^n_{/\mathbf{x}}$, have been developed for univariate situations allowing for exact computation of the permutation distribution in polynomial time (Pagano and Tritchler, 1983; Zimmerman, 1985a,b; Mehta and Patel, 1980, 1983, 1999; Barabesi, 1998, 2000, 2001; Hemerik and Goeman, 2018; Segal et al., 2018; Koning and Hemerik, 2022). It is also worth noting that there are computer packages, a well-known is StatXact® – which provide exact computations on many univariate problems. Approximate computations in the univariate context are provided, for instance, by Berry and Mielke (1985) and Mielke and Berry (2007), where a suitable parametric distribution sharing the same few moments of the exact one is considered. Such computing methods for approximating

permutation distributions in some specific cases of multivariate hypotheses are straightforward (Section 4.1.1).

For practical reasons, especially for multivariate purposes, in order to obtain appropriate and reliable evaluations of the involved distributions, in this book, we suggest using *Conditional Monte Carlo* (CMC) procedures on permutation spaces $\mathcal{X}^n_{/\mathbf{x}}$ or $\Pi(\mathbf{x})$. Although in principle it is always possible to carry out exact calculations by means of specific computing routines based on complete examination, in practice the use of CMC algorithms is required by the too large cardinality of permutation spaces, especially by the method of NPC of dependent permutation tests in multivariate situations. We underline that CMC methods are carried out by means of without-replacement resampling procedures on the data $\mathbf{x}$. It must be emphasized that these procedures are substantially different from the *bootstrap techniques* (Property 2.3; Remarks 2.51 and 2.52). Actually, to consider random permutations, in CMC resampling, replicates are done without replacement on data $\mathbf{x}$, considered as playing the role of a finite sub-population, provided that sample sizes are finite. Hence, they correspond to a random sampling from the space of data permutations. In this sense, to some extent, the name CMC has the meaning of *without-replacement resampling*. Of course, CMC procedures provide good and reliable *statistical estimates* of desired permutation distributions, the accuracy of which depends on the number R of runs.

The increasing availability of relatively inexpensive and fast computers and good software has made permutation tests more and more accessible. In fact, concerned distributions can be effectively approximated by statistical estimations, more than by numeric evaluations, without compromising their desirable statistical properties.

Most well-known statistical packages include specific routines for CMC simulation of permutation distributions; as for instance, MATLAB©, Python, R, SAS, SC©, S-PLUS©, SPSS©, Statistica©, StatXact®, etc. Some of these also include routines for exact computations.

1.8 A Problem with Paired Observations

As an initial example, let us consider a testing problem on the effectiveness of training in the reduction of anxiety in a sample of $n = 20$ subjects (Pesarin, 2001). At a first glance, subjects of the experiment are presumed to be 'homogeneous' with respect to the most important experimental conditions, the so-called covariates, such as sex, age and health status. In other terms, subjects are assumed to be *a sample from population P* (Remark 1.2).

Anxiety Y is assessed by an Institute for Personality and Ability Testing (IPAT) psychological test whose responses are quantitative scores corresponding to the

Table 1.1 IPAT data on anxiety in 20 individuals.

i	Y_1	Y_2	X	i	Y_1	Y_2	X
1	19	14	5	11	16	17	−1
2	22	23	−1	12	25	20	5
3	18	13	5	13	22	18	4
4	18	17	1	14	19	17	2
5	24	20	4	15	27	22	5
6	30	22	8	16	23	21	2
7	26	30	−4	17	24	21	3
8	28	21	7	18	18	15	3
9	15	11	4	19	28	24	4
10	30	29	1	20	27	22	5

sum of sub-responses to a set of items. Each unit is observed before treatment, occasion 1, also called the *baseline observation*, and one week after a fixed number of training sessions, occasion 2, that are administered with the aim to stochastically reducing baseline values.

Of course, within unit bivariate responses are dependent because they are measured on the same unit at different times, whereas the n pairs are assumed to be independent because they are related to different units. Moreover, due to the assumed homogeneity of units, the data pairs $\{(Y_{1i}, Y_{2i}), i = 1, \dots, n\}$ may be viewed as a random sample of n i.i.d. pairs from the bivariate variable $Y = (Y_1, Y_2)$. Formally, data are represented by a matrix of n pairs $(\mathbf{Y}_1, \mathbf{Y}_2) \in \mathcal{X}$, where $\mathcal{X}$ is the sample space of the experiment.

Observed data are listed in Table 1.1, where the fourth column contains individual differences $\{X_i = Y_{1i} - Y_{2i}, i = 1, \dots, 20\}$. The dataset is available at http://www.wiley.com/go/npc in `examples_chapters_1-4` folder.

1.8.1 Modelling Responses

The expected treatment effect is that training may generate a stochastic reduction of anxiety. Therefore, we can write the hypotheses as

$$H_0 : \{Y_1 \stackrel{d}{=} Y_2\} = \{P_1(t) = P_2(t), \forall t \in \mathcal{R}^1\}$$

against $H_1 : \{Y_1 \stackrel{d}{>} Y_2\}$, where P_1 and P_2 are the marginal distributions of Y_1 and Y_2.

Note that H_0 asserts the distributional – stochastic – equality of responses and that this is coherent with the hypothesis that training is completely ineffective. Moreover, it should be noted that the *stochastic dominance* of Y_1 with respect to Y_2, stated by H_1 and denoted by the symbol $\overset{d}{>}$, may be specified in several ways according to proper side-assumptions. Most common specifications of response models are:

- (M.i) With *fixed additive effects*. $Y_{1i} = \mu + Z_{1i}$, $Y_{2i} = \mu - \delta + Z_{2i}$, $i = 1, \dots, n$, where μ is a constant common to all units; δ is the fixed treatment effect, assumed to be finite and strictly positive in H_1; Z_{1i} and Z_{2i} are identically distributed *centred random deviates*, the so-called *error terms*, which are assumed to be not independent within but independent between units.
- (M.ii) With *fixed additive effects but with non-homogeneous units*. $Y_{1i} = \mu + \eta_i + Z_{1i}$, $Y_{2i} = \mu + \eta_i - \delta + Z_{2i}$, $i = 1, \dots, n$, where η_i are unknown components specific to the ith unit assumed to be not dependent on treatment levels; all other components have the same meaning as in (M.i).
- (M.iii) With *individually varying additive effects*, i.e. fixed effects specific to each unit. $Y_{1i} = \mu + \eta_i + \sigma_i Z_{1i}$, $Y_{2i} = \mu + \eta_i - \delta_i + \sigma_i Z_{2i}$, $i = 1, \dots, n$, where σ_i are the scale coefficients and δ_i the treatment effects both specific to the ith unit; under H_1, the δ_i are fixed non-negative finite quantities at least one of which is positive; it is worth noting that bivariate data are independent but not identically distributed.
- (M.iv) With *generalized stochastic effects* . $Y_{1i} = \mu + \eta_i + \sigma_i Z_{1i}$, $Y_{2i} = \mu + \eta_i + \sigma_i Z_{2i} - \Delta_{2i}$, $i = 1, \dots, n$, where under H_1, random effects Δ_{2i}, which may depend in some way on $(\mu, \eta_i, \sigma_i, Z_{1i}, Z_{2i})$, are non-negative stochastic quantities, at least one of which is strictly positive.

Model (M.i) is the standard model for homogeneous homoscedastic observations. Other models extend standard conditions. In particular, model (M.ii) assumes homoscedasticity of responses but non-homogeneous units. (M.iii) is consistent with situations in which relevant covariates are not observed, e.g. when some individuals are male and others female with possibly different associated effects even with different scale coefficients. (M.iv) is consistent with any form of stochastic dominance for quantitative responses, in particular with those with fixed or stochastic multiplicative forms. In this chapter, we mainly refer to the additive fixed model as in (M.i). We leave the extension of the main results to other models to Chapters 2 and 3 and to some suggested exercises.

The null hypothesis may also be written as $H_0 : \{\Pr(Y_1 - Y_2 \leq -t) = \Pr(Y_1 - Y_2 \geq t), \forall t \in \mathcal{R}^1\}$, where it is assumed that these probability statements are well-defined and are related to distribution P of (Y_1, Y_2). Thus, under H_0, the difference $X = Y_1 - Y_2 = \delta + Z_1 - Z_2$ is symmetrically distributed around 0,

whereas under H_1, where, in particular, we have $\Pr\{X > 0|H_1\} > 1/2$, X is symmetrically distributed around the location, i.e. the treatment effect $\delta > 0$. Of course, δ corresponds to a suitable indicator, i.e. a functional or pseudo-parameter, for the effect; usually it is taken as the mean, or the trimmed mean, or the median, etc.

Remark 1.2 When using differences X, models (M.i) and (M.ii) become equivalent. Indeed, both become $\{X_i = Y_{1i} - Y_{2i} = \delta + Z_{1i} - Z_{2i}, i = 1, \dots, n\}$. This means that when covariates are assumed to influence only individual specific components η_i, differences $\mathbf{X}$ become covariate free. A nice consequence of this is that when adopting model (M.ii), it is not required that units are homogeneous with respect to experimental conditions.

1.8.2 Symmetry Induced by Exchangeability

A formal proof of the symmetry property around 0 of $X = Z_1 - Z_2$, under H_0, may easily be achieved by observing that two error deviates Z_1 and Z_2, although often not independent, are exchangeable within units. Exchangeability within units implies both $F_1(t) = F_2(t), \forall t \in \mathcal{R}^1$ and

$$F_{1|t}(z|Z_2 = t) = F_{2|t}(z|Z_1 = t), \forall (t, z) \in \mathcal{R}^2,$$

where F_1, F_2, $F_{1|t}$ and $F_{2|t}$ represent, respectively, the CDFs of variables Z_1, Z_2, $(Z_1|Z_2 = t)$, and $(Z_2|Z_1 = t)$. Of course, all these CDFs are associated with P. Hence,

$$\Pr\{(Z_1 - Z_2) \le z\} = \int_{-\infty}^{+\infty} F_{1|t}(z + t|Z_2 = t) \cdot \mathrm{d}F_2(t)$$

and

$$\Pr\{(Z_2 - Z_1) \le z\} = \int_{-\infty}^{+\infty} F_{2|t}(z + t|Z_1 = t) \cdot \mathrm{d}F_1(t),$$

thus $\Pr\{X > z\} = \Pr\{X < -z\}, \forall z \in \mathcal{R}^1$, which is the condition for symmetry of X around 0.

One consequence of this property is that, in H_0, $\Pr\{X < 0\} = \Pr\{X > 0\}$. Thus, assuming $\mathbb{E}(Z)$ is finite, X has a null mean value; moreover, assuming the median $\mathbb{M}d(Z)$ has only one value, X has a null median. Instead, under H_1, $\Pr\{X < 0\} > (<) \Pr\{X > 0\}$, according to whether the responses are such that $Y_1 \overset{d}{<} (\overset{d}{>}) Y_2$. One more consequence is that, under H_0, the vectors of signs $(X_i/|X_i|, i = 1, \dots, n)$ and of differences $(X_i = Y_{1i} - Y_{2i},\ i = 1, \dots, n)$ are stochastically independent, when $X_i = 0$, the difference X_i and related sign $X_i/|X_i|$ are excluded from testing (Randles and Wolfe, 1979, p. 50, for a proof).

1.8.3 Further Aspects

The distribution P of X is assumed to be unknown in some of its parts, i.e. in some of its parameters, in its analytic form, or in P as a whole, provided that it belongs to a family of non-degenerate distributions $\mathcal{P}$ (Section 1.2). Moreover, assuming the mean value $\mathbb{E}(Z)$ is finite, so that we may consider the sample mean $\overline{X} = \sum_i X_i/n$ as a proper indicator of training effect δ, the hypotheses can equally be written as $H_0 : \{\delta = 0\}$ against $H_1 : \{\delta > 0\}$.

In this setting, the vector of pairs $\{(Y_{1i}, Y_{2i}), i = 1, \dots, n\}$ may be viewed as a random sample of n pairs, where exchangeability is intended within each individual pair (Remark 2.4).

Note that the existence of $\mathbb{E}(X)$ is a sufficient condition to get consistency (Section 2.5.3). If $\mathbb{E}(X)$ cannot be assumed finite, then we may refer to a suitable and possibly robust indicator of δ, such as the median $\mathbb{M}d(X)$ or the trimmed mean. Also note that if it were convenient for analysis, we might consider non-degenerate data transformations φ, such as $\varphi(Y_1) - \varphi(Y_2)$ or more generally $\varphi(Y_1, Y_2) = -\varphi(Y_2, Y_1)$, so that the sample mean of the transformed data is a proper indicator of effect. These transformations generally modify the distributions of concerned variables and may better fit one of the additive models (M.i) to (M.iv) in Section 1.8.1, so it is possible to obtain good power behaviour of the resulting test statistics and to improve the interpretation of the results (Remarks 1.4 and 2.36).

Remark 1.3 The one-sample-matched pairs problems, where independent units are paired according to some known covariates, are formally equivalent to that of paired observations. The only inessential difference is that error components Z_{1i} and Z_{2i}, instead of simply exchangeable, being related to independent units are now independent (Problem 11, Section 1.9.4).

Remark 1.4 The testing problem with paired observations may be solved in several parametric and nonparametric ways, according to explicit assumptions concerning the distribution P. Moreover, determining the best data transformation φ so as to obtain a best test of the form $\sum_i \varphi_i$ for finite sample sizes is still an open problem in the nonparametric context (Runger and Eaton, 1992); Section 2.6).

1.8.4 Student's *t*-Paired Solution

A first well-known parametric solution may be found if the response variable X were assumed to be normally distributed with unknown variance. Accordingly,

the response model with fixed additive effects can now be written as $\{Y_{1i} = \mu + \sigma \cdot Z_{1i}, Y_{2i} = \mu - \delta + \sigma \cdot Z_{2i}, i = 1, \dots, n\}$, where μ is a population constant; δ the treatment effect; $\sigma \in \mathcal{R}^+$ the unknown standard deviation assumed to be common to units and unaffected by treatment; random errors $Z_{ji} \sim \mathcal{N}(0,1)$, $j = 1, 2$, are assumed to be normally distributed, with null means and unit variances, independent among units but not necessarily independent within units.

In this setting, the alternative being one-sided, an optimal solution – Uniformly Most Powerful (UMP) similar invariant – is based on the well-known Student's t test for paired observations, $T = \overline{X} \cdot \sqrt{n}/\hat{\sigma}$, where $\hat{\sigma}^2 = \sum_i (X_i - \overline{X})^2/(n-1)$ and $\overline{X} = \sum_i X_i/n$, because differences are i.i.d. normal: $X_i \sim \mathcal{N}(\delta, \sigma_X^2)$.

Under H_0, the distribution of T is central Student's t with $n-1$ degrees of freedom (d.f.). Under H_1, it is distributed as a non-central Student's t with – positive – non-centrality parameter $\sqrt{n}\delta/\sigma_X$, so that large values are significant. Note that unknown standard deviation σ_X is the only nuisance entity for the problem and T is an invariant statistic on both σ and σ_X. Note also that $\hat{\sigma}$ is a minimal – complete – sufficient statistic for σ_X, either under H_0 or H_1. Using the data of the example, we have $T^o = 4.84$ with 19 d.f., which would lead to the rejection of H_0 at $\alpha = 0.001$.

One somewhat slightly more efficient parametric solution may be found via covariance analysis, if the pairs (Y_{1i}, Y_{2i}), $i = 1, \dots, n$, were i.i.d. bivariate normally distributed and if baseline Y_1 were considered as covariates for the problem. However, it is worth noting that although the bivariate normality of (Y_1, Y_2) implies normality of X, the converse is not true. Actually, normality of X is more frequently valid than the bivariate normality of (Y_1, Y_2). In our specific case, however, normality of X is an assumption which is difficult to justify because IPAT data are aggregates of a finite number of discrete scores, each related to an aspect of anxiety. Of course, the assumption of bivariate normality for the pair (Y_1, Y_2) is even more questionable.

Remark 1.5 When responses follow model (M.ii) in Section 1.8.1, i.e., $Y_{1i} = \mu + \eta_i + \sigma \cdot Z_{1i}$, $Y_{2i} = \mu + \eta_i - \delta + \sigma \cdot Z_{2i}$, $i = 1, \dots, n$, then the covariance analysis method becomes much more difficult or even impossible. Moreover, when $Y_{1i} = \mu + \eta_i + \sigma_i \cdot Z_{1i}$, $Y_{2i} = \mu + \eta_i - \delta + \sigma_i \cdot Z_{2i}$, $i = 1, \dots, n$, i.e. when unknown standard deviations are dependent on units, then no parametric solution can be obtained unless the σ_i and the within-unit correlation coefficients ρ_i, $i = 1, \dots, n$, were all known.

Remark 1.6 Student's t can also be applied to response models such as $Y_{1i} = \mu + \eta_i + \sigma_1 \cdot Z_{1i}$, $Y_{2i} = \mu + \eta_i - \delta + \sigma_2 \cdot Z_{2i}$, $i = 1, \dots, n$, where the two scale coefficients σ_1 and σ_2 may not be equal with respect to measurement occasions but

pairs (σ_1, σ_2) are invariant with respect to units, provided that underlying errors are normal. This solution has been used by Scheffé (1943c) in a randomized test for the Behrens–Fisher problem.

1.8.5 The Signed Rank Solution

Let us assume that P is unknown and X is a continuous variable, so ties in the observations are assumed to occur with probability zero. In this situation, P as a whole must be considered a nuisance entity for the testing problem, and a solution must be found either by using group-invariance arguments or by conditioning on a set of sufficient statistics for P.

Applying invariance arguments and assuming homoscedasticity with respect to units, a suitable solution based on ranks of absolute values of differences is provided by the well-known Wilcoxon signed rank test (e.g. Randles and Wolfe, 1979; Hollander and Wolfe, 1999; Lehmann, 2006, etc.). Notably, in this context, we need not assume that $\mathbb{E}(X)$ is finite.

Wilcoxon's test is based on the statistic $W = \sum_i R_i \cdot w_i$, where $R_i = \mathbb{R}(|X_i|) = \sum_{1 \le j \le n} \mathrm{I}(|X_j| \le |X_i|)$, where $\mathrm{I}(\cdot) = 1$ if relation $(\cdot)$ is satisfied and 0 elsewhere, are the ordinary ranks of the absolute values of differences $|X_i|$, $w_i = 1$ if $X_i > 0$ and $w_i = 0$ if $X_i \le 0$, $i = 1, \dots, n$, and $\mathbb{R}$ is the rank operator. If there are no ties, we have $\mathbb{E}(W|H_0) = n(n+1)/4$ and $\mathbb{V}(W|H_0) = n(n+1)(2n+1)/24$, whereas, in H_1, the mean value is larger than $n(n+1)/4$. Moreover, if n is not too small, in H_0, the distribution of $\{W - \mathbb{E}(W|H_0)\}/\sqrt{\mathbb{V}(W|H_0)}$ is well approximated by a standard normal variable.

A test statistic which is equivalent to W is $T = \sum_i R_i \cdot Sg(X_i)$, where $Sg(X_i) = 1$ if $X_i > 0$ and -1 if $X_i < 0$, that in H_0 has mean value $\mathbb{E}(T) = 0$ and variance $\mathbb{V}(T) = \sum_i R_i^2 = n(n+1)(2n+1)/6$ (Randles and Wolfe, 1979).

Unfortunately, the data of the example do not allow the direct use of this test because of the excessive number of ties due to the non-continuity of X. In this case, the permutation distribution of Wilcoxon's signed rank test cannot be approximated by its asymptotic counterpart, so it must be directly evaluated through specific calculations. However, it is worth noting that, assuming continuity of X, the test is distribution-free, and hence also P-invariant.

Remark 1.7 If, in place of ordinary ranks, a version of the so-called *generalized scores* is used, $\varphi_i = \varphi(R_i)$, $i = 1, \dots, n$, we can obtain other nonparametric solutions. Among the many, each used for specific purposes, those related to the standard normal distribution are the most popular. They consist of replacing ordinary ranks R_i with the related normal scores, $\zeta_i = \Phi^{-1}(R_i/(n+1))$ or $\varsigma_i = \mathbb{E}(Z_{(R_i)})$, $i = 1, \dots, n$, respectively, for the well-known van der Waerden and Fisher–Yates solutions, where Φ is the standard normal CDF and $Z_{(R_i)}$ is the R_ith-order statistic of n i.i.d. random elements from a standard normal variable.

1.8.6 The McNemar Solution

If no assumption regarding the continuity of X can be made, we must still consider P as a nuisance entity. With this relaxed assumption, an invariant solution can be found via the binomial test.

Let $U = \#(X_i > 0) = \sum_i \mathrm{I}(X_i > 0)$ and $\nu = \#(X_i \neq 0)$ be, respectively, the numbers of positive differences and non-null differences. Thus, in H_0, the statistic U is binomially distributed with parameters ν and $1/2$, $U \sim \mathcal{B}n(\nu, 1/2)$ say. In H_1, U is still binomially distributed, but with parameters ν and $\vartheta = \Pr\{X > 0\} > 1/2$, so that large values of U are significant. This solution essentially corresponds to the one-sided McNemar test, also called the sign test.

With the data from our problem, we have $\nu = 20$, $U = 17$, and $\Pr(U \geq 17|\mathbf{X}) = \sum_{i \geq 17} \binom{20}{i} 2^{-20} = 0.0013$, which is significant at $\alpha = 0.005$. Note that, when responses are binary, this test is UMP conditional for one-sided alternatives.

It is worth observing that McNemar's solution depends essentially on the number ν of non-null differences, in the sense that removing from the analysis all units and responses with null differences leads to the same result. One problem, which immediately arises, is concerned with how it is possible to obtain solutions by also including the $n - \nu$ null differences – for suggestions regarding external/auxiliary randomization procedures, see Lehmann (1986) and Randles (2001); for further suggestions, see Remark 2.41 and Problem 8, Sections 1.9.4 and Problem 1, Section 2.8.1.

These two testing solutions do look slightly different in that, in the latter, null differences are assumed to be informative of a substantially null treatment effect, whereas in the former, they seem to be considered as totally non-informative. However, it should be noted that this argument appears as rather apparent. On the one hand, if we consider the permutation confidence interval for the treatment effect, we see that all null differences play their part in the analysis as well as all other differences (Remark 2.41). On the other, in the multivariate case, all observed data vectors must be processed in order to maintain underlying dependence relations among responses (Chapters 4, 7, and 10).

Remark 1.8 Of course, McNemar's solution can also be used, in an obvious way, to test for a median in one-sample problems. Indeed, suppose that variable X is continuous, $\mathbf{X} = \{X_i, i = 1, \ldots, n\}$ are the data, and $H_0 : \{\mathbb{M}d(X) = \tilde{\mu}\}$. Let $\nu = \sum_{i \leq n} \mathrm{I}(X_i \leq \tilde{\mu})$. Thus, under H_0, $\nu \sim \mathcal{B}n(n, 1/2)$ and so the solution is clear. Further extensions are presented in Example 3.6.

Remark 1.9 McNemar's test may be applied in the case of non-homogeneity in distribution of experimental units: $P_i \neq P_j$, $i \neq j$, i.e. when the components of $\mathbf{X}$ are independent but not identically distributed. In terms of the response models of Section 1.8.1, this means that the distributions of random errors (Z_{1i}, Z_{2i}) may

vary with respect to units and that, in particular, they may have non-constant scale coefficients σ_i, $i = 1, \dots, n$. This fact allows this test to be used even in some cases under lack of homogeneity of experimental conditions. Therefore, it can be applied when there are censored paired observations (Good, 1991). Of course, non-constant scale coefficients may have influence on power behaviour. McNemar's solution may also be used when responses are ordered categorical and differences correspond to either positive or negative variations (Problems 7 and 8, Section 1.9.4; Examples 3.6–3.9; Chapter 6).

1.9 The Permutation Solution

1.9.1 General Aspects

Roughly speaking, permutation solutions are conditional on the whole set of observed data $\mathbf{X}$, which is always a set of sufficient statistics for any kind of underlying non-degenerate distribution P (Section 1.3). Let us now examine one solution to our problem under the assumption that P is unknown and that the nonparametric family $\mathcal{P}$ of distributions only contains non-degenerate distributions including discrete, continuous and mixed. Note that because of conditioning and assumed independence of the n units, the multivariate distribution P is $\Pi_i P_i$, where P_i is specific to the ith unit. In terms of response models of Section 1.8.1, this is consistent with the fact that the error deviates (Z_{1i}, Z_{2i}) might be not equally distributed with respect to units (Section 1.8.3). Thus, the conditioning on a set of sufficient statistics allows relaxation of the identical distribution condition for all units. Moreover, we note that to get consistency, we need to assume that $\mathbb{E}(X)$ and $\mathbb{E}(Z)$ are finite (Section 2.5.3).

1.9.2 The Permutation Sample Space

Let us proceed heuristically. So, first observe that the null hypothesis $H_0 : \{Y_1 \stackrel{d}{=} Y_2\}$ implies that the two variables Y_1 and Y_2 are exchangeable within each unit. This means that, in H_0, the two observed values of each unit behave as if they were randomly assigned to two occasions. In other words, the sign of each difference X_i, $i = 1, \dots, n$, is considered as if it were randomly assigned with probability $1/2$. Thus, one way to solve the testing problem is to consider a test statistic of the form $T = \sum_i X_i$, whose conditional distribution $F_T(t|\mathbf{X})$, when the observed points $\mathbf{X} = \{X_i, i = 1, \dots, n\}$ are held fixed, is obtained under the assumption that the assignment of the plus or minus sign to each difference with equal probability in all possible ways – a formal derivation of this statement within the conditional approach is given in Remark 2.4. This may be done by considering the

distribution of $T^* = \sum_i X_i^*$, where X_i^* is obtained by attributing the sign + or − to X_i, $i = 1, \ldots, n$, with probability 1/2. Observe that the probability distribution of $\mathbf{X}^* = \{X_i^*,\ i = 1, \ldots, n\}$, conditional on $\mathbf{X}$, is uniform within the permutation sample space $\mathcal{X}_{/\mathbf{X}}$. That is, all points of $\mathcal{X}_{/\mathbf{X}}$ are equally likely (Property 2.1).

The permutation sample space $\mathcal{X}_{/\mathbf{X}}$ of our example then contains $M^{(n)} = 2^n$ points, in spite of the fact that the permutation of signs is ineffective on the $n - \nu$ null differences. Apparently, it seems that this solution takes into consideration only the non-null differences. We shall see in Sections 2.7 and 2.8 that when determining a conditional power function or a conditional confidence interval for treatment effect δ, null differences enter the process as well as the non-null, and so they cannot be discarded from analysis.

Let us denote by $F(t|\mathbf{X}) = \Pr\{T^* \leq t|\mathbf{X}\}$, $t \in \mathcal{R}^1$, the permutation conditional CDF induced by T, given $\mathbf{X}$. Observe that $F(t|\mathbf{X})$ always exists because, by assumption, $\mathbf{X}$ is a measurable entity with respect to measurable space $(\mathcal{X}, \mathcal{A})$, which is assumed to exist and to be well defined.

Remark 1.10 In H_1, the permutation CDF of T is stochastically larger than that of T in H_0; so, large values of T are significant and the test is unbiased – formal proofs of these ordering properties are reported in Sections 2.2.4 and 2.5.2. In practice, by using $T^o = T(\mathbf{X})$ to indicate the observed value of T, if the p-value-like statistic $\lambda = \Pr\{T^* \geq T^o|\mathbf{X}\}$ is larger than α, whatever the fixed value of α, then H_0 is accepted, according to traditional testing rules – refer to Section 2.2.4 for a formal justification of the use of p-value-like statistics.

Remark 1.11 When the underlying model is $Y_{1i} = \mu + \eta_i + \sigma_i \cdot Z_{1i}$, $Y_{2i} = \mu + \eta_i - \delta + \sigma_i(\delta) \cdot Z_{2i}$, $i = 1, \ldots, n$, so that location and scale coefficients are both not invariant on units and treatment levels, the permutation solution remains effective (Problem 11, Section 2.5.4). Instead, when the response model is $Y_{1i} = \mu + \eta_i + \sigma_1 \cdot Z_{1i}$, $Y_{2i} = \mu + \eta_i - \delta + \sigma_2 \cdot Z_{2i}$, $i = 1, \ldots, n$, where the two scale coefficients σ_1 and σ_2 are not equal, due to the lack of exchangeability within units, the permutation solution based on $T^* = \sum_i X_i^*$ is generally not exact. However, if under this model, the error terms Z_{1i} and Z_{2i} are both symmetrically distributed around zero, exact permutation solutions exist (Section 3.1).

1.9.3 The Conditional Monte Carlo Method

1.9.3.1 An Algorithm for Inspecting Permutation Sample Spaces

In the case of our example, as well as in all cases where sample sizes are not small, the cardinality $M^{(n)} = \#[\mathbf{X}^* \in \mathcal{X}_{/\mathbf{X}}]$ of the permutation sample space $\mathcal{X}_{/\mathbf{X}}$ – which is finite if n is finite – is too large to examine all its points. According to many authors since Dwass (1957), to obtain reliable estimates of functionals of interest, we can

inspect this space by means of a random sample from it – the idea of random sampling from $X_{/\mathbf{X}}$ goes back to Eden and Yates (1933) in the spirit of sampling from finite populations (Cochran, 1952). This idea is realized by a simulation of the testing problem conditional on the observed data $\mathbf{X}$, i.e. by a *without replacement experiment* (WORE) (Pesarin, 1992, 2001, also Chapter 2). In some simple cases (Hemerik and Goeman, 2018), it is possible to find exact evaluations by random sampling on $\mathcal{X}_{/\mathbf{X}}$.

Note that the term *Conditional Monte Carlo* (CMC) is used to emphasize that an ordinary Monte Carlo simulation is carried out on the permutation sample space $\mathcal{X}_{/\mathbf{X}}$, where the set of observed points $\mathbf{X}$ is held fixed. The term *conditional resampling* procedure underlines a without replacement resampling from the observed dataset, considered as a finite population.

Essentially, the CMC algorithm operates according to the steps:

- (S.a) Calculate, on the given data $\mathbf{X}$, the observed value T^o of the test statistic T: i.e. $T^o = T(\mathbf{X})$.
- (S.b) For each of the n differences in $\mathbf{X}$, consider a random assignment of signs, obtaining the data permutation $\mathbf{X}^* = \{X_i \cdot S_i^*, i = 1, \dots, n\}$.
- (S.c) Calculate $T^* = T(\mathbf{X}^*)$.
- (S.d) Independently, repeat R times steps (S.b) and (S.c).
- (S.e) The R data permutations $\mathbf{X}^*$ are a random sample from the permutation sample space $\mathcal{X}_{/\mathbf{X}}$.

Thus, the R corresponding values of T^* simulate the permutation distribution of T. Therefore, they permit the statistical estimation of the conditional CDF $F_T(t|\mathcal{X}_{/\mathbf{X}})$ and of the significance level – or survival – function (SLF) $L_T(t|\mathcal{X}_{/\mathbf{X}}) = \Pr\{T^* \geq t|\mathcal{X}_{/\mathbf{X}}\}$ by, respectively, their empirical versions: the EDF $\hat{F}_T^*(t) = \sum_{1\leq r\leq R} \mathrm{I}(T_r^* \leq t)/R = \#(T^* \leq t)/R$ and the Empirical Survival Function (ESF) $\hat{L}_T^*(t) = \sum_{1\leq r\leq R} \mathrm{I}(T_R^* \geq t)/R, \forall t \in \mathcal{R}^1$. The subscript T in $F_T(\cdot)$, $L_T(\cdot)$, etc. can be omitted when the context suffices to avoid misunderstandings.

Step (S.b) may be easily obtained by the rule $X_i^* = X_i \cdot S_i^*$, $i = 1, \dots, n$, where the random variables S_i^* are i.i.d. taking value -1 or $+1$ with equal probability, according to the function $S^* = 2 \cdot \lfloor 2 \cdot \texttt{Rnd} \rfloor - 1$, and `Rnd` is a pseudo-random number in the open interval $(0, 1)$ with $\lfloor(\cdot)\rfloor$ being the integer part of $(\cdot)$.

Estimates in step (S.e) are such that the higher the number R of CMC runs, the more closely in probability $\hat{F}_T^*(\cdot)$ and $\hat{L}_T^*(\cdot)$ estimate $F_T(\cdot|\mathcal{X}_{/\mathbf{X}})$ and $L_T(\cdot|\mathcal{X}_{/\mathbf{X}})$, respectively. In any case, $\hat{F}_T^*(\cdot)$ or $\hat{L}_T^*(\cdot)$ may conveniently be used in place of $F_T(\cdot|\mathcal{X}_{/\mathbf{X}})$ or $L_T(\cdot|\mathcal{X}_{/\mathbf{X}})$ for evaluating the agreement of the observed data to H_0. In practice, the estimated p-value statistic, which in turn corresponds to the ESF evaluated at the observed value T^o, is given by

$$\hat{\lambda}_T(\mathbf{X}) = \hat{\lambda} = \hat{L}_T^*(T^o) = \sum_r \mathrm{I}(T_r^* \geq T^o)/R.$$

Table 1.2 Conditional Monte Carlo method.

$\mathbf{X}$	$\mathbf{X}_1^*$	…	$\mathbf{X}_r^*$	…	$\mathbf{X}_R^*$	$\rightarrow \hat{\lambda}(X) = \sum_{r=1}^{R} \mathrm{I}(T_r^* \geq T^o)/R$
T^o	T_1^*	…	T_r^*	…	T_R^*	

If $\hat{\lambda} \leq \alpha$, we may conclude that the empirical evidence disagrees with H_0, which should be rejected in accordance with traditional rules – in Section 2.5.1 and beyond, quantity $\hat{\lambda}(\cdot)$ are called p-value-like statistics.

Table 1.2 summarizes the CMC procedure: the first line contains the data $\mathbf{X}$ and the R permutations $\mathbf{X}^*$ randomly chosen from $\mathcal{X}_{/\mathbf{X}}$; the second contains the corresponding values of T.

Remark 1.12 If, in place of R random permutations, all possible permutations were considered, then the functions $F_T(t|\mathcal{X}_{/\mathbf{X}})$, $L_T(t|\mathcal{X}_{/\mathbf{X}})$, and p-value statistic λ were exactly determined. However, due to the well-known Glivenko–Cantelli theorem, the estimated value $\hat{\lambda}$, as R tends to infinity, tends almost surely to the true λ (Section 2.2.4). Of course, the greater the number R of CMC runs, the closer in probability the estimate $\hat{\lambda}$ is to its true value; further, R can be stated in an obvious way so that $\Pr\{|\hat{L}_T^*(t) - L_T(\cdot|\mathcal{X}_{/\mathbf{X}})| < \varepsilon\} > \eta$, for any suitable choice of $\varepsilon > 0$ and $0 < \eta < 1$. When H_0 is not true, with $\delta > 0$, the rejection probability power, $\Pr\{\lambda(\mathbf{X}(\delta)) \leq \alpha\} \geq \alpha$ say, monotonically increases in δ (formal proofs are in Problem 8, Section 2.4.2 and Property 2.9), and so T is unbiased.

In the specific example, with $R = 10\,000$ CMC runs, we obtain $\hat{\lambda} = 0.0003$, which leads to rejection of H_0 at $\alpha = 0.001$. Note that the number of CMC runs might be smaller than 10 000, for instance, 2000 or 1000, without appreciable changes in the conclusions.

1.9.3.2 Approximating the Permutation Distribution

If the sample size n is large, a permutation central limit theorem (PCLT) (Section 2.12) may be applied in order to approximate the permutation CDF $F_T(\cdot|\mathcal{X}_{/\mathbf{X}})$ of T. To this end, according to our experience, we observe that:

- (a) If n is smaller than about 25, it is possible, by using proper computing tools which are easy to download from the internet or to implement on desktop computers, to exactly calculate T^* on all points of the permutation sample space and then to exactly obtain $F_T[t|\mathcal{X}_{/\mathbf{X}}]$ and $L_T[t|\mathcal{X}_{/\mathbf{X}}]$.
- (b) If n is greater than about 200, σ_X is assumed to be finite and the ratio $(\sum_i X_i^4)/(\sum_i X_i^2)^2$ is small (Section 2.12), then $F_T[t|\mathcal{X}_{/\mathbf{X}}]$ can be approximated by the PCLT. To this end, let us observe that the expectation and variance of S^* are, respectively, $\mathbb{E}(S^*) = 0$ and $\mathbb{V}(S^*) = 1$. Hence, $\mathbb{E}\left\{ \left(\sum_i X_i \cdot S_i^*/n\right) \middle| \mathcal{X}_{/\mathbf{X}} \right\} = 0$ and

$\mathbb{V}\left\{\left(\sum_i X_i \cdot S_i^*/n\right)\middle|\mathcal{X}_{/\mathbf{X}}\right\} = \sum_i X_i^2/n^2$ because, conditionally on $\mathbf{X}$, quantities X_i in T^* play the role of fixed quantities. Therefore, the standardized version

$$K^* = \left(\sum_i X_i \cdot S_i^*\right) \Big/ \left(\sum_i X_i^2\right)^{1/2},$$

being the standardized sum of n independent variables, is approximately standard normally distributed – further details on the asymptotic permutation behaviour of test statistics are given in Sections 2.11 and 2.12.

- (c) In all other cases, $F_T[t|\mathcal{X}_{/\mathbf{X}}]$ and $L_T[t|\mathcal{X}_{/\mathbf{X}}]$ may be approximated, to the desired degree of accuracy, by means of a CMC-based R runs.

Although in this example the sample size n is not sufficiently large for normal approximation, the standardized observed value is $K^o = 3.324$, with p-value $\lambda = 0.000\,44$, that is very close to the CMC estimate $\hat{\lambda}$.

Remark 1.13 Observe that test K^* is approximately normally distributed independently of the underlying population distribution P, whereas Student's t requires normality of P.

1.9.4 Problems and Exercises

1 Compare the standardized permutation test statistic K, introduced in Section 1.9.3.2, to Student's t appropriate for the same problem and find the main differences. In particular, show that the two are asymptotically equivalent.

2 Prove that the two test statistics, K as above and Student's t, are asymptotically equivalent under both H_0 and H_1.

3 Extend the test solution for paired observations to the one-sample problem of testing for symmetry. Note that (i) X is symmetric with respect to δ if $X - \delta$ is symmetric with respect to 0; (ii) X is symmetric with respect to 0 if and only if $\Pr\{X < -z\} = \Pr\{X > z\}, \forall z \in \mathcal{R}^1$; (iii) if X is symmetric with respect to 0, then $\Pr\{X < 0\} = \Pr\{X > 0\}$ (Section 2.7).

4 Extend the test solution for paired observations when the model for responses is of multiplicative form, $Y_{2i} = \rho \cdot Y_{1i} + \varepsilon_i$, $i = 1, \ldots, n$, so that $H_0 : \{\rho = 1\}$, whereas $H_1 : \{\rho > 1\}$.

5 Discuss the permutation solution for paired observations in the case where $\sigma_1 \neq \sigma_2$.

6 Draw a block diagram for a test of symmetry in a one sample-problem, according to Problem 4.

7 Show that when there are ties in the ordered categorical data, i.e. the number of zero variations is positive, a solution not conditional on non-null differences should imply auxiliary – external – randomization (Lehmann, 1986).

8 With reference to Section 1.9 and taking account of Problem 7, show that one way to take into consideration the $n - \nu$ null differences is by using auxiliary randomization, according to Lehmann (1986).

9 Prove that the CMC method for testing with paired quantitative observations, illustrated in Section 1.9, may also be used in the case of paired binary-ordered categorical observations.

10 Prove that, with reference to the same testing problem for paired observations, the permutation test $T_S^* = \sum_i \left(X_i \cdot S_i^*/|X_i|\right)$, which in the spirit of Anderson–Darling – corresponding to the sum of standardized summands – because $\mathbb{V}\{(X_i \cdot S_i^*/|X_i|)|\mathcal{X}_{/\mathbf{X}}\} = 1,\ i = 1, \dots, n,$ and where $X_i \cdot S_i^*/|X_i| = 0$ if $X_i = 0$, coincides with the binomial or McNemar test – note that this solution may be taken into consideration when individual distributions P_i are considerably different from each other.

11 Prove that the matched pairs problem is equivalent to that with paired observations (Remark 1.3).

12 Show that the McNemar test is no more than a test on paired binary observations, either ordered categorical or quantitative.

1.10 A Two-Sample Problem

Let us now discuss, as a second example, a problem (Pesarin, 2001) concerning the comparison of locations of two populations. In a psychological experiment to assess the job satisfaction in two samples of 20 workers, assumed to be homogeneous in respect of most important covariates, sex, age, general health, social status, etc. were examined through the response variable X, corresponding to the perceived job satisfaction. Variable X was evaluated by a proper psychological index consisting of a sum of a finite number of items each related to a specific sub-aspect. Twelve units – sample 1 – were classified as 'extroverted', X_1; the remaining 8 units – sample 2 – as 'introverted', X_2. So, the sample data from

distributions P_1 and P_2 were $\mathbf{X}_1 = \{X_{1i},\ i = 1, \dots, 12\}$ and $\mathbf{X}_2 = \{X_{2i},\ i = 1, \dots, 8\}$, respectively. The testing problem was to show whether the data better agree to the null hypothesis of no difference in distribution, or to the alternative of a difference in favour of 'extroverted'. The test is then one sided – i.e. for restricted or dominance alternative. It is worth noting that since subjects are assigned to symbolic treatment levels – extroverted and introverted – after they were observed, so subjects were not randomized to treatments. Thus, this looks as a typical observational study where the treatment is merely a post-hoc classification (Remark 2.5 for some post-classification and post-stratification problems). However, since the null hypothesis assumes that there is no distributional difference between two treatment levels, instead of permuting subjects we are allowed to permute the observed data (Section 1.5).

1.10.1 Modelling Responses

Data are reported in Table 1.3. Formally, the hypotheses being tested are $H_0 : \{X_1 \overset{d}{=} X_2\}$ against $H_1 : \{X_1 \overset{d}{>} X_2\}$. Note that H_1 asserts the stochastic dominance of X_1 over X_2. This stochastic dominance may be specified according to several response models, the most important of which are:

- (M.i) A model with fixed additive effects: $X_{1i} = \mu + \delta + \sigma \cdot Z_{1i}, X_{2i} = \mu + \sigma \cdot Z_{2i}$, $i = 1, \dots, n_j, j = 1, 2$, where δ is the treatment effect, i.e. the *location functional*, Z_{ji} are exchangeable errors with null location and unit scale coefficient that is not dependent on units and treatment levels – note homoscedasticity.
- (M.ii) A model with generalized stochastic effects: $X_{1i} = \mu + \Delta_{1i} + \sigma \cdot Z_{1i}, X_{2i} = \mu + \sigma \cdot Z_{2i}, i = 1, \dots, n_j, j = 1, 2$, where μ is a population constant, Z_{ji} and σ are as in (M.i), and $\Delta_{1i} \geq 0$ are non-negative random quantities representing individually specific effects, which may depend on (μ, Z_{1i}) but are independent with respect to units, even though not identically distributed.

Model (M.ii) is compatible with any kind of stochastic dominance, i.e. $H_1 : \{P_1 < P_2\}$. In particular, with: (a) fixed effect model (M.i) when, with probability one, $\Delta_{1i} = \delta$; (b) a multiplicative effect model for positive responses, where $X_{1i} = \delta \cdot (\mu + \sigma \cdot Z_{1i})$, with $\delta \geq 1$; (c) an individually varying fixed effect model, where $X_{1i} = \mu + \sigma \cdot Z_{1i} + \delta_{1i}$, when $\Delta_{1i} = \delta_{1i}$ with probability one; (d) a model where the treatment may influence both location and scale coefficients, $X_{1i} = \mu + \delta + \sigma(\delta) \cdot$

Table 1.3 Job satisfaction of extroverted and introverted groups.

X_1 :	66	57	81	62	61	60	73	59	80	55	67	70
X_2 :	64	58	45	43	37	56	44	42				

Z_{1i}, where $\sigma(\delta)$ is any monotonic function of δ or of $|\delta|$, provided that the associated CDFs satisfy the stochastic dominance condition $P_1(x) \leq P_2(x)$, $x \in \mathcal{R}^1$ (Sections 2.2.4 and 2.5.2, for further details).

Also worth noting is that model (M.ii) is consistent with the so-called *placebo effect*. Under placebo effect, the null treatment typically assigned to units in the second sample is supposed to produce an effect, say δ_P with error W_{Pi}. Thus, we may model responses – with obvious notation – as $X_{ji} = (\mu + \delta_P) + (W_{Pi} + \sigma Z_{ji}) + \Delta_{ji}$, $i = 1, \ldots, n_j$, $j = 1, 2$. From this representation, we see that μ becomes $\mu + \delta_P$ and σZ changes to $\sigma Z + W_P$, so that the placebo effect is included in the population constant and errors.

Furthermore, when the response model becomes $X_{1i} = \mu + \delta + \sigma_1 \cdot Z_{1i}$, $X_{2i} = \mu + \sigma_2 \cdot Z_{2i}$, $i = 1, \ldots, n_j$, $j = 1, 2$, where homoscedasticity is not assumed even under H_0, then we refer to the generalized Behrens–Fisher problem in which, of course, the exchangeability condition is violated and so we have to look for approximate solutions (Example 4.13).

Along with this introductory chapter, we refer to the fixed effect model (M.i).

Remark 1.14 Under the alternative H_1, the generalized effect model (M.ii) does not imply homoscedasticity of responses. We recall that homoscedasticity implies $P_1(x + \delta) = P_2(x)$, $\forall x \in \mathcal{R}^1$, where δ is also named the *size-effect* or *location shift pseudo-parameter*. In the nonparametric setting, the terms *functional, size-effect*, and *pseudo-parameter* for the treatment effect δ are commonly used (Section 1.3.2).

Remark 1.15 With some testing problems, it is common to refer to two-sided alternatives, which are usually stated as $H_1 : \{X_1 \overset{d}{\neq} X_2\}$. This notation is quite ambiguous (Hubbard et al., 2019; Pesarin, 2019). What is usually intended is that the alternative H_1 is **either** $F_1(x) \leq F_2(x)$ **or** $F_1(x) \geq F_2(x)$, with strict inequality in a set of positive probability – sometimes referred to as two CDFs do not cross and not $F_1(x) \neq F_2(x)$, where two distributions are not equal. In the former notation, it is presumed that the effect, fixed or random, is either positive or negative, *but not both*. In the latter, instead, it is presumed that the effect can be positive on some set of points and negative on others. Such testing problems, named *multi-sided*, are much more intriguing and are partially discussed in Example 4.10, after the introduction of NPC and multi-aspect testing. There are, however, situations in which the two-sided testing assumes a clear meaning. For instance, two of such cases are: testing for normality, where $H_0 : \{P \textit{ is normal}\}$ against $H_1 : \{P \textit{ is not normal}\}$; testing for independence in contingency tables, where $H_0 : \{P(X, Y) = P_1(X) \cdot P_2(Y)\}$ against $H_1 : \{P(X, Y) \neq P_1(X) \cdot P_2(Y)\}$; and so on.

1.10.2 The Student *t* Solution

If we assume that responses of two populations are homoscedastic and normally distributed with σ unknown, this problem is efficiently solved – UMP similar invariant – by the one-sided Student's t for comparing two means. That is

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\left[\sum_{ji}(X_{ji} - \overline{X}_j)^2\right]^{1/2}} \sqrt{\frac{n_1 n_2 (n-2)}{n}},$$

where $\overline{X}_j = \sum_{i \le n_j} X_{ji}/n_j, j = 1, 2$, are sample averages, $n = n_1 + n_2$ the pooled sample size. The null distribution is Student's t with $n - 2$ d.f.

Assuming normality in the present case is rather unnatural because responses are integer numbers whose empirical distributions look slightly asymmetric. Thus, we can say that Student's t can only provide an approximate solution for which it is difficult to assess the rate. Moreover, if an underlying continuous model for responses is assumed, then observed integer values can be considered as truncated data, so resulting estimate of population variance becomes biased downwards and consequent inference conservative. However, the results are $t = 4.237$ with 18 d.f. which would be significant at $\alpha = 0.001$.

1.10.3 The Permutation Solution

Maintaining the assumption of homoscedasticity in the null hypothesis, with reference to the additive fixed effect model (M.i), we can relax normality and assume that the data distributions P_1 and P_2 are non-degenerate from the nonparametric family $\mathcal{P}$. Accordingly, assuming population means are finite, we can write the hypotheses as $H_0 : \{X_1 \overset{d}{=} X_2\} = \{\delta = 0\}$ against $H_1 : \{X_1 \overset{d}{>} X_2\} = \{\delta > 0\}$.

Remark 1.16 In the permutation context, in order to apply a test statistic based on comparison of sampling averages, we need only assume that means of involved responses are finite (Section 2.5.3). If by chance we cannot assume population means are finite, we must use a test statistic based, for instance, on comparison of sampling medians, trimmed means or empirical distribution functions (EDFs). What is essential is that there is a location functional δ playing the role of treatment effect and that a proper empirical indicator is available for it. Also note that H_0 implies exchangeability of observed data with respect to treatment levels.

Since under H_0 the common distribution P is unknown, we have to proceed conditionally on a set of sufficient statistics for P. Such a set is $\mathbf{X} = \mathbf{X}_1 \uplus \mathbf{X}_2 = \{X(i), i = 1, \dots, n; n_1, n_2\}$, where n_1, n_2 and $n = n_1 + n_2$ are the sample sizes of

the two samples and pooled sample, respectively. The proof of the sufficiency of the pooled data $\mathbf{X}$ in H_0 is left as an exercise (Sections 1.2, 2.1.2, and 2.1.3). The pooled array $\{X(i), i = 1, \dots, n; n_1, n_2\}$ is called the *unit-by-unit representation* of the data $\mathbf{X}$. This assumes that $X(i)$ belongs to sample 1 if $1 \le i \le n_1$, otherwise it belongs to sample 2. Remember that we use the same symbol $\mathbf{X}$ to indicate both the list of sample data, considered as an n-dimensional variable, and the pooled vector of actual data, the distinction being clear from the context.

Remark 1.17 Observe that if in the null hypothesis $\sigma_1 \neq \sigma_2$, then a set of sufficient statistics is the vector of sampling data $(\mathbf{X}_1; \mathbf{X}_2)$. Here, it should be emphasized that the data is partitioned into two subsets $\mathbf{X}_1$ and $\mathbf{X}_2$, so that in this case, we are not allowed to exchange data between samples (see Example 4.13 for a discussion).

Conditioning on the whole data is equivalent to conditioning with respect to the EDF for P, which is also sufficient (note that in $H_0, P_1 = P_2$; Definition 2.2). Hence, in H_0, observed data may be viewed as if they were randomly assigned to two treatments. Thus, for this kind of problem, the permutation sample space $\mathcal{X}_{/\mathbf{X}}$ is exactly the set of all permutations of data $\mathbf{X}$, i.e. $\mathcal{X}_{/\mathbf{X}} = \Pi(\mathbf{X})$, the cardinality of which is $M^{(n)} = n!$ – in the example $n = 12 + 8 = 20$, so $M^{(n)} \approx 2.4329 \cdot 10^{18}$. Of course, if we use symmetric test statistics such as $T = \sum_i X_{1i}/n_1 - \sum_i X_{2i}/n_2$ or the like, which in turn are differences of two symmetric functions, being invariant on rearrangements of data entry, then this cardinality becomes $M^{(n)} = C_{n,n_1} = \binom{n}{n_1}$, leading to $C_{20,8} = 125\,970$, since in $\mathcal{X}_{/\mathbf{X}}$, there are $n_1! \cdot n_2!$ points sharing the same value of T^*.

If we assume that sampling means are proper indicators for the treatment effect δ, a suitable permutation test statistic is $T^*_{1,2} = \overline{X}^*_1 - \overline{X}^*_2$, where $\overline{X}^*_j = \sum_i X^*_{ji}/n_j$, $j = 1, 2$, are the sample averages of permuted data $\mathbf{X}^*$. However, $T^*_{1,2}$ is permutationally equivalent to $T^* = \sum_i X^*_{1i}$ because there is an increasing one-to-one relationship between the two statistics since the data $\mathbf{X}$ is held fixed (Section 2.4). Indeed, $\sum_{ji} X^*_{ji}$ and $\sum_{ji} X_{ji}$ are permutation invariant quantities. Thus, T^* and $T^*_{1,2}$ are related with a one-to-one increasing relationship.

In the framework of this problem, it may be seen that T^* is unbiased and consistent (see Chapter 2 for a discussion).

Remark 1.18 In a permutation framework, we need not consider standardized forms for the concerned test statistics because standardization is an increasing one-to-one relationship of main statistic. Hence, standardized and non-standardized forms are always permutationally equivalent (Section 2.4). Moreover, in order to get unbiasedness, we must assume that two CDFs, F_1 and F_2, do not cross each other (Sections 2.2.4 and 2.5.2 for further comments).

One way of inspecting $\mathcal{X}_{/\mathbf{X}}$ for a two-sample problem is by modifying step (S.2) of Section 1.9.3 (also Section 2.2.4) into:

- (S.2′) (i) Take a random permutation $(u_1^*, \dots, u_n^*)$ of unit labels $(1, \dots, n)$. (ii) According to the unit-by-unit representation, assign the first n_1 corresponding data to sample 1 and the other n_2 to sample 2, thus obtaining the data permutation $\mathbf{X}^* = \{X(u_i^*),\ i = 1, \dots, n;\ n_1, n_2\}$. (iii) Calculate the test statistic $T^* = T(\mathbf{X}^*)$.

The dataset and the corresponding software codes are available at http://www.wiley.com/go/npc in `examples_chapters_1-4` folder. With $R = 10\,000$ CMC runs, we obtained $\hat{\lambda} = 0.0002$, so the null hypothesis has to be rejected at $\alpha = 0.001$.

Remark 1.19 As permutations are without replacement random samples from $\mathbf{X}$, when sample sizes are sufficiently large, the permutation CDF $F_T(t|\mathcal{X}_{/\mathbf{X}})$ of $T^* = \sum_i X_{1i}^*$ may be approximated (see Section 2.12 for required conditions) by that of a normal distribution with mean value $\mathbb{E}(T^*|\mathcal{X}_{/\mathbf{X}}) = n_1 \cdot \sum_{ji} X_{ji}/n$ and variance $\mathbb{V}(T^*|\mathcal{X}_{/\mathbf{X}}) = n_1 \cdot n_2 \cdot \sigma_X^2/(n-1)$, where $\sigma_X^2 = \sum_{ji} X_{ji}^2/n - \left(\sum_{ji} X_{ji}/n\right)^2$ is the variance of pooled data $\mathbf{X}$ considered as a finite population.

1.10.4 Rank Solutions

If we assume P_1 and P_2 are continuous, the same problem might be solved by the well-known Wilcoxon–Mann–Whitney rank test. This is actually a permutation test based on ranks, $MW = \sum_i R_{1i}$, where $R_{ji} = \mathbb{R}(X_{ji}) = \sum_{gh} \mathrm{I}(X_{gh} \le X_{ji})$ are the ranks of X_{ji}, $i = 1, \dots, n_j$, $j = 1, 2$, in the pooled set $\mathbf{X}$. Alternatively, one of its permutationally equivalent forms may be considered. Since the mean value and variance of $\sum_i R_{1i}$ in H_0 are $\mathbb{E}(\sum_i R_{1i}) = n_1(n+1)/2$ and $\mathbb{V}(\sum_i R_{1i}) = n_1 n_2 (n+1)/12$, respectively, the standardized version of MW is $T_{MW} = [\sum_i R_{1i} - n_1(n+1)/2]/[n_1 n_2 (n+1)/12]^{1/2}$. If sample sizes are not too small, the null distribution of T_{MW} is well approximated by a standard normal distribution. With the data from the example, as $\sum_i R_{1i} = 164$, we have $T_{MW} = 4.146$, which is significant at $\alpha = 0.001$ – this can be compared with the standard normal distribution because sample sizes $n_1 = 12$ and $n_2 = 8$ suffice for normal approximation.

Note that in general, rank transformations are not one-to-one with respect to data $\mathbf{X}$, so they lose the sufficiency property for P, although, under continuity of P, MW is a *maximal invariant test*. This often, but not always, implies some power decay – for the problem of a reasonable choice of a test statistic for finite sample sizes, see the discussion in Section 2.6. Of course, other nonparametric solutions are available, according to specific assumptions regarding P.

Furthermore, if assumptions suggest that medians or any other robust statistic are to be preferred as proper indicators of treatment effects, in place of mean

values, one solution is to use a permutation test statistic of the form $\tilde{T} = \tilde{X}_1 - \tilde{X}_2$, where $\tilde{X}_j, j = 1, 2$, represents the median or any other robust statistics calculated on the jth data sample. Alternatively, but not equivalently, another solution is to use the so-called Mood median test. Also, if no assumption regarding the equality of scale parameters with respect to treatment levels or more generally if no assumption of stochastic dominance can be made, another solution is to use a permutation Behrens–Fisher or a multi-sided kind of test (see Examples 4.10 and 4.13 for a discussion).

1.10.5 Fisher's Exact Probability Solution

While using the empirical global median $\mathbb{M}d(\mathbf{X}) = \tilde{X} = 60$ as a classification point, the data in the example provide for $f_{10} = 3$ units smaller than 60 in the first sample and $f_{20} = 7$ in the second – respectively, with $f_{11} = 9$ and $f_{21} = 1$ not smaller than $\tilde{X}$. Fisher's exact probability test calculates the probability of finding tables that are as extreme as the given one under the condition of taking the marginals $(n_1, n_2; f_{\cdot 0}, f_{\cdot 1}) = (12, 8; 10, 10)$ fixed. It is known that such a conditioning gives rise to an hypergeometric distribution (Randles and Wolfe, 1979). It is to be emphasized that if data are transformed according to $Y = \mathrm{I}(X \geq 60)$, i.e. $Y = 1$ if $X \geq 60$ and 0 otherwise, marginal data are sufficient statistics for the underlying distribution and exchangeable under the (sub) null hypothesis $H_0 : \Pr\{Y_1 = 0\} = \Pr\{Y_2 = 0\}$. Thus, Fisher's exact probability test is nothing else than a permutation test on two-sample binary data. In this setting, the six digits exact p-value is $\lambda = 0.009883$, significant at $\alpha = 0.01$. Of course, if in place of $\tilde{X} = 60$ we choose a different classification point, consequent inference might be different. For instance, choosing $X' = 61$, so the table would be $(f'_{10} = 4, f'_{20} = 7; f'_{11} = 8, f'_{21} = 1)$ with marginals $(12, 8; 11, 9)$, we would have $\lambda' = 0.024887$, significant at $\alpha' = 0.05$. It is worth noting that the inferential result, depending on the chosen classification point, e.g. 60 or 61, is slightly unstable. This makes the binary transformation quite questionable in this case.

It is worth noting that in Fisher's exact test, the alternative is one-sided, i.e. restricted, $H_1 : \Pr\{Y_1 = 0\} < \Pr\{Y_2 = 0\}$; whereas the alternative with the standard χ^2 test is two-sided, i.e. unrestricted, $H'_1 : \Pr\{Y_1 = 0\} \neq \Pr\{Y_2 = 0\}$. In Example 4.12, we will see that multivariate testing for restricted alternatives, so including the extension of Fisher's exact test, are straightforward within the NPC; whereas parametric approaches may become extremely difficult.

1.10.6 Problems and Exercises

1 Discuss the permutation median test $T^*_{Md} = \tilde{X}^*_1 - \tilde{X}^*_2$ for the two-sample problem, where $\tilde{X}^*_j = \mathbb{M}d(\mathbf{X}^*_j) = X^*_{((n_j+1)/2)}$ if n_j is odd and $(X^*_{(n_j/2)} + X^*_{(1+n_j/2)})/2$ if n_j is even, where $X^*_{(1)} \leq X^*_{(2)} \leq \cdots \leq X^*_{(n_j)}$ are the order statistics of $\mathbf{X}^*_j, j = 1, 2$.

2 In the Behrens–Fisher problem, where the response model is $X_{ji} = \mu + \delta_j + \sigma_j \cdot Z_{ji}$, $i = 1, \dots, n_j, j = 1, 2$, with $\sigma_1 \neq \sigma_2$, and where Z_{ji} are exchangeable random deviates with null mean value, in which the hypotheses are $H_0 : \{\mathbb{E}(X_1) = \mathbb{E}(X_2)\}$ and $H_1 : \{\mathbb{E}(X_1) > \mathbb{E}(X_2)\}$, prove that the dominance of means, i.e. $\mu_1 > \mu_2$, does not imply dominance of responses: $X_1 \overset{d}{>} X_2$.

3 Show that if the response variable is binary, then the test statistic for testing $H_0 : \{X_1 \overset{d}{=} X_2\}$ against $H_1 : \{X_1 \overset{d}{>} X_2\}$ corresponds to Fisher's exact probability test, that rejects H_0 if $\Pr\{\sum_i X^*_{1i} \geq \sum_i X_{1i} | \mathcal{X}_{/\mathbf{X}}\} \leq \alpha$.

4 With reference to the two-sample problem for $H_0 : \{X_1 \overset{d}{=} X_2\}$, in which the two sample data are $\mathbf{X}_j = \{X_{ji},\ i = 1, \dots, n_j\}, j = 1, 2$, prove that the pooled set $\mathbf{X} = \mathbf{X}_1 \biguplus \mathbf{X}_2$ is a set of sufficient statistics for whatever underlying distribution P (Sections 1.2 and 2.1.2).

5 Using the unit-by-unit representation for the binary transformation of Fisher's test solution, find that the whole dataset $\mathbf{Y} = \{Y_i, i = 1, \dots, n; n_1, n_2\}$, except for an irrelevant sequence of units, under the sub-null hypothesis, $H_0 : \Pr\{Y_1 = 0\} = \Pr\{Y_2 = 0\}$ is one-to-one with respect to the 2×2 table representation $(n_1, n_2; f_{\cdot 0}, f_{\cdot 1}) = (12, 8; 10, 10)$ and that both are a set of minimal sufficient statistics for the problem.

1.11 One-Way ANOVA

As a third introductory example, let us consider a one-way ANOVA problem. It is well-known that this corresponds to testing for the equality in distribution of $C \geq 2$ samples of data, where C is the number of treatment levels in a symbolic experiment. In this setting, units belonging to the jth sample, $j = 1, \dots, C$, receive treatment at the jth level. When specific side-assumptions ensure that two parameter responses have finite means and are homoscedastic, i.e. $\mathbb{E}(|X_j|) < \infty$ and $\mathbb{V}(X_j) = \sigma^2$, $j = 1, \dots, C$, the equality of C distributions becomes that of C means.

To introduce this problem and justify a permutation solution for it, let us consider the data in Table 1.4 (from Pollard, 1977, p. 169). The related problem is concerned with the length of worms in three different samples, where the purpose is to test whether the mean lengths of the worms are the same. Formally, we may write $H_0 : \{\mu_1 = \mu_2 = \mu_3\}$ against the alternative $H_1 : \{$*at least one equality is false*$\}$.

Table 1.4 Length of worms in three groups.

Group		
1	**2**	**3**
10.2	12.2	9.2
8.2	10.6	10.5
8.9	9.9	9.2
8.0	13.0	8.7
8.3	8.1	9.0
8.0	10.8	
	11.5	

1.11.1 Modelling Responses

In the fixed effects additive response model, data are $\mathbf{X} = \{X_{ji} = \mu + \delta_j + \sigma \cdot Z_{ji}, i = 1, \dots, n_j, j = 1, \dots, C\}$, where μ is a population constant, δ_j are the fixed treatment effects which satisfy the contrast condition $\sum_j \delta_j = 0$, Z_{ji} are exchangeable random deviates with zero mean and unit scale parameter, σ is a scale coefficient which is assumed to be invariant with respect to samples and C is the number of samples into which the data are partitioned. Note that responses are assumed to be homoscedastic and that scale coefficients are assumed to be unaffected by the treatment levels even under H_1. If, in addition, data are normally distributed, this problem is solved by Fisher–Snedecor's well-known F test on the one-way ANOVA layout where $H_0 : \{\delta_1 = \delta_2 = \delta_3\}$ against $H_1 : \{H_0 \textit{ is not true}\}$. That is, with clear meaning of the symbols, by the test statistic:

$$F = \frac{\sum_{j=1}^{C} (\overline{X}_j - \overline{X}_\cdot)^2 n_j}{\sum_{ji} (X_{ji} - \overline{X}_j)^2} \cdot \frac{n - C}{C - 1},$$

whose null distribution is central Fisher's F with $C - 1$ and $n - C$ d.f.s for numerator and denominator, respectively.

Observe that, within homoscedasticity condition, the null hypothesis is equivalent to equality of three distributions: $H_0 : \{X_1 \stackrel{d}{=} X_2 \stackrel{d}{=} X_3\}$. Also observe that this equality implies that the data $\mathbf{X}$ are exchangeable; in particular, they may be viewed as if they were randomly assigned to samples.

While maintaining homoscedasticity, but not normality, and assuming the existence, in H_0, of a common non-degenerate, continuous, unknown distribution P, the problem may be solved by the Kruskal–Wallis rank test, or by any analogous test statistic based on generalized ranks, or even by conditioning on a set of

sufficient statistics, i.e. by a permutation procedure. Note that due to conditioning, the latter allows for relaxation of continuity for P and for relaxation of finite scale coefficients for responses. It only requires the existence of location coefficients and proper empirical indicators for them.

The permutation solution also allows for relaxation of some forms of homoscedasticity for responses in H_1. In fact, the generalized one-way ANOVA model allowing for unbiased permutation solutions assumes that the hypotheses are $H_0 : \{X_1 \overset{d}{=} X_2 \overset{d}{=} X_3\}$ against $H_1 : \{X_1 \overset{d}{\neq} X_2 \overset{d}{\neq} X_3\}$, with the restriction that, for every pair $h \neq j$, $h,j = 1,2,3$, the corresponding response variables are stochastically ordered pairwise dominance according to either $X_h \overset{d}{>} X_j$ or $X_h \overset{d}{<} X_j$, in such a way that, $\forall t \in \mathcal{R}^1$, the associated CDFs do not cross (Remark 1.15), i.e. $F_h(x) \leq F_h(x)$ or $F_h(x) \geq F_h(x)$, $\forall x \in \mathcal{R}^1$, with strict inequality in a set of positive probability.

Remark 1.20 Such a pairwise dominance assumption may correspond to a model in which treatment may affect both location and scale coefficients, as for instance in $\{X_{ji} = \mu + \delta_j + \sigma(\delta_j) \cdot Z_{ji},\ i = 1, \ldots, n_j,\ j = 1, \ldots, C\}$, where $\sigma(\delta_j)$ are monotonic functions of treatment effects δ_j or of their absolute values $|\delta_j|$, provided that $\sigma(0) = \sigma$ and pairwise stochastic ordering on CDFs are preserved. The latter model is consistent with the randomization notion (Section 1.5). Indeed, (i) units are assumed to be randomly assigned to treatment levels, so that H_0 implies exchangeability of responses; (ii) in the alternative, treatment may jointly affect location and scale coefficients, so that resulting permutation distributions become either stochastically larger or smaller than the null. Also note that the pairwise dominance assumption is consistent with a generalized model with random effects of the form $\{X_{ji} = \mu + \sigma \cdot Z_{ji} + \Delta_{ji},\ i = 1, \ldots, n_j,\ j = 1, \ldots, C\}$, where Δ_{ji} are the stochastic effects that satisfy the pairwise ordering condition that for every pair $h \neq j$, $i,j = 1,2,3$, either it is $\Delta_h \overset{d}{>} \Delta_j$ or $\Delta_h \overset{d}{<} \Delta_j$ – formal proofs are provided in Sections 2.2.4 and 2.5.2.

1.11.2 Permutation Solutions

Formalizing the testing problem for a C-sample one-way ANOVA layout, we assume that $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_C\}$ represents the data partitioned into C samples, where $\mathbf{X}_j = \{X_{ji},\ i = 1, \ldots, n_j\},\ j = 1, \ldots, C$, are i.i.d. observations from non-degenerate distributions P_j, respectively. It is helpful to use the unit-by-unit representation $\mathbf{X} = \{X(i), i = 1, \ldots, n; n_1, \ldots, n_C\}$, where it is assumed that $X(i) \in \mathbf{X}_1$ if subscript i satisfies the condition $1 \leq i \leq N_1$, $X(i) \in \mathbf{X}_2$ if $N_1 + 1 \leq i \leq N_2$, and so on, where $N_j = \sum_{h \leq j} n_h, j = 1, \ldots, C$, are cumulative sample sizes. We also assume that the sample means are proper indicators of treatment effects.

Under the homoscedastic model, the hypotheses are

$$H_0 : \{X_1 \stackrel{d}{=} \cdots \stackrel{d}{=} X_C\} = \{\delta_1 = \cdots = \delta_C = 0\}$$

against $H_1 : \{H_0 \textit{ is not true}\}$.

If it is suitable for analysis, we may consider a monotonic data transformation φ, so that related sampling means become proper indicators for treatment effects. According to the CMC procedure, runs are now on the pooled data $\mathbf{X} = \mathbf{X}_1 \biguplus \ldots \biguplus \mathbf{X}_C$, which is always sufficient for the problem. If symmetric test statistics are used, the related permutation sample space $\mathcal{X}_{/\mathbf{X}}$ contains $n!/(n_1! \cdot \ldots \cdot n_C!)$ distinct points, where $n = \sum_j n_j$ is the pooled sample size.

According to the above assumptions and Remark 1.18, a reasonable test statistic is the deviance among sample means:

$$T_C^* = \sum_{j=1}^{C} (\bar{Y}_j^* - \bar{Y}_\bullet)^2 \cdot n_j,$$

where $\bar{Y}_j^* = \sum_i \varphi(X_{ji}^*)/n_j$ and $\bar{Y}_\bullet = \sum_j \bar{Y}_j \cdot n_j/n$. Note that $\bar{Y}_\bullet$, being the sum of all observed data, is a permutationally invariant quantity. Hence, statistic T_C^* is permutationally equivalent to $T^* = \sum_{j=1}^{C} n_j \cdot (\bar{Y}_j^*)^2$ (Example 2.2). In Example 4.12 and Chapter 6, quite different multisample problems, one related to stochastic ordering –isotonic inference – the other on pair-wise comparisons, are discussed.

The dataset and the corresponding software codes are available at http://www.wiley.com/go/npc in `examples_chapters_1-4` folder. With $R = 10\,000$ CMC runs, we obtain $\hat{\lambda} = 0.0106$, which leads to the rejection of H_0 at $\alpha = 0.025$. This result fits with those obtained by Pollard (1977) with the parametric Fisher–Snedecor's F test: $F = 6.30$, with 2 and 15 d.f., and the Kruskal–Wallis KW rank test: $KW = 7.76$, the null distribution of which is approximately central χ^2 with 2 d.f., both significant at $\alpha = 0.025$. We recall that the Kruskal–Wallis permutation rank test is based on the statistic

$$KW = \left\{ \frac{12}{n(n+1)} \cdot \sum_{j=1}^{C} n_j \cdot \left[\bar{R}_j - \frac{n+1}{2} \right]^2 \right\},$$

where R_{ji} is the rank of X_{ji}, $j = 1, \ldots, C$, $i = 1, \ldots, n_j$, within the pooled data $\mathbf{X}$, and $\bar{R}_j = \sum_i R_{ji}/n_j, j = 1, \ldots, C$, is the jth sample mean rank. For moderately large sample sizes n_j, the null distribution of KW is approximated by that of a central χ^2 with $C - 1$ d.f.

1.11.3 Problems and Exercises

1 Discuss a solution to the one-way ANOVA when, in place of sample means $\overline{X}_j$, sample medians $\tilde{X}_j$ are assumed to be proper indicators for treatment effects.

2 Discuss Mood's median test for the one-way ANOVA and find that it is a permutation test.

3 Express the heuristic motivations for the choice of test statistic T in the one-way ANOVA.

4 Compare the previous permutation solution T to Fisher–Snedecor's F based on homoscedastic normal responses, the Kruskal–Wallis KW based on rank transformations and Mood's test based on frequencies above and below the pooled median. Discuss conditions in which one is better than the others.

5 Prove that the Kruskal–Wallis rank test is permutationally equivalent to $\sum_j n_j \cdot \overline{R}_j^2$ (Section 2.4).

6 Prove that for the one-way ANOVA problem, the permutation sample space $\mathcal{X}_{/\mathbf{X}}$ associated with $\mathbf{X}$ contains $n!/(n_1! \cdot \ldots \cdot n_C!)$ distinct permutations.