
# Permutation Tests

Jaime Turón Rodríguez

*Supervisor:*

Ana María Muñoz Reyes

Bachelor Thesis

Faculty of Mathematics

University of Seville

July 2021

*To Josefa, as I have longed for you to be alive to witness this.*

# Contents

# Abstract

In this Bachelor dissertation permutation tests are reviewed both theoretically and practically. Chapter 1 is a brief introduction to set the reader on the starting point of this project. Non-parametric methods and some basic concepts would be defined. Chapter 2 is the theoretical core of this dissertation. Permutation techniques are exhaustively introduced, as well as the construction of tests based on this idea, the behaviour of $p$-values, a succinct review of the critical region and an overview of permutational equivalence. Chapter 3 complements the previous chapter as theoretical framework development continues. Test unbiasedness, conditional power function, confidence intervals and general asymptotic properties are illustrated here. Finally, Chapter 4 shows some practical experiments where theoretical concepts approached in Chapters 2 and 3 are applied to real-world scenarios.

# Resumen

En este trabajo se presentan los tests de permutaciones de forma tanto teórica como práctica. El Capítulo 1 es una breve introducción para situar al lector en el punto de partida del proyecto. Se definen los métodos no paramétricos y algunos conceptos claves. El Capítulo 2 es el núcleo teórico del ensayo. Las técnicas de permutaciones se explican en profundidad, así como la construcción de tests basados en esta idea, el comportamiento de su $p$-valor, una breve introducción de la región crítica y el concepto de equivalencia permutacional. El Capítulo 3 complementa el capítulo anterior continuando el desarrollo teórico. La insesgadez de los tests, la funcion potencia condicionada, intervalos de confianza y algunas propiedades asintóticas generales son los protagonistas de esta parte. Finalmente, en el Capítulo 4 se exponen algunos experimentos prácticos donde los conceptos teóricos analizados previamente en los capítulos 2 y 3 se aplican a casos reales.

# Chapter 1

# Introduction

Permutation tests were simultaneously introduced by Fisher and Pitman in 1937.

Fisher introduced it in [1] by studying a previous experiment carried out by Charles Darwin and Francis Galton which consisted in measuring and seeking evidence for mean differences in a group of plants which were crossed and self fertilised respectively (in fact, this scenario would be our general framework: two populations with one of them having been through some kind of treatment). Fisher made a similar experiment by measuring crossed and self fertilised plants again. His purpose was to determine if both populations have the same mean, just as Darwin and Galton did. Fisher stated an equivalent hypothesis by assuming that both populations were drawn from the same distribution and compared the difference of means without making distinction between cross and self fertilised plants: he took two subsets and rearranged them in every way he could, studying the behaviour of the mean in both of them.

Though Fisher was a pioneer by using this kind of statistical method in an experiment, he did not formalise it theoretically. It was Pitman in [2] who formally introduced the concept of permutation tests. He wanted to propose statistical methods with no assumptions about the underlying population (this is known as non-parametric framework, more on that later). He introduced the notion of

splitting data and rearranging it by considering the resulting separations neutral, concordant or discordant depending on its behaviour with respect to the original data. The final decision on the hypothesis testing would be based on the number of each kind of separations.

Despite not working together, both Fisher and Pitman intended to introduce new statistical methods with little assumptions about the form of the distributions. This is know as non-parametric tests, or more generally, non-parametric statistics.

## 1.1 The metaphysical duality of non-parametric statistics

The definition of the non-parametric framework leads to a clear division between authors. This division is clearly illustrated in [1], where Fisher approached both trends.

*"In recent years tests using the physical act of randomisation to supply (on the Null Hypothesis) a frequency distribution, have been largely advocated under the name of "Non-parametric" tests. They assume less knowledge, or more ignorance, of the experimental material than do the standard tests, and this has been an attraction to some mathematicians who often discuss experimentation without personal knowledge of the material."*

*Ronald A. Fisher*

The first current of this concept is more anthropocentric, presenting non-parametric statistics as the researcher's own election. It is presented in [3] as follows:

*"The basic idea of nonparametric inference is to use data to infer an unknown quantity while making as few assumptions as possible. Usually, this means using statistical models that are infinite-dimensional. Indeed, a better name for non-*

*parametric inference might be infinite-dimensional inference. But it is difficult to give a precise definition of nonparametric inference, and if I did venture to give one, no doubt I would be barraged with dissenting opinions."*

*Larry Wasserman*

However, he clarifies the problem that stems from defining this concept. As it can be observed, the non-parametric framework is presented as pure decision and not as a result of ignorance or mathematical inability. This is the second approach when defining it. There are situations where data does not stick to some known population or statistical method. This is a more sceptical standpoint where ignorance and limitations are courageously assumed. This could be found in [4]:

*"However, in a real-world problem everything does not come packaged with labels of populations of origin. A decision must be made as to what populations properties may judiciously be assumed for the model."*

*J.D. Gibbons*

These are the two main currents of non-parametric statistics definitions. Let us now set our conditions to develop permutation tests.

## 1.2 Prerequisites

Permutation tests would be strongly based on conditional distributions. We will abuse of the use of sufficiency, so let us introduce it:

**Definition 1.2.1.** *Suppose that* $(X_1, ..., X_n)$ *have a joint distribution that depends on a vector of parameters* $\theta$ *for* $\theta \subset \Theta$ *where* $\Theta$ *is the parameter space. A statistic* $T(X_1, ..., X_n)$ *is a sufficient statistic for* $\theta$ *if the conditional distribution of* $(X_1, ..., X_n)$ *given* $T = t$ *does not depend on* $\theta$ *for any value of* $t$ *in the support of* $T$*.*

The classical interpretation of a sufficient test statistic is that no information about the distribution is lost when using the statistic instead of the sample itself.

We thus can make a one-dimensional reduction through sufficient statistics.

However, there would be some situations where this one-dimensional reduction would not be possible, and we will select a set of statistics $\{T_1, ..., T_p\}$ which would be jointly sufficient. This would be our later scenario: conditioning with respect to a set of sufficient statistics in order to free the distribution of any possible unknown parameters.

Another concept which would be relevant when working in this framework would be exchangeability.

**Definition 1.2.2.** *Let $X_1, ..., X_n$ be a finite set of random variables, with $p(X_1, ..., X_n)$ denoting their joint distribution. The variables $X_1, ..., X_n$ are said to be exchangeable if:*

$$p(X_1, ..., X_n) = p(X_{\pi(1)}, ..., X_{\pi(n)})$$

*for every permutation $\pi \in S(n)$, with $S(n)$ the symmetric group.*

For example, in case we are dealing with continuous variables (the existence of the density function is then assumed), exchangeability could be written as

$$\int_{\Omega} f(x_1, ..., x_n) = \int_{\Omega} f(\pi(x_1), ..., \pi(x_n))$$

in every non-null measure subset $\Omega$.

Exchangeability is a weaker notion than independence. We just want the order of the variables not to be relevant, which is slightly permissive than the strict notion of independence. What is more, it is obvious that independence implies exchangeability whereas it is not always true the other way.

This would be a basic concept when working with permutation tests. It would be necessary the null distribution to be exchangeable, as it seems pretty logic not to be able to differentiate between observations under two equally distributed populations.

No more early concepts would be needed in the development of this project, as new items would be gradually introduced and reviewed.

# Chapter 2

# Theory of Permutation Tests

In this chapter we will introduce the basic concepts of permutation tests. We will use the two-sample design as a guide, following the spirit of [7]. Conditionality and exchangeability would be our cornerstone to develop the theoretical body as described in 1.2.

As it was also said in advance, the conditioning will be done with respect to a set of sufficient statistics.

Let $\mathbf{X}$ be a random variable taking values on the sample space $\mathcal{X}$, and $P \in \mathcal{P}$ the underlying distribution belonging to a non-parametric family of distributions (if $\mathcal{P}$ were a parametric family, classical parametric methods could be employed). The existence of the density function of $P$, $f_P(x)$, is assumed. This function would sometimes be interpreted as the density function itself whereas it could be also used to refer to the likelihood function. No distinction will be made as the context will be self-explanatory to determine which one we are referring to.

Let $\mathbf{X_1} = \{X_{11}, ..., X_{1n_1}\} \in \mathcal{X}^{n_1}$ and $\mathbf{X_2} = \{X_{21}, ..., X_{2n_2}\} \in \mathcal{X}^{n_2}$ be two independent identically distributed samples data from the models $P_1$ and $P_2$ respectively, both $P_1, P_2 \in \mathcal{P}$. We will write $\mathbf{X} = \{X_{11}, ..., X_{1n_1}, X_{21}, ..., X_{2n_2}\} \in \mathcal{X}^n$, uniting the two i.d.d sample data; whose model is $P = P_1^{n_1} \cdot P_2^{n_2}$, as $X_{11}, ..., X_{1n_1}$ are i.i.d from $P_1$, $X_{21}, ..., X_{2n_2}$ are i.d.d. from $P_2$, and the models $P_1$ and $P_2$

present independence; and where $n = n_1 + n_2$. From now on we will write $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}) = \{X(i), i = 1, ..., n;\ n_1, n_2\}$ to lighten notation, where the first $n_1$ data come from the first sample, and $X_{n_1+1}, ..., X_n$ belong to the second one. Finally, we will note $\mathbf{X}^* = \{X(u_i^*), i = 1, ..., n;\ n_1, n_2\}$ the resulting data sample from applying the permutation $u^* = (u_1^*, ..., u_n^*)$ to $\mathbf{X}$.

Remark that when permuting mixing observations between both populations is permitted. In fact, this would be the key point of permutation tests. If $P_1$ and $P_2$ were identically distributed, this exchange should not be of relevance. However, if it is not the case, that is, $P_1 \neq P_2$, changing data would be relevant when measuring statistics on the permuted data set.

We develop the theoretical body in this project standing on testing problems for stochastic dominance generated by some possible fixed effects as a result of the application of a treatment. Stochastic dominance implies one sided alternatives. Two sided alternatives would be discussed later. Under this assumption, our null hypothesis is

$$H_0 : \{P_1 = P_2\}$$

whereas the alternative hypothesis is

$$H_1 : \{X_1 + \delta_1 > X_2 + \delta_2\}$$

with $\delta_1$ and $\delta_2$ the respective fixed effects in each population.

In case we had quantitative variables, the null hypothesis could equivalently be stated as

$$H_0 : \{F_1(t) = F_2(t)\ \forall\ t \in \mathcal{R}\}$$

and the alternative as

$$H_1 : \{F_1(t) \leq F_2(t)\ \forall\ t \in \mathcal{R}\}$$

where the strict inequality is given in a set of non null probability for both distributions. It is important to underline that under the null hypothesis data are

exchangeable. Although it is an assumption, it naturally stems from the fact that we should not be able to distinguish whether data come from the first or from the second sample as both populations behave equally; and that under the alternative, due to stochastic dominance, it is obvious that CDFs do not cross. Without loss of generality, we can assume that $\delta_1 = \delta > 0$ and $P[\delta_2 = 0] = 1$. This might be interpreted as the first sample receiving a real treatment, where we want to test whether the treatment leads or not to some effects, while the second sample is under a placebo effect. Now, our hypotheses can be written as

$$H_0 : \{\delta = 0\} \quad \text{vs.} \quad H_1 : \{\delta > 0\}$$

This alternative implies that $\delta_i > 0$ for at least one individual of the first sample. We will use $\mathbf{X}(\delta) = \{X_{11} + \delta_1, ..., X_{1n_1} + \delta_{n_1}, X_{21}, ..., X_{2n_2}\}$ to note the data set under the alternative, whereas $\mathbf{X}(0)$ denotes the data set in $H_0$ (with no effects produced).

It is also important to remark that the alternative hypothesis $H_1 : \{\delta > 0\}$ does not imply $\delta_i > 0 \ \forall \ i = 1, .., n_1$, so the treatment may have some effects on some subjects of the first simple whereas it might not produce any effects on others.

## 2.1 Conditional Aspects of Permutation Tests

We now introduce the first concept related to permutation tests: the conditional reference space.

**Definition 2.1.1.** *The conditional reference space is the set of points of the sample space $\mathcal{X}$ which carry the same information as $\mathbf{X}$ in terms of the likelihood function. It is noted $\mathcal{X}_{|\mathbf{X}}$*

According to the definition, this space contains the points $\mathbf{X}^*$ such that the ratio $f_P(\mathbf{X})/f_P(\mathbf{X}^*)$ does not depend on $P$. Under the null hypothesis the density function $f_P(\mathbf{X}) = \prod_{j,i} P(X_{ji})$ is assumed to be exchangeable in its arguments as

$f_P(\mathbf{X}) = f_P(\mathbf{X}^*)$ and thus all permutations $\mathbf{X}^*$ of $\mathbf{X}$ are in $\mathcal{X}_{|\mathbf{X}}$. That is, the conditional reference space can be expressed as $\mathcal{X}_{|\mathbf{X}} = \{\bigcup_{u^*} X(u_i^*), i = 1, ..., n\}$, the set of all possible permutations of the observed data set $\mathbf{X}$. One important consequence is that every element $\mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}$ is a set of sufficient statistics for the distribution $P$ in $H_0$.

Another consequence is that the sample space $\mathcal{X}$ is partitioned into orbits $\mathcal{X}_{|\mathbf{X}}$ so that any point $\mathbf{X} \in \mathcal{X}$ belongs to one and only one orbit. We then have two disjunctive cases: if $\mathbf{X}_1 \in \mathcal{X}_{|\mathbf{X}_2}$ then the orbits are equal, $\mathcal{X}_{|\mathbf{X}_1} = \mathcal{X}_{|\mathbf{X}_2}$, whereas $\mathbf{X}_2 \notin \mathcal{X}_{|\mathbf{X}_1}$ instantly implies that orbits of $\mathbf{X}_1$ and $\mathbf{X}_2$ do not intersect, that is, $\mathcal{X}_{|\mathbf{X}_1} \bigcap \mathcal{X}_{|\mathbf{X}_2} = \emptyset$.

These consequences allow us to think about permutations as one element when working in the permutation approach, which is naturally consistent. We would not difference between points in the same orbit, but between orbits. Attending to this and enhancing the idea, the sample space could be written as $\mathcal{X} = \bigcup \mathcal{X}/\mathcal{X}_{|\mathbf{X}}$ where $\mathcal{X}/\mathcal{X}_{|\mathbf{X}}$ denotes the partition of the sample space into orbits (the sample space is perceived as a quotient space, which embodies the notion of conditional reference spaces acting as units). Conditional reference spaces $\mathcal{X}_{|\mathbf{X}}$ are also known as *permutation sample spaces*.

The next aspect related to conditional procedures is conditioning with respect to a set of sufficient statistics in $H_0$. When we condition with respect to $\mathcal{X}_{|\mathbf{X}}$ the null conditional probability of an event $A \in \mathcal{A}$ given $\mathcal{X}_{|\mathbf{X}}$ is independent of $P$ due to its sufficiency. We then have that the permutation distribution induced by any statistic $T : \mathcal{X} \to \mathbb{R}$, noted $F_T(t|\mathcal{X}_{|\mathbf{X}}) = F_{T^*}(t) = \mathbb{P}[T^* \leq t|\mathcal{X}_{|\mathbf{X}}]$ does not depend of $P$, so it is $P$-invariant. This is a great advantage as any conditional inference is free of any distribution belonging to nonparametric families. Furthermore, we can define and give a closed form of the permutation probability of every $A \in \mathcal{A}$ for finite sample sizes thanks to the finiteness of points in $\mathcal{X}_{|\mathbf{X}}$, whose cardinal number is $M = \sum_{\mathcal{X}_{|\mathbf{X}}} \mathbb{I}(X^* \in \mathcal{X}_{|\mathbf{X}}) < \infty$, and to the independence on $P$:

$$\mathbb{P}[\mathbf{X}^* \in A|\mathcal{X}_{|\mathbf{X}}] = \frac{\sum_{\mathbf{X}^* \in A} f_P(\mathbf{X}^*)}{\sum_{\mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}} f_P(\mathbf{X}^*)} = \frac{\sum_{\mathcal{X}_{|\mathbf{X}}} \mathbb{I}(\mathbf{X}^* \in A)}{M}$$

because for every permutation $\mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}$ we have $f_P(\mathbf{X}) = f_P(\mathbf{X}^*)$. We can thus define the *permutation measurable space* $(\mathcal{X}_{|\mathbf{X}}, \mathcal{A}_{|\mathbf{X}})$, where $A_{|\mathbf{X}} = \mathcal{A} \bigcap \mathcal{X}_{|\mathbf{X}}$ (the projection of events on the permutation sample space).

Attending to the permutation probability introduced before, the following proposition is stated.

**Proposition 2.1.1.** *Let us assume that* $\sum_{\mathcal{X}_{|\mathbf{X}}} \mathbb{I}(\mathbf{X}^* = \mathbf{x}) = 1$ *if* $\mathbf{x} \in \mathcal{X}_{|\mathbf{X}}$ *and* 0 *elsewhere. Then permutations* $\mathbf{X}^*$ *are equally likely in* $H_0$*:*

$$\mathbb{P}[\mathbf{X} = \mathbf{x}|\mathcal{X}_{|\mathbf{X}}] = \begin{cases} 1/M & \text{if } \mathbf{x} \in \mathcal{X}_{|\mathbf{X}} \\ 0 & \text{if } \mathbf{x} \notin \mathcal{X}_{|\mathbf{X}} \end{cases}$$

*and the elements of the same orbit* $\mathcal{X}_{|\mathbf{X}}$ *are conditionally uniformly distributed over it.*

This proposition allows us to conclude that under the null hypothesis, the permutation distribution $\mathbb{P}[\mathbf{X}^* = \mathbf{x}|\mathcal{X}_{|\mathbf{X}}]$ just depends on the resulting data set $\mathbf{X}$. The data set $\mathbf{X}$ can be interpreted as the $n$-dimensional parameter for the permutation cumulative distribution function $F_{T^*}$.

Here we have an important difference in the alternative. A set of sufficient statistics is $(\mathbf{X_1}, \mathbf{X_2})$ being the data exchangeable within but not between samples, and so the observed data $\mathbf{X}$ is not uniformly distributed over $\mathcal{X}_{|\mathbf{X}}$ conditionally.

We now introduce the empirical probability measure (EPM) and its analogous in case of quantitative variables, the empirical distribution function (EDF).

**Definition 2.1.2.** *Let* $\mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}$ *be a permutation of the observed data set* $\mathbf{X}$*. The empirical probability measure of an event* $A \in \mathcal{A}$ *is defined as:*

$$\hat{P}_{\mathbf{X}^*}(A) = \sum_{i \leq n} \mathbb{I}(X_i^* \in A)/n = \sum_{i \leq n} \mathbb{I}(X_i \in A)/n = \hat{P}_{\mathbf{X}}(A) \tag{2.1}$$

*which is a permutation invariant function. In case we have quantitative variables, we can also definde the empirical distribution function (EDF) as:*

$$\hat{F}_{\mathbf{X}^*}(t) = \sum_{i \leq n} \mathbb{I}(X_i^* \leq t)/n = \sum_{i \leq n} \mathbb{I}(X_i \leq t)/n = \hat{F}_{\mathbf{X}}(t) \tag{2.2}$$

*which is permutationally invariant as well .*

An interesting result related to these concepts is stated. Although the EPM is the general version of the empirical function, we will deal with quantitative variables, so we state it just for the EDF.

**Proposition 2.1.2.** *Let $\mathbf{X} \in \mathcal{X}$. Then $\hat{F}_{\mathbf{X}}$ is a permutation invariant function which characterises $\mathcal{X}_{|\mathbf{X}}$. Thus, the conditional sample space can be defined as the set of points in $\mathcal{X}$ which share the same EDF.*

This proposition is consequence of the partition of $\mathcal{X}$ conformed by the different orbits $\mathcal{X}_{|\mathbf{X}}$.

A consequence of this result is that the EDF is a sufficient statistical function for $P$ in the null hypothesis. Hence, conditioning on the conditional reference space is equivalent to doing so on $\hat{F}_{\mathbf{X}}$. Another consequence is that for any statistic $T$,

$$\mathbb{P}[T(\mathbf{X}^*) \leq t | \mathcal{X}_{|\mathbf{X}}] = \mathbb{P}[T(\mathbf{X}^*) \leq t | \hat{F}_{\mathbf{X}}] = \mathbb{P}[T(\mathbf{X}^*) \leq t | \hat{F}_{\mathbf{X}^*}]$$

for all $t \in \mathbb{R}$ and all permutation $\mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}$. According to this equivalence, the permutation null distribution of $T$ can be interpreted as a process of without replacement random experiment from a uniformly distributed population whose distribution is the EDF. This interpretation is clearly different to that of bootstrap methods, where the selection is made with replacement.

An asymptotic result with mild assumptions is obtained for the null permutation distribution of $T$. If we have sufficiently large $n_1$ and $n_2$ then the null permutation distribution $\mathbb{P}[T(\mathbf{X}) \leq t | \hat{F}_{\mathbf{X}}]$ approximates its unconditional counterpart $\mathbb{P}[T(\mathbf{X}) \leq t]$

## 2.2 What is a Permutation Test

Let us suppose a test statistic $T$ for which, without loss of generality, large values of $T$ are evidence against $H_0$. Of course, if it is not the case, we can consider a suitable transformation of the test to be in this situation.

**Definition 2.2.1.** *On the previous conditions, the permutation support induced by (T,***X***) is defined as the set* $\mathcal{T}_{\mathbf{X}} = \{T(\mathbf{X}^*) : \mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}\}$

This set is essentially conformed by all the possible values which the statistic $T$ assumes by permuting the data set $\mathbf{X}$.

We now assume the null hypothesis as true. In light of 2.1.1, $\mathbf{X}^*$ is uniformly distributed over the conditional space. Let us put $M$ as the cardinal number of $\mathcal{X}_{|\mathbf{X}^*}$, and let us consider $T^*_{(1)} \leq T^*_{(2)} \leq \cdots \leq T^*_{(M)}$, the members of $\mathcal{T}_{\mathbf{X}}$ in a non-decreasing order. If we now fix a value $\alpha \in (0,1)$, $T_\alpha(\mathbf{X}) = T_\alpha = T^*_{(M_\alpha)}$ denotes the permutation critical value related to the statistic $T$ and the data set $\mathbf{X}$, where $M_\alpha = \sum_{\mathcal{X}_{|\mathbf{X}}} \mathbb{I}[T(\mathbf{X}^*) < T_\alpha]$ is the number of values assumed by $T$ in the permutation support lower than $T_\alpha$. Note that the behaviour of the critical value not only depends on the data set observed $\mathbf{X}$ but also on $\mathcal{X}_{|\mathbf{X}}$, as $M_\alpha$ would vary depending on the values assumed by $T$ in the respective permutation sample space of $\mathbf{X}$. That is, if we get another observed data set $\mathbf{X}^*$ verifying $\mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}$, the critical value satisfies $T_\alpha = T_\alpha(\mathbf{X}) = T_\alpha(\mathbf{X}^*)$ as the two orbits coincide and so do the permutation supports. The critical value is orbit-invariant, so for each $\alpha \in (0,1)$, $T_\alpha$ is a fixed value in $\mathcal{T}_{\mathbf{X}}$ which varies as $\mathbf{X}$ varies in $\mathcal{X}$ just in case we jump into a different orbit.

### 2.2.1 Constructing Permutation Tests

We now move on to construct the tests based on permutational aspects. As it is widely known, there are two kind of tests: randomized and non-randomized tests.

**Randomized Permutation Tests**

We first focus on randomized tests. Let us note $\phi_R$ the test associated with the pair $(T, \mathbf{X})$:

$$\phi_R = \begin{cases} 1 & \text{if } T^0 > T_\alpha \\ \gamma & \text{if } T^0 = T_\alpha \\ 0 & \text{if } T^0 < T_\alpha \end{cases}$$

where $T^0 = T(\mathbf{X})$ is the value of the test statistic $T$ examined on the observed data set $\mathbf{X}$ and $\gamma$ is a probability given by

$$\gamma = \frac{\alpha - \mathbb{P}[T^0 > T_\alpha | \mathcal{X}_{|\mathbf{X}}]}{\mathbb{P}[T^0 = T_\alpha | \mathcal{X}_{|\mathbf{X}}]}$$

The test itself is not complete. When $T^0 = T_\alpha$, we should define an auxiliary rule of decision based on a independent experiment. This can be made by running a variable $U \sim \mathcal{U}(0,1)$ and rejecting $H_0$ in case $U \geq \gamma$.

Let us write the conditional expectation in $H_0$:

$$\begin{aligned} \mathbb{E}[\phi_R(\mathbf{X}) | \mathcal{X}_{|\mathbf{X}}] &= \mathbb{P}[T^0 > T_\alpha | \mathcal{X}_{|\mathbf{X}}] + \gamma \cdot \mathbb{P}[T^0 = T_\alpha | \mathcal{X}_{|\mathbf{X}}] \\ &= \mathbb{P}[T^0 > T_\alpha | \mathcal{X}_{|\mathbf{X}}] + \frac{[\alpha - \mathbb{P}[T^0 > T_\alpha | \mathcal{X}_{|\mathbf{X}}]] \cdot \mathbb{P}[T^0 = T_\alpha | \mathcal{X}_{|\mathbf{X}}]}{\mathbb{P}[T^0 = T_\alpha | \mathcal{X}_{|\mathbf{X}}]} \\ &= \alpha \end{aligned}$$

what shows that randomized permutation tests are exact tests for all $\mathbf{X} \in \mathcal{X}$ and any $\alpha \in (0,1)$. Indeed, they verify a stronger condition, the *uniform similarity* property:

**Proposition 2.2.1.** *Under exchangeability of data* $\mathbf{X}$*, the conditional rejection probability of a randomized test* $\phi_R$ *is* $\mathbf{X}$*-*$P$*-invariant in* $H_0$ *for all* $P \in \mathcal{P}$ *and all* $\mathbf{X} \in \mathcal{X}$*.*

The proof can be found in [5]. It can be appreciated once again how robust permutation test are in a nonparametric framework, freeing not only the distribution but also the test and its critical region from any underlying unknown population and observed data.

**Non-Randomized Permutation Tests**

Let us study now the other type of tests, non-randomized ones. These tests have the form:

$$\phi = \begin{cases} 1 & \text{if } \ T^0 \geq T_\alpha \\ 0 & \text{if } \ T^0 < T_\alpha \end{cases}$$

In contrast with the randomized version of permutation tests, the conditional expectation in $H_0$ (and coinciding with the type I error rate in non-randomized tests) is:

$$\mathbb{E}[\phi(\mathbf{X})|\mathcal{X}_{\mathbf{X}}] = \mathbb{P}[T^0 \geq T_\alpha|\mathcal{X}_{|\mathbf{X}}] = \sum_{\mathcal{X}_{|\mathbf{X}}} \mathbb{I}[T(\mathbf{X}^*) \geq T_\alpha]/M = \alpha_a \geq \alpha$$

Considering the significance level function $L_{\mathbf{X}}(t) = \mathbb{P}[T^* \geq t|\mathcal{X}_{|\mathbf{X}}]$, and given a pair $(T, \mathbf{X})$, the possible $\alpha$-values for the test, called *attainable $\alpha$-values*, are those where the significance level function is discontinuous, that is, when $l_{\mathbf{X}}(t) > 0$, with $l_{\mathbf{X}}(t)$ the derivative of $L_{\mathbf{X}}(t)$. Thus, the set $\Lambda_{\mathbf{X}} = \{L_{\mathbf{X}}(t) : l_{\mathbf{X}}(t) > 0\}$ (step points of the significance level function) contains these attainable $\alpha$-values. This is always a discrete set which depends on $T$, $\mathbf{X}$ and $n$. Because of this, non-randomized permutation tests do not admit all values of type I error rate in practice.

Let us study the behaviour of $\Lambda_{\mathbf{X}}$. In case that it exists a constant $c \geq 1$ verifying that for all values in the permutation support $\sum_{\mathcal{X}_{|\mathbf{X}}} \mathbb{I}(T^* = t) = c$ (i.e., $\Lambda_{\mathbf{X}}$ does not contain multiple points or these points have constant multiplicity), $\Lambda_{\mathbf{X}}$ becomes $\mathbf{X}$-invariant, though it does not become free of its dependence on $n$. In this case, the set $\Lambda_{\mathbf{X}}$ has the form $\Lambda_{\mathbf{X}} = \{mc/M,\ m = 1, ..., M/c\}$, and $\alpha_a$-values have constant jumps of $c/M$. Due to this, non-randomized tests become conservative when we choose a desired type I error rate $\alpha_d$ which does not belong to $\Lambda_{\mathbf{X}}$ and an attainable $\alpha$-value with $\alpha_d \geq \alpha_a$. Of course, if the desired rate is in the set of attainable $\alpha$-values, exactness is obtained.

However, if $\mathcal{T}_{\mathbf{X}}$ does not verify the condition previously showed, the set $\Lambda_{\mathbf{X}}$ is $\mathbf{X}$-variant. This might lead to an apparent decrease of the power function when increasing sample size.

### 2.2.2 Analyzing the p-Value

Determining the critical values $T_\alpha$ is not a trivial task, so we resort to the $p$-value associated with $(T, \mathbf{X})$. Recall the $p$-value is the probability of the statistic $T$ assuming values as extreme or more than the obtained with the data observed $\mathbf{X}$, that is, $T(\mathbf{X})$. Attending to this, we define the $p$-value in the permutation context as

$$\lambda = \lambda_T(\mathbf{X}) = \mathbb{P}[T^* \geq T^0 | \mathcal{X}_{|\mathbf{X}}]$$

Note the $p$-value can be also expressed in terms of the significance level function as $L_{\mathbf{X}}(T^0)$.

The $p$-value is a non-increasing function of $T^0$ (it is obvious that if $T^0$ increases, the probability of obtaining bigger values than $T^0$ decreases or remains equal). It is also in bijection with the attainable $\alpha$-values of the test, verifying that if $\lambda_T > \alpha$, then $T^0 < T_\alpha$ and vice versa. This relationship solves the problem of obtaining the critical values $T_\alpha$ to determine whether to reject or not the null hypothesis. In effect, once the $p$-value has been calculated, we can exclusively attend to it to determine the decision. Hence, the test can also be expressed as

$$\phi = \begin{cases} 1 & \text{if} \quad \lambda_T(\mathbf{X}) \leq \alpha \\ 0 & \text{if} \quad \lambda_T(\mathbf{X}) > \alpha \end{cases}$$

Now the attainable $\alpha$-values play the role of critical values, and the $p$-value can be used as a test statistic to construct the hypothesis testing.

If $X$ is a continuous variable and $T$ is a regular function, these tests verify the similarity property in the *almost sure* form.

**Proposition 2.2.2.** *If $X$ is continuous and $T$ is a regular function, then the*

*attainable $\alpha$-values of $\phi$ are independent of both the data set $\mathbf{X}$ and the underlying population $P$ for almost all $\mathbf{X} \in \mathcal{X}$ with probability one with respect to $P$.*

This is given because in case $X$ is continuous, the probability of finding ties in the data set is zero. Moreover, this is an important result as the attainable $\alpha$-values are just depending on the statistic $T$, so we just need to compute the calculus to obtain them once.

However, for discrete or mixed variables, the case where ties in the data set may have positive probability, the attainable values of significance depend on $\mathbf{X}$, losing the similarity property. In spite of this lost of properties, this is just in the finite sample case, being valid for large samples as it is verified asymptotically.

Under the assumptions of the previous proposition, we could state a property about the distribution of $p$-values in the invariant set $\Lambda$.

**Proposition 2.2.3.** *If $X$ continuous and $T$ is a regular function, the $p$-values $\lambda_T(\mathbf{X})$ are uniformly distributed in $\Lambda$.*

This proposition is deeply related to the uniform distribution of the permutations over the conditional reference space (2.1.1), as the $p$-values are in one-to-one relationship with the data set.

**Corollary 2.2.1.** *Let us suppose that there are not repeated elements in $\mathcal{T}_{\mathbf{X}}$. Then under $H_0$ the elements of $\mathcal{T}_{\mathbf{X}}$ are equally likely.*

The corollary is straightforward from the uniform distribution of attainable $\alpha$-values over $\Lambda$, because if no elements are repeated in the permutation support, there is a bijection between $\Lambda$ and $\mathcal{T}(\mathbf{X})$.

### Computing the *p*-value

Properties related to $p$-values of permutation tests have been exposed. However, a problem obviously arises when obtaining the $p$-value of a test. Recall the $p$-value was defined as the probability of finding values more extreme than the

observed, that is , $\mathbb{P}[T^* \geq T^0 | \mathcal{X}_{|\mathbf{X}}]$. This implies evaluating the test statistic $T$ on all possible permutations $\mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}$. But when $n$ is too large this is obviously impractical.

We then go to a Monte Carlo algorithm to approximate the $p$-value. The algorithm is described as follows:

1. In first place, calculate the observed value of $T$, $T^0 = T[\mathbf{X}(\delta)]$.
2. Randomly permute the observed data set $\mathbf{X}^*(\delta)$ and assess $T(\mathbf{X}^*(\delta))$.
3. Repeat last step $B$ times
4. Finally, the $p$-value is estimated as $\hat{\lambda} = \sum_{i=1}^{B} \mathbb{I}[T_i^*(\mathbf{X}^*) \geq T^0]/B$, that, is, the proportion of permutations where the observed value $T^0$ is exceeded.

The set created by resampling, $\{\mathbf{X}_b^* \ b = 1, ..., B\}$, is a random sample from the permutation space, so the corresponding values of $T$, $\{T_b^*, \ b = 1, ..., B\}$ resemble the null permutation of $T$. Furthermore, based on Glivenko-Cantelli theorem, the estimation of the $p$-value $\hat{\lambda}$ converge to its real value.

Naturally, if $n$ is sufficiently small, we can exactly compute the value of the $p$-value as

$$\lambda = \sum_{\mathcal{X}_{|\mathbf{X}}} \mathbb{I}[T^* \geq T^0]/M$$

## 2.3 Statistics Permutationally Equivalent

With a view to simplify the most possible the structure of the tests itself, we introduce a concept which will allow us to use simpler statistics instead of complex data functions.

**Definition 2.3.1.** *Let us suppose $\mathbf{X} \in \mathcal{X}$ and two statistics $T_1, T_2$. $T_1$ and $T_2$ are said to be permutationally equivalent if $[T_1(\mathbf{X}^*) \leq T_1(\mathbf{X})]$ is verified if and only if $[T_2(\mathbf{X}^*) \leq T_2(\mathbf{X})]$ for all $\mathbf{X} \in \mathcal{X}$, $\mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}$. It is noted $T_1 \approx T_2$.*

Then two statistics are permutationally equivalent if the values assumed by them in the sample space respect an order notion. As this relation is verified pointwise (that is, once considered an orbit $\mathcal{X}_{|\mathbf{X}}$ this is verified for any $\mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}$), we can give an accurate description of the permutation support of two equivalent statistic.

Let us assume $(T_1, \mathbf{X})$ and $(T_2, \mathbf{X})$, with $T_1$ and $T_2$ permutationally equivalent. These two pairs would respectively induce the permutation supports $\mathcal{T}^1_{\mathbf{X}}$ and $\mathcal{T}^2_{\mathbf{X}}$. As the relationship $[T_1(\mathbf{X}^*) \leq T_1(\mathbf{X})]$ implies $[T_2(\mathbf{X}^*) \leq T_2(\mathbf{X})]$ and vice versa, it is straightforward that $[T_1(\mathbf{X}^*) \geq T_1(\mathbf{X})]$ if and only if $[T_2(\mathbf{X}^*) \geq T_2(\mathbf{X})]$. Attending to this, the respective $p$-values are $\lambda_1 = \mathbb{P}[T_1(\mathbf{X}^*) \geq T_1^0]$ and $\lambda_2 = \mathbb{P}[T_2(\mathbf{X}^*) \geq T_2^0]$. Assuming finiteness in the sample space, $\lambda_1 = \sum_{\mathcal{X}_{|\mathbf{X}}} \mathbb{I}[T_1(\mathbf{X}^*) \geq T_1^0]$ and $\lambda_2 = \sum_{\mathcal{X}_{|\mathbf{X}}} \mathbb{I}[T_2(\mathbf{X}^*) \geq T_2^0]$.

From the relationship previously exposed, it is trivially concluded that $\lambda_1 = \lambda_2$. This let us go from certain complex test using a convoluted statistic to other pretty much simpler in case we find an equivalent statistic to the first one.

In conclusion, permutation equivalence may be understood as a transformation of the permutation support which does not alter the order of its elements nor their ranks.

Recall that, if we want to simplify a test, we should be capable of finding some easier (beforehand) statistic permutationally equivalent to the one we are handling. We state some results which ease this task.

**Theorem 2.3.1.** *Let us suppose that it exists an increasing bijection between $T_1$ and $T_2$. Then $T_1 \approx T_2$ and $\mathbb{P}[T_1(\mathbf{X}^*) \leq T_1(\mathbf{X})|_{\mathcal{X}_{|\mathbf{X}}}] = \mathbb{P}[T_2(\mathbf{X}^*) \leq T_2(\mathbf{X})|_{\mathcal{X}_{|\mathbf{X}}}]$.*

*Proof.* Let us consider $\varphi : \mathbb{R} \to \mathbb{R}$ the relationship between the two statistics, with $T_2 = \varphi(T_1)$. Then we have:

$$[T_1(\mathbf{X}^*) \leq T_1(\mathbf{X})] \leftrightarrow [\varphi(T_1(\mathbf{X}^*)) \leq \varphi(T_1(\mathbf{X}))] = [T_2(\mathbf{X}^*) \leq T_2(\mathbf{X})]$$

On the other hand, as $\varphi$ is a bijection, considering $\varphi^{-1}$,

$$[T_2(\mathbf{X}^*) \leq T_2(\mathbf{X})] \leftrightarrow [\varphi^{-1}(T_2(\mathbf{X}^*)) \leq \varphi^{-1}(T_2(\mathbf{X}))] = [T_1(\mathbf{X}^*) \leq T_1(\mathbf{X})]$$

□

This theorem fits with the notion we explained in advance. $\mathcal{T}^2_{\mathbf{X}}$ is just a translation of $\mathcal{T}^1_{\mathbf{X}}$. A similar result is given when there is a decreasing relationship.

**Corollary 2.3.1.** *Let us suppose that it exists a decreasing bijection between $T_1$ and $T_2$. Then $T_1 \approx T_2$ in the sense that $[T_1(\mathbf{X}^*) \leq T_1(\mathbf{X})]$ if and only if $[T_2(\mathbf{X}^*) \geq T_2(\mathbf{X})]$.*

In this case, we are mirroring the permutation support with respect to $T^0$. It is an inversion which also invert ranks, but under a bijection. That is, let us assume again finiteness in the sample space $\mathcal{X}_{|\mathbf{X}}$, say $M$ is its cardinal. Then the th-element of $\mathcal{T}^1_{\mathbf{X}}$ would be the $(M - th + 1)$-element of $\mathcal{T}^2_{\mathbf{X}}$.

Because of this, $p$-values do not behave as before. Now the relationship between both critical values is $\lambda_2 = 1 - \lambda_1$. We can then have a case where $H_0$ is rejected using $T_1$ but we would clearly not reject it when using $T_2$. Though this could appear a contradiction, recall we are interpreting large values of the statistic as evidence against the null hypothesis, and large values of $T_1$ mean little values of $T_2$. We just need to reconsider the critical region of our test if $T_2$ is desired.

It is also worth noting that the permutationally-equivalent relation is in fact an equivalence relation. This could help us to find a not-so-complex but neither not-so-simple statistic $T_2$ as an intermediate equivalence between $T_3$ and $T_1$, where $T_3$ is intended to be a difficult statistic to compute whereas $T_1$ is a pretty appealing one.

We end the section with the most practical result about this concept.

**Proposition 2.3.1.** *If $T_1$ and $T_2$ are one-to-one related when restricted to the data set $\mathbf{X}$, then $T_1 \approx T_2$.*

The proof of this is straightforward from the proof of the previous theorem. This result is frequently used when proving the equivalence of test statistics.

Let us illustrate the usefulness of this proposition. Consider $T_1^* = \bar{X}_1^* - \bar{X}_2^*$, and $T_2^* = \sum_i X_{2i}^*$. Using that $\sum_{ji} X_{ji}^* = \sum_{ji} X_{ji} = KX$, we can then write

$$T_1^* = \frac{1}{n_1}\sum_i X_{1i}^* - \frac{1}{n_2}[KX - \sum_i X_{1i}^*] = [\frac{n}{n_1 n_2}]\sum_i X_{1i}^* - \frac{1}{n_2}KX \approx \sum_i X_{1i}^*$$

as $\frac{n}{n_1 n_2}$ and $\frac{1}{n_2}KX$ are permutationally invariant. Thus $T_1$ and $T_2$ are one-to-one related, and we can conclude they are permutationally equivalent. Observe that we have passed from the difference of both sample means to the sum of the individuals of just one sample.

## 2.4 The Test Statistic and the Critical Region

The first discussion to approach in this section is the preference for statistics based on divergence.

We have developed the theory standing on a test statistic based on divergence. However, no rational reason has been explained to justify this selection. Let us present the arguments which would vindicate that.

In first place, following [5], we have two important results related to the choice of an optimal permutation tests.

**Lemma 2.4.1.** *Let $\psi$ be any test of invariance hypothesis testing whose size is less or equal than $\alpha$. Then there exists a permutation test $\phi$ verifying*

$$\int \phi \geq \int \psi$$

*for any distribution $P \in \mathcal{P}$*

This lemma shows that permutation tests are better as any arbitrary test when testing invariance.

The second result is related to how choosing the most powerful tests when testing invariance hypothesis again.

**Lemma 2.4.2.** *Let $H_0$ be an invariance hypothesis. Let $f_{H_1}$ be the density of $P$ in $H_1$, and $\mathbf{X} \in \mathcal{X}$. We order the points of the sample space in terms of $f_{H_1}$ resulting in $f_{H_1}(\mathbf{X}^*_{(1)}) \geq f_{H_1}(\mathbf{X}^*_{(2)}) \geq \cdots \geq f_{H_1}(\mathbf{X}^*_{(M)})$. Then the most powerful test of size $\alpha$ is*

$$\phi_R = \begin{cases} 1 & \text{if} \quad f_{H_1} > f_\alpha \\ \gamma & \text{if} \quad f_{H_1} = f_\alpha \\ 0 & \text{if} \quad f_{H_1} < f_\alpha \end{cases}$$

*where for any fixed $\alpha \in (0,1)$, $f_\alpha$ is the critical value, $M_\alpha$ is the number of points which lay within the critical region and*

$$\gamma = \frac{\alpha - \mathbb{P}[f(\mathbf{X}) > f_\alpha(\mathbf{X})]}{\mathbb{P}[f(\mathbf{X})]} = f_\alpha(\mathbf{X})$$

With this result, along with the previous one, we can think that we do have solved the problem of choosing a suitable test. However, the disadvantage of this result is clear: we need to know the analytical form of the distribution $P$, and since we are working in a nonparametric framework, this results impracticable.

We need to make a different approach to justify the selection of our test statistics.

In fact, this selection is mainly heuristic. As $P$ is assumed to be unknown in the non-parametric framework, we can not make use of the previous results to establish a best statistic. In spite of this, there is an analogy in permutation tests and the parametric solution when we are working with large sample sizes. As stated in [6],

**Proposition 2.4.1.** *Let $T$ be a best statistic for the parametric family $\mathcal{P}$, and assume that the unconditional critical region does not depend on any specific*

*alternative. Then its permutation counterpart is asymptotically equivalent (and the permutation version of $T$ is asymptotically a best statistic for $\mathcal{P}$).*

Though it is a powerful result as we are able to determine a best (asymptotically) permutation statistic for the family of distributions, this is a very particular case where $\mathcal{P}$ is parametric. Since we have introduced permutation tests to work in a non-parametric framework, following the nature of the previous results to determine an optimal tests, this result is very restrictive.

In light of these previous results, the choice of divergence turns out to be pretty heuristic. Results claiming optimal properties for tests are so rigid that the hypotheses are scarcely verified. What is more, permutation tests are presented to cover nonparametric situations, so meeting the requirements to apply some optimality results seems impractical.

The other question to analyse is why large values are evidence against $H_0$, a fact which we have arbitrarily imposed.

As well as the selection of $T$, it is mainly heuristic, though we have some reasons backing up this fact. The first one is that this choice is supported on a suitable transformation which provides equivalence of tests. We can transform our test into another one in which the alternative distribution stochastically dominates the null distribution. What is more, we would prove in 3.1 that the null distribution of a statistic $T$ is dominated by its alternative distribution in one-sided critical regions. Thanks to this, power functions of tests would be monotonic increasingly with respect to the effects we are testing (more on this later in 3.2).

# Chapter 3

# More Properties of Permutation Tests

We continue on the same conditions previously stated, that is, we assume a one-dimensional variable $X$ taking values on sample space $\mathcal{X}$ with probability distribution $P$, and that in $H_1$ the distribution of $X_1$ is shifted by a constant quantity $\delta$ with respect to that of $X_2$. We are in stochastic dominance scenario, with $F_1(x) \leq F_2(x) \ \ \forall\ x \in \mathbb{R}$.

Moreover, we continue assuming that the constant effect $\delta$ is positive. We would write the variable as $X(\hat{\delta}) = \{X_{ji} = \hat{\delta}_{ji} + Z_{ji},\ i = 1, ..., n_j,\ j = 1, 2\}$ in the alternative from now on, where $Z_{ji}$ are random deviates exchangeable which follow an unknown distribution $P$. We also assume $\hat{\delta}_{2i} = 0$ for $i = 1, ..., n_2$ as we previously did, so there are no effects on the second group (recall the placebo effect). Under this assumptions, we can represent the data set as $\mathbf{X}(\delta) = (\mathbf{Z_1} + \delta, \mathbf{Z_2})$, with $\delta = (\delta_{11}, ..., \delta_{1n_1})$. This notation emphasizes the exclusively activity of the treatment on the first sample, with no effects on the second one. It is important to remark that $\delta = (\delta_{11}, ..., \delta_{1n_1})$ is such that $\delta_{1i} > 0$ for at least one $i = 1, .., n_1$, and thus the null and the alternative state different situations.

In these conditions, out hypotheses would be

$$H_0 : \{\delta = 0\} \text{ vs. } H_1 : \{\delta > 0\}$$

## 3.1 Unbiasedness

Firstly let us introduce the concept of unbiasedness for a test statistic. We recover this definition from [5].

**Definition 3.1.1.** *Let $\phi$ be a test. $\phi$ is said to be unbiased if it verifies that*

$$\beta_\phi(\theta_0) \leq \alpha \leq \beta_\phi(\theta_1)$$

*for every size $\alpha \in (0, 1)$ and every specific alternative in $H_1$, where $\beta_\phi(\theta)$ is the power function of the test.*

The notion of test unbiasedness is related to a comparison between the rejection behaviour of the test under the null and the alternative hypothesis respectively. It essentially states that the probability of rejecting is always higher when the alternative hypothesis is true than when the null hypothesis is true. Due to our determination of the rejection region for permutation tests (which is of the form $[T_\alpha, \infty)$ as large values of $T$ were evidence against the null hypothesis), this definition can be equivalently stated as

$$\mathbb{P}[T(\mathbf{X}) \geq T_\alpha | H_0] \leq \alpha \leq \mathbb{P}[T(\mathbf{X}(\delta)) \geq T_\alpha | H_1]$$

Owing to our test $\phi$ is based on the test statistic $T$, we will say that $T$ *possesses unbiasedness* instead of the test itself. We should not confuse this with the notion of unbiasedness related to the expected value of $T$. However, as the latter concept is defined in the parametric framework, there would be no chance of provoking confusion since we are working in a nonparametric scenario.

We have a sufficient condition for determining the unbiasedness of $T$.

**Proposition 3.1.1.** *Let $T$ be a test statistic. If the null distribution of $T$, $T_{H_0}$, is dominated by every distribution from $H_1, T_{H_1}$, then the test $\phi$ based on $T$ is unbiased.*

*Proof.* The proof is straightforward from the definition of stochastic dominance. Let $T_{H_1}$ dominate $T_{H_0}$ for any alternative, and let us fix a type I error rate $\alpha \in (0, 1)$. Then we have

$$\mathbb{P}[T_{H_0} \geq T_\alpha] \leq \alpha \leq \mathbb{P}[T_{H_1} \geq T_\alpha]$$

□

However, this concept has been defined in an unconditional scenery. Since we are developing permutation tests from a conditional standpoint, we introduce the conditional counterpart of this notion.

**Definition 3.1.2.** *Let $T$ be a test statistic. $T$ is said to be conditionally or permutationally unbiased if the p-values verify*

$$\mathbb{P}[\lambda_T(\mathbf{X}(\delta)) \leq \alpha_a | \mathcal{X}_{|\mathbf{X}(\delta)}] \geq \mathbb{P}[\lambda_T(\mathbf{X}(0)) \leq \alpha_a | \mathcal{X}_{|\mathbf{X}(0)}] = \alpha_a$$

*for every $\mathbf{X} \in \mathcal{X}$ and any $\delta \in H_1$, where $\mathbf{X}(\delta) = (\mathbf{Z_1} + \delta, \mathbf{Z_2})$, $\mathbf{X}(0) = (\mathbf{Z_1}, \mathbf{Z_2})$, and $\alpha_a$ is any attainable $\alpha$-value.*

Similarly to the classic definition of test unbiasedness, the interpretation of the definition is basically that rejecting is more likely when the alternative hypothesis is true than doing so when it is the null hypothesis the one which is true.

In a similar way to the unconditional case, we can introduce a sufficient condition to obtain conditional unbiasedness.

**Proposition 3.1.2.** *Let $T$ be a test statistic. If $T$ is such that*

$$\begin{aligned}\lambda_T(\mathbf{X}(\delta)) &= \mathbb{P}[T(\mathbf{X}^*(\delta)) \geq T^0(\delta) | \mathcal{X}_{|\mathbf{X}(\delta)}] \\ &\leq \mathbb{P}[T(\mathbf{X}^*(0)) \geq T^0(0) | \mathcal{X}_{|\mathbf{X}(0)}] = \lambda_T(\mathbf{X}(0))\end{aligned}$$

*then $T$ is conditionally unbiased.*

We just have to use the monotony of the distribution function to get the result.

Note that this sufficient condition just involves the respective $p$-values under both hypotheses. If we have smaller $p$-values in the alternative hypothesis, it would be more likely to reject using the same levels of significance than in the null.

Of course, the reader could instinctively think that conditional unbiasedness is a more restrictive notion than the traditional one. This is in fact true, as we state in this proposition.

**Proposition 3.1.3.** *If $T$ is conditionally unbiased, then it is also unconditionally unbiased.*

This result naturally stems from the definition of conditional probability, as the unconditional rejection probability of the permutation test based on $T$ is

$$\int_{\mathcal{X}} \mathbb{P}[\lambda_T(\mathbf{X}(\delta)) \leq \alpha | \mathcal{X}_{|\mathbf{X}(\delta)}] f_P(\mathbf{X}(\delta)) \geq \alpha$$

because under the assumption that $T$ is conditionally unbiased, it is verified that $\mathbb{P}[\lambda_T(\mathbf{X}(\delta)) \leq \alpha | \mathcal{X}_{|\mathcal{X}(\delta)}] \geq \alpha$ and $\int_{\mathcal{X}} f_P = 1$ as it is a density function.

The converse, however, is not true. This reinforces the idea of conditional unbiasedness being a more stringent concept than the unconditional one.

### 3.1.1 Characteristics of Conditional Unbiasedness

Let us dig into the concept previously introduced. Given an observed data set $\mathbf{X}$, the two observed values of $T$ are $T^0(0) = T^0(\mathbf{X}(0))$ (under $H_0$) and $T^0(\delta) = T^0(\mathbf{X}(\delta))$ (under $H_1$). Consider a permutation of the observed data set, resulting into two different values of $T$, $T^*(0) = T(\mathbf{X}^*(0))$ and $T^*(\delta) = T(\mathbf{X}^*(\delta))$ respectively. We now define the increment of $T$ due to the fixed effects $\delta$ as

$\Delta_T(\mathbf{X}(\delta)) = T(\mathbf{X}(\delta)) - T(\mathbf{X}(0))$, and the difference of values resulted from permuting as $\Delta_T(\mathbf{X}^*(\delta)) = T(\mathbf{X}^*(\delta)) - T(\mathbf{X}^*(0))$. These increments allow us to compare the behaviour of $T$ under both hypotheses.

Before continuing, we should make some assumptions on the statistic $T$ we are working with. The first one is that $T$ will have the form

$$T = S_1(\mathbf{X}_1) - S_2(\mathbf{X}_2) \tag{3.1}$$

We assume $S_j$ to be symmetric functions, which implies invariance under rearrangement of data ($S_j(\mathbf{X}_j) = S_j(\mathbf{X}_j^*)$, with $\mathbf{X}_j^*$ any rearrangement of the data set $\mathbf{X}_j$). It is typical to consider $S_j$ some kind of sample mean or median. The second assumption will be the non-decreasing monotony of $S_j$, so $S_j(\mathbf{X} + \mathbf{X}') \geq S_j(\mathbf{X})$ for any data set $\mathbf{X}$ and any non-negative $\mathbf{X}' \geq 0$. This second assumption enhances the idea of large values of $T$ being evidence against $H_0$ as effects would be such that $\delta > 0$. A final observation is that $T$ is non-increasing in its second $n_2$ arguments.

Back on the increments, we have that $\Delta_T(\mathbf{X}(\delta)) = T(\mathbf{X}(\delta)) - T(\mathbf{X}(0)) \geq 0$, because $T$ is not decreasing in its first $n_1$ arguments along with the observation that it is non-increasing in its second $n_2$ arguments.

We can also stablish a one-to-one pointwise relationship between the conditional permutation spaces of $H_0$ and $H_1$. If we fix a permutation $\mathbf{u}^* = (u_1^*, ..., u_n^*)$ of $\mathbf{u} = (u_1, ..., u_n)$, then $\mathbf{X}^*(0) = \{\mathbf{Z}(u_i^*),\ i = 1, ..., n;\ n_1, n_2\}$ gives a unique point in the alternative sample space, which is $\mathbf{X}^*(\delta) = \{\mathbf{Z}(u_i^*) + \delta(u_i^*),\ i = 1, ..., n;\ n_1, n_2\}$. It is also true the other way.

Therefore, $p$-values in $H_0$ are

$$\lambda_T(\mathbf{X}(0)) = \mathbb{P}[T(\mathbf{X}^*(0)) \geq T^0(0) | \mathcal{X}_{|\mathbf{X}(0)}]$$

whereas in $H_1$, they can be expressed as

$$\begin{aligned}\lambda_T(\mathbf{X}(\delta)) &= \mathbb{P}[T(\mathbf{X}^*(\delta)) \geq T^0(\delta)|\mathcal{X}_{|\mathbf{X}(\delta)}] \\ &= \mathbb{P}[T(\mathbf{X}^*(0)) + \Delta_T(\mathbf{X}^*(\delta)) - \Delta_T(\mathbf{X}(\delta)) \geq T(\mathbf{X}(0))|\mathcal{X}_{|\mathbf{X}(0)}]\end{aligned}$$

Two key points here. The first has been continuously repeated and it is that the effects $\delta$ are only active on the first sample, i.e., on the first $n_1$ observations of $\mathbf{X}$. The second one is that, due to the previous assumptions, the difference in any data permutation $\Delta(\mathbf{X}^*(\delta))$ tends to be smaller than $\Delta(\mathbf{X}(\delta))$ because some effects are exchanged between the first and the second sample and $T$ is assumed to be non-decreasing on its first $n_1$ arguments. Let us clarify this.

The data set originally is $(\mathbf{Z}_1 + \delta, \mathbf{Z}_2) = (Z_{11} + \delta_{11}, ..., Z_{1n_1} + \delta_{1n_1}, Z_{21}, ..., Z_{2n_2})$ in the alternative. When we consider a permutation $\mathbf{u}^* = (u_1^*, ..., u_n^*)$, the new data set is $(\mathbf{Z}_1^* + \delta^*, \mathbf{Z}_2^*) = \mathbf{u}^*(Z_{11} + \delta_1 1, ..., Z_{1n_1} + \delta_{1n_1}, Z_{21}, ..., Z_{2n_2}) = (u_{11}^*(Z_{11}) + u_{11}^*(\delta_{11}), ..., u_{1n_1}^*(Z_{1n_1}) + u_{1n_1}^*(\delta_{1n_1}), u_{21}^*(Z_{21}), ..., u_{2n_2}^*(Z_{2n_2}))$, where some values of the first sample, which are larger than those of the second sample in the alternative due to positive fixed effects, could have been permuted to the second sample.

Therefore, $\Delta(\mathbf{X}^*(\delta)) - \Delta(\mathbf{X}(\delta))$ is expected to assume non-positive values. Depending on how this occurs, different types of conditional unbiasedness are presented:

1. If $\Delta(\mathbf{X}^*(\delta)) - \Delta(\mathbf{X}(\delta)) \leq 0$ for all $\mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}$ and for all data sets $\mathbf{X} \in \mathcal{X}$, then $\lambda_T(\mathbf{X}(0)) \geq \lambda_T(\mathbf{X}(\delta))$. This is denominated *strictly uniform conditional unbiasedness*.

2. If $\Delta(\mathbf{X}^*(\delta)) - \Delta(\mathbf{X}(\delta)) \leq 0$ is verified in terms of permutation distribution instead of pointwise, then the $p$-values also verify $\lambda_T(\mathbf{X}(0)) \geq \lambda_T(\mathbf{X}(\delta))$ distributionally talking.

3. If for some data set $\mathbf{X} \in \mathcal{X}$ and some permutation $\mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}$ $\Delta(\mathbf{X}^*(\delta)) - \Delta(\mathbf{X}(\delta)) > 0$ is given, then we may not have conditional unbiasedness. This could happen when the two CDF cross.

We will make use of the first kind of conditional unbiasedness, the strict uniform one, which is the most practical one.

### 3.1.2 Strictly Uniform Conditional Unbiasedness

Let us prove strictly uniform conditional unbiasedness for test statistics of the form we have proposed earlier, $T(\mathbf{X}) = S_1(\mathbf{X}_1) - S_2(\mathbf{X}_2)$.

The observed value in $H_0$ is

$$T^0(0) = S_1(\mathbf{Z}_1) - S_2(\mathbf{Z_2})$$

whereas, under $H_1$, the statistic $T$ assumes the value

$$T^0(\delta) = S_1(\mathbf{Z}_1 + \delta_1) - S_2(\mathbf{Z}_2)$$

Permutating the data set under $H_0$ and $H_1$, we obtain

$$T^*(0) = S_1(\mathbf{Z}_1^*) - S_2(\mathbf{Z}_2^*)$$

and

$$T^*(\delta) = S_1(\mathbf{Z}_1^* + \delta_1^*) - S_2(\mathbf{Z}_2^* + \delta_2^*) = T^*(0) + \Delta_S(\mathbf{Z}_1^*, \delta_1^*) - \Delta_S(\mathbf{Z}_2^*, \delta_2^*)$$

We introduce the term $\delta_2$ as some active effects of the first $n_1$ subjects may have been permuted to the second sample.

Now let us make the following appreciations:

- $\Delta_S(\mathbf{Z}_2^*, \delta_2^*) \geq \Delta_S(\mathbf{Z}_2^*, 0) = 0 = \Delta_S(\mathbf{Z}_2, 0)$ as effects $\delta_2^*$ are non-negative and $S_j$ are invariant over rearrangement.
- $\Delta_S(\mathbf{Z}_1^*, \delta_1^*) \leq \Delta_S(\mathbf{Z}_1^*, \delta_1)$ as it may be some values of the effects produced on the second sample in $\delta_1^*$, which are null.
- Finally, observe that $\Delta_S(\mathbf{Z}_1^*, \delta_1) = \Delta_S(\mathbf{Z}_1, \delta_1)$ in distribution because the points of the conditional reference space are uniformly distributed.

Due to all of this, $\Delta_S(\mathbf{Z}_1^*, \delta_1^*) - \Delta_S(\mathbf{Z}_2^*, \delta_2^*) \leq \Delta_S(\mathbf{Z}_1, \delta_1)$ pointwise, so the $p$-value verifies

$$\begin{aligned}
\lambda_T(\mathbf{X}(\delta)) &= \mathbb{P}[T(\mathbf{X}^*(\delta)) \geq T^0(\mathbf{X}(\delta))|\mathcal{X}_{|\mathbf{X}(\delta)}] \\
&= \mathbb{P}[T(\mathbf{X}^*(\delta)) - T(\mathbf{X}^*(0)) + T(\mathbf{X}^*(0)) + T^0(\mathbf{X}(0)) - T^0(\mathbf{X}(0)) \geq T^0(\delta))|\mathcal{X}_{|\mathbf{X}(\delta)}] \\
&= \mathbb{P}[T(\mathbf{X}^*(0)) + \Delta_S(\mathbf{Z}_1^*, \delta_1^*) - \Delta_S(\mathbf{Z}_2^*, \delta_2^*) - \Delta_S(\mathbf{Z}_1, \delta_1) \geq T^0(\mathbf{X}(0))|\mathcal{X}_{|\mathbf{X}(0)}] \\
&\leq \mathbb{P}[T^*(\mathbf{X}(0)) \geq T^0(\mathbf{X}(0))|\mathcal{X}_{|\mathbf{X}(0)}] = \lambda_T(\mathbf{X}(0))
\end{aligned}$$

It has been proved that permutation tests are conditionally unbiased for every attainable level of significance, independently of the population $P$ and for all data set $\mathbf{X} \in \mathcal{X}$.

Another important result is that, given two fixed effects $\delta$ and $\delta'$ such that $\delta \leq \delta'$, the respective $p$-values behave as $\lambda(\mathbf{X}(\delta)) \geq \lambda(\mathbf{X}(\delta'))$. We can regard $p$-values as non-decreasingly quantities with respect to the fixed effects of the treatment. In terms of effects, the larger the effect, the smaller the $p$-value (and so the more likely to reject null effects), which is heavily consistent with the theory we are presenting.

This aspect would help us when it comes to obtain confidence intervals for fixed effects $\delta$.

**Two-Sided Alternatives**

Now, instead of one-sided alternative, i.e., $H_1 : \{\delta \leq \delta_0\}$ or $H_1 : \{\delta \geq \delta_0\}$, we will focus on two-sided alternatives, where

$$H_0 : \{X_1 = X_2\} \text{ vs. } H_1 : \{X_1 \neq X_2\}$$

The first problem presented is that if fixed effects verify $\delta \neq 0$, a situation where $\delta$ provoking positive effects on some subjects while negative ones on others is perfectly compatible. If a suitable transformation $\phi$ of the sample space is such that $\phi(\mathbf{X}(\delta)) > \phi(\mathbf{X}(0))$, it is totally analogous to the ideas presented so far. However, this transformation $\phi$ is not easy to determine, and this makes the analysis substantially difficult.

As a solution, a relaxation on the sign of the fixed effects is assumed. According to this, we will accept that if $\delta$ is such that $|X_1+\delta-X_2| > 0$, this would imply that $|\delta| > 0$ in the alternative. To tackle this test, we will choose statistics which consider not only the difference between two samples but also its absolute value. For example, some useful statistics for this tests would be the absolute mean divergence, $T = |\sum_i X_{1i}(\delta)/n_1 - \sum_i X_{2i}/n_2|$, being some non-degenerate regular function of the sample means a possible approach, the squared mean divergence, $T = [\sum_i X_{1i}(\delta)/n_1 - \sum_i X_{2i}/n_2]^2$ , or the absolute divergence of a statistic $T$ following the form of 3.1.

Confidence intervals for $\delta$ would be another suitable option to approach these tests which will be discussed later.

Furthermore, according to this sign relaxation, the $p$-value would be calculated as

$$\lambda_T(\mathbf{X}(\delta)) = 1 - \mathbb{P}[-|T^0| \leq T \leq |T^0|])$$

due to the unknown direction of the effects (observe that the statistic $T$ could assume extreme values in a negative direction if the alternative is such that

$$H_1 : \{|\delta| > 0\}$$

with $\delta < 0$

## 3.2 Conditional Power Function

Let us now discuss the power function of these tests. Recall the power of a test is the probability of rejecting the null hypothesis when it is the alternative the one which is true. We thus look for tests which have a high power function. In a similar way to previous concepts which have been introduced, we give a conditional version of this notion.

**Definition 3.2.1.** *In a permutation test framework, the conditional power function is defined as*

$$W[(\delta,\alpha,T)|\mathcal{X}_{|\mathbf{X}}] = \mathbb{P}[\lambda_T(\mathbf{X}(\delta)) \leq \alpha|\mathcal{X}_{|\mathbf{X}(\delta)}]$$
$$= \mathbb{E}[\mathbb{I}[\lambda_T(\mathbf{X}^*(\delta))]|\mathcal{X}_{|\mathbf{X}^*(\delta)}]$$

Observe that this function depends on the considered level of significance, the statistic involved, the supposed effects and the sample size. The behaviour of $W$ with respect to the statistic $T$ is not uniform, whereas how the sample size affects would be discussed later.

With respect to $\delta$, attending to the comment of the inverse order which effects $\delta, \delta'$ and their respective $p$-values verify (recall the last paragraph of 3.1.2), it is straightforward that if $\delta < \delta'$ then

$$W[(\delta,\alpha,T)|\mathcal{X}_{|\mathbf{X}}] \leq W[(\delta',\alpha,T)|\mathcal{X}_{|\mathbf{X}}]$$

for every $\mathbf{X} \in \mathcal{X}$ and any attainable $\alpha$-value (in fact, this is highly consistent because the larger the effects are, the more likely to reject null effects.)

As well as the $p$-value, computing the conditional power function is not a trivial task. Remark that $\lambda_T(\mathbf{X}^*(\delta))$ is the $p$-value calculated on $\mathbf{X}^*(\delta) = (\mathbf{Z}_1^* + \delta, \mathbf{Z}_2^*)$, so we need to compute it on all possible permutations of the random deviations $\mathbf{Z}$.

Following the same philosophy of $p$-value estimation, we propose a Monte Carlo algorithm to get an approximation of the conditional power function:

1. First consider the pooled deviates $\mathbf{Z}$ and the effects $\delta$.

2. Randomize $\mathbf{Z}$ obtaining $\mathbf{Z}^*$, resulting $\mathbf{X}_l^*(\delta) = (\mathbf{Z}_{l1}^* + \delta, \mathbf{Z}_{l2}^*)$ and apply the Monte Carlo algorithm for obtaining an approximation of the $p$-value $\hat{\lambda}_T(\mathbf{X}_l^*(\delta))$, all this for $N$ times.

3. The approximation is $\hat{W} = \sum_l \mathbb{I}[\hat{\lambda}(\mathbf{X}_l^*(\delta)) \leq \alpha]/N$

Remark that the conditional power function is a non-decreasing function of $\alpha$. In effect, if we fix $0 < \alpha_1 \leq \alpha_2 \ < 1$, it is easily followed that

$$W[(\delta,\alpha_1,T)|\mathcal{X}_{|\mathbf{X}}] \leq W[(\delta,\alpha_2,T)|\mathcal{X}_{|\mathbf{X}}]$$

due to

$$\mathbb{P}[\lambda_T(\mathbf{X}(\delta)) \leq \alpha_1 | \mathcal{X}_{|\mathbf{X}(\delta)}] \leq \mathbb{P}[\lambda_T(\mathbf{X}(\delta)) \leq \alpha_2 | \mathcal{X}_{|\mathbf{X}(\delta)}]]$$

just attending to monotony (of course, the more relaxed the level of significance, the more likely to reject the null hypothesis).

However, a striking detail flourishes when examining the previous algorithm. It uses the random deviates $\mathbf{Z}$, but in practice, these deviations are homogeneous along with the supposed effects $\delta$. We should be able to separate them in order to apply the mechanism, so we introduce an alteration in the algorithm which contemplates an estimation of these deviations.

It is as follows:

1. Consider a suitable statistic $T$ for estimating $\delta$, obtaining a point estimation $\hat{\delta}$ from the observed data set $\mathbf{X}(\delta)$ . A point estimation of the random deviations is $\hat{\mathbf{Z}} = (\mathbf{X}_1 - \hat{\delta}, \mathbf{X}_2)$.

2. Randomize the estimated deviations $\hat{\mathbf{Z}}^*$, with $\hat{\mathbf{X}}_l^*(\delta) = \{\hat{\mathbf{Z}}_{l1}^* + \delta, \hat{\mathbf{Z}}_{l2}^*\}$ and compute the $p$-value $\hat{\lambda}_T(\mathbf{X}_l^*(\delta))$ for $N$ times.

3. The estimation is $\hat{W} = \sum_l \mathbb{I}[\hat{\lambda}_T(\mathbf{X}_l^*(\delta)) \leq \alpha]/N$.

In case the underlying population $P$ is known (including its analytical form) we can obtain the unconditional power function as

$$\begin{aligned} W(\delta, \alpha, T, P, n) &= \mathbb{E}_{\mathcal{X}}[W(\delta, \alpha, T, n) | \mathcal{X}_{|\mathbf{X}}] \\ &= \int_{\mathcal{X}} \mathbb{I}[\lambda_T(\mathbf{X}(\delta)) \leq \alpha | \mathcal{X}_{|\mathbf{X}}] f_P(\mathbf{X}(\delta)) \end{aligned}$$

Remark that due to averaging with respect the sample space $\mathcal{X}$ we first have to take the mean with respect to the conditional distributions over $\mathcal{X}_{|\mathbf{X}}$ and then take the mean of these with respect to the partition of the sample space into orbits. A similar Monte Carlo algorithm to the one explained for the conditional power function could be applied:

1. Choose a value of $\delta$ and simulate a sample of size $n$ from $P$. Add the effects to the first $n_1$ deviates, resulting in $\mathbf{X}_l(\delta) = (\mathbf{Z}_{l1} + \delta, \mathbf{X}_{l2})$. Do this $N$ times.

2. Compute the $p$-value $\hat{\lambda}_T(\mathbf{X}_l^*(\delta))$ for each sample obtained in 1.
3. The estimation is $\hat{W} = \sum_l \mathbb{I}[\hat{\lambda}_T(\mathbf{X}_l^*(\delta)) \leq \alpha]/N$.

To obtain a function in $\delta, \alpha, T$ and $n$, we just need to compute the algorithm for different values of these elements.

The unconditional power function could be used to determine if a statistic $T$ is unbiased checking if it verifies $\hat{W}(\delta, \alpha, T, P, n) \geq \alpha$.

This function can also be interpreted as a least squares estimation, due to its obtaining through an expectancy.

### 3.2.1 Let's make it unconditional!

In light of the results previously described we set out the possibility of obtaining unconditional decisions based on the conditional ones we have already obtained.

Let us consider a non-randomized permutation test in the previous conditions. As we have seen, these tests verify two important properties which will be extremely handy to develop this topic. The first one is the conditionally unbiasedness they present, while the second one is the similarity property. Thanks to these two facts, conditional inference conclusions might be extended to a non conditional framework.

Due to the similarity property, for any attainable $\alpha$-value the unconditional power function in $H_0$ verifies

$$W(0, \alpha, T, n) = \int_{\mathcal{X}} \mathbb{P}[\lambda_T(\mathbf{X}(0)) \leq \alpha | \mathcal{X}_{\mathbf{X}}] f_P(\mathbf{X}) = \alpha$$

whereas, attending to unbiasedness, for each $\delta > 0$,

$$W(\delta, \alpha, T, n) = \int_{\mathcal{X}} \mathbb{P}[\lambda_T(\mathbf{X}(\delta)) \leq \alpha | \mathcal{X}_{\mathbf{X}}] f_P(\mathbf{X}) \geq \alpha$$

This way, we are permitted to extend our conditional results to populations $P$ from which our samples have been taken, that is, unconditionally. For example,

suppose we are testing a drug on the data set $\mathbf{X}$. Owing to this result, in case that evidence against $H_0$ is presented and thus rejected, we can conclude that effects are non-null not only on the individuals of our concrete sample but also from every person belonging to $P$. However, this extension is not valid if the density function $f_P$ of a population $P$ is almost zero, as similarity property or conditional unbiasedness may not be guaranteed.

## 3.3 Constructing Confidence Intervals

This section would illustrate one of the most practical and useful utilities. We want to propose confidence intervals for the fixed effects $\delta$ of the treatment.

Let us suppose two samples $\mathbf{X}_1 = \delta + \mathbf{Z}_{1i}$ for $i = 1, ..., n_1$ and $\mathbf{X}_2 = \mathbf{Z}_{2i}$ for $i = 1, ..., n_2$, and $\mathbf{X}$ the concatenation of the two samples we have been using all along this work. Our goal is to find two values $\delta_{low}$ and $\delta_{up}$ verifying

$$\mathbb{P}[\delta_{low} \leq \delta \leq \delta_{up} | \mathcal{X}_{|\mathbf{X}(\delta)}] = 1 - \alpha$$

for any $\alpha \in (0, 1)$.

In fact, these limits are functions of the data set $\mathbf{X}(\delta)$. We now announce a handy proposition for finding these quantities.

**Proposition 3.3.1.** *Let $IC(\delta)_{1-\alpha} = (\delta_{low}, \delta_{up})$ be a confidence interval for $\delta$ of confidence level 1 - $\alpha$. Then all the values $\delta'$ for which the null hypothesis $H_0 : \{\mathbf{X}_1(\delta) - \delta' = \mathbf{X}_2\}$ against $H_1 : \{\mathbf{X}_1(\delta) - \delta' \neq \mathbf{X}_2\}$ is accepted at level $\alpha$ are contained in $IC$.*

The proof of this result is basically due to the relationship between confidence intervals at level $1 - \alpha$ and tests based on rejecting when the null value does not fall within the interval. That is, considering the test

$$\phi = \begin{cases} 1 & \text{if } \delta \notin (\delta_{low}, \delta_{up}) \\ 0 & \text{if } \delta \in (\delta_{low}, \delta_{up}) \end{cases}$$

we know this is an $\alpha$ size test. If we subtract a certain quantity $\delta'$ for which the null hypothesis stated before is not rejected, then it is because that value clearly falls within the $1-\alpha$ confidence interval for the real value of $\delta$.

Then based on this proposition, for a given statistic $T$ and an observed data set $\mathbf{X}$, our aim is to determine the set of values for which the null hypothesis is not rejected. The fact of $T$ being permutationally equivalent to an estimator $\hat{\delta}$ for $\delta$ ease this task (for example, $T = \sum_i X_{1i}/n_1 - \sum_i X_{2i}/n_2 = \bar{X}_1 - \bar{X}_2$).

Once again we resort to Monte Carlo simulation for finding the lower limit $\delta_{low}$ (respectively, an algorithm for $\delta_{up}$ is presented later).

1. In first place, once $\varepsilon$ (the desired width of the interval) and the level of confidence $\alpha$ are fixed, pick a negative number $\eta < 0$ and consider

$$\mathbf{X}_1(\eta) = \{X_{1i}(\delta) - (\hat{\delta} + \eta),\ i = 1, ..., n_1\}$$

2. Using also a Monte Carlo procedure based on $B$ iterations, compute the empirical distribution function of $T$, $\hat{F}^*_B(T^0_\eta)$, on $T^*(\eta) = \bar{X}^*_1(\eta) - \bar{X}^*_2(\eta)$ (with $T^0_\eta = T[\mathbf{X}(\eta)]$).

3. Repeat the previous steps varying the $\eta$ value until $|1-\hat{F}^*_B(T^0_\eta)-\alpha/2| < \varepsilon/2$, resulting in $\delta_{low} = \hat{\delta} + \eta$.

The algorithm for obtaining the upper limit is similar, just introducing a little variation. We will want to satisfy the condition $|\hat{F}^*_B(T^0_\eta) - \alpha/2| < \varepsilon/2$, and $\delta_{up} = \hat{\delta} + \eta$ with $\eta > 0$ in this case.

The initial selection of $\eta$ seems to be pretty heuristic. To make that approximation, we can attend to some values resulting from some parametric method if possible.

In case the permutation support $\mathcal{T}(\mathbf{X})$ and the related attainable $\alpha$-values are defined in discrete sets, the lower and upper limits are predetermined by the values in the permutation support whereas the probability of the interval is determined by the jumps of the permutation distribution function on the permutation

support. These intervals will be of the form $T_{(k)} < t < T_{(k+1)}$, leaving inside a probability of

$$F([T_{(k+1)}|\mathcal{X}_{|\mathbf{X}}]) - F([T_{(k)}|\mathcal{X}_{|\mathbf{X}}])$$

## 3.4 Asymptotic Properties

### 3.4.1 Consistency

Keeping our aim of exploring different properties of permutation tests, we now move on to the notion of consistency, which is intimately related to the power function of a test. Recall a statistic $T$ is said to be consistent if, for fixed values of $\alpha \in (0, 1)$ and $\delta > 0$, the unconditional power function verifies

$$\lim_{n\to\infty} W(\delta, \alpha, n) = 1$$

This concept remarks that the test based on the test statistic $T$ is almost infallible when the sample size becomes extremely large, as we are providing an asymptotic probability of 1 for rejecting the null hypothesis when it is false.

### 3.4.2 Asymptotic Behaviour of the Critical Value

In contrast with parametric tests, where the critical value is a fixed quantity, this now could vary as $\mathbf{X}$ varies in $\mathcal{X}$, that is, they are random variables (recall 2.2.2). Let us study how does this variable $T_\alpha$ behave when the sample size goes to infinity following [6].

Let us assume $\mathbf{X}_n \in \mathcal{X}^n$, $M_n$ and $T(\mathbf{X}_n)$, all of them considered successions. We shall also assume that $M_n \to \infty$ and $\sum_{\mathcal{X}_{|\mathbf{X}}}(T^* < T_\alpha)/M_n \to 1 - \alpha$ as $n \to \infty$ (that is, the test is of size $\alpha$).

Consider the randomized test

$$\phi_R = \begin{cases} 1 & \text{if } T(\mathbf{X}_n) > (t_\alpha)_n \\ a(x) & \text{if } T(\mathbf{X}_n) = (t_\alpha)_n \\ 0 & \text{if } T(\mathbf{X}_n) < (t_\alpha)_n \end{cases}$$

with $0 \leq a(x) < 1$ and suppose that the following two conditions are satisfied:

(1) There exists a constant $l$ such that $(T_\alpha)_n \to l$ in probability.

(2) There exists a function $J(y)$ continuous at $y = l$ such that for every $y$ at which $J(y)$ is continuous,

$$\mathbb{P}[T(\mathbf{X}_n) \leq y] \to J(y)$$

Under these assumptions, we have

$$\mathbb{P}[T(\mathbf{X}_n) > (T_\alpha)_n | \mathbf{X}_n] \leq \mathbb{E}_{\mathcal{X}_{|\mathbf{X}}}[\phi_R(\mathbf{X}_n)] \leq \mathbb{P}[T(\mathbf{X}_n) \geq (T_\alpha)_n | \mathbf{X}_n]$$

which, thanks to (1) and (2), clearly leads to

$$\mathbb{E}_{\mathcal{X}_{|\mathbf{X}}}[\phi_R(\mathbf{X}_n)] \to 1 - J(l)$$

Both $l$ and $J(y)$ depends on the sequence $P_n$. However, the dependence of $J(y)$ is stronger than that of $l$. Due to this, we may introduce an alternative condition for (1):

(1') $F(y|\mathcal{X}_{|\mathbf{X}}) \to F(y)$ in probability for every $y$ at which $F(y)$ is continuous, where $F(y)$ is a distribution function, the equation $F(y) = 1 - \alpha$ has a unique solution $y = l$ and $F(y)$ is continuous at $y = l$.

Actually this latter assumption implies (1). In effect, by definition

$$\mathbb{P}[(T_\alpha)_n \leq y] = \mathbb{P}[F(y|\mathcal{X}_{|\mathbf{X}}) \geq \sum_{\mathcal{X}_{|\mathbf{X}}} (T^* < T_\alpha)/M_n]$$

for every $y \in \mathbb{R}$. Let $y$ be a point of continuity of $F(y)$. As we have assumed, $\sum_{\mathcal{X}_{|\mathbf{X}}} (T^* < T_\alpha)/M_n \to 1 - \alpha = F(l)$, and $y < l$ implies $F(y) < F(l)$, so the right-hand side tends to 0 if $y < l$. It analogously tends to 1 in case $y > l$, and then $(T_\alpha)_n \to l$ in probability.

What is more, let $\mathbf{X}_n^*$ be a permutation of $\mathbf{X}_n$, and $\mathbf{X}'_n$ another element in the sample space, the three of them mutually independent. If $T(\mathbf{X}_n^*)$ and $T(\mathbf{X}'_n)$ have the limiting joint distribution $F(y) \cdot F(y')$, then for every $y$ at which $F(y)$

is continuous $F(y|\mathcal{X}_{|\mathbf{X}}) \to F(y)$ in probability and, if the equation $F(y) = 1 - \alpha$ has a unique solution $y = l$, it is also verified that $(T_\alpha)_n \to l$ in probability.

We now define the test $\phi'$:

$$\phi' = \begin{cases} 1 & \text{if } T(\mathbf{X}_n) > (t_\alpha)_n \\ \gamma_n & \text{if } T(\mathbf{X}_n) = (t_\alpha)_n \\ 0 & \text{if } T(\mathbf{X}_n) < (t_\alpha)_n \end{cases}$$

assuming that the test has size $\alpha$. If the previous assumption (1') is satisfied, then we can firmly claim $(t_\alpha)_n \to l$. If (2) is also satisfied, we have $\mathbb{E}[\phi'(\mathbf{X}_n)] \to 1 - J(l)$.

Considering the subfamily of distributions $\hat{\mathcal{P}} \subseteq \mathcal{P}$ for which (1') and (2) holds, if $\hat{\mathcal{P}}$ contains all sequences induced by $H_0$, then $\phi_R$ and $\phi'$ are asymptotically equivalent in terms of power.

An interesting application of this could be seen when testing symmetry. When approximating by a standard normal distribution the permutation distribution of a test statistic, (1') holds. Because of this, the test based on the sample mean asymptotically behaves as the one-sided Student's $t$-test of size $\alpha$.

# Chapter 4

# Applying Theory

Once we have introduced the main theoretical aspects of permutation tests, we move on to illustrate them in some practical cases. The first case we will be studying is the one which embodies the theoretical framework that has been developed in this project. This is the hypothesis testing determining the existence of some kind of fixed effects on two populations versus the absence of these possible effects. The second case would present another hypothesis testing environment on which permutation approach would be useful. In this case we will be testing whether some distribution is symmetric or not. Finally, a third case would be introduced to show a possible permutation approach in a different nature problem, as we would be dealing with categorical variables.

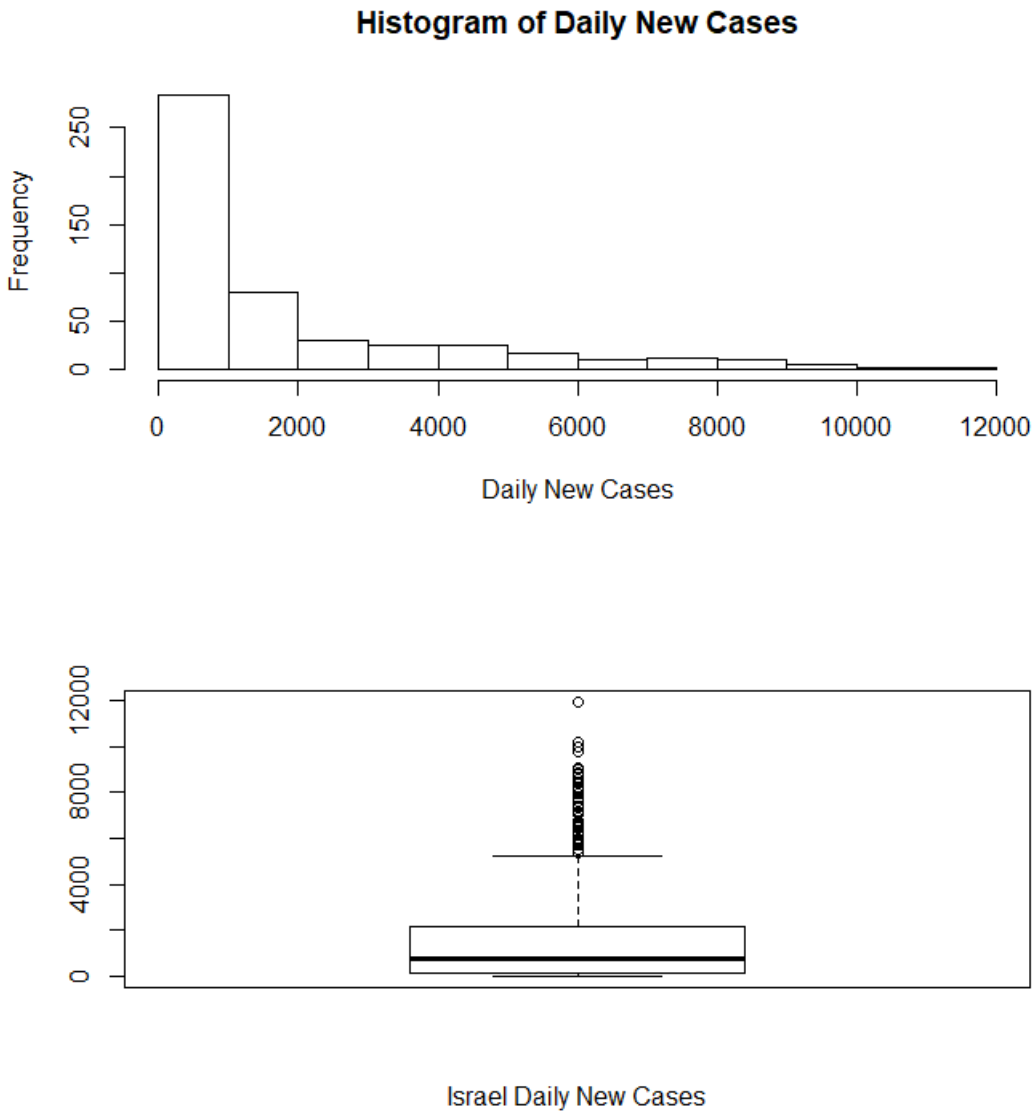
## 4.1 Testing Fixed Effects

As it was introduced, the first applied case concerning our study is testing the existence of non-null fixed effects. The dataset used for this task is related to the COVID-19 pandemic ([9]). This file contains 59 variables which gather information about the pandemic around the world, such as the date when data was collected, the country it belongs to; and other variables which are directly related to the disease itself by measuring some pandemic features such as daily new cases, total deaths, or the basic reproduction number of the disease (usually

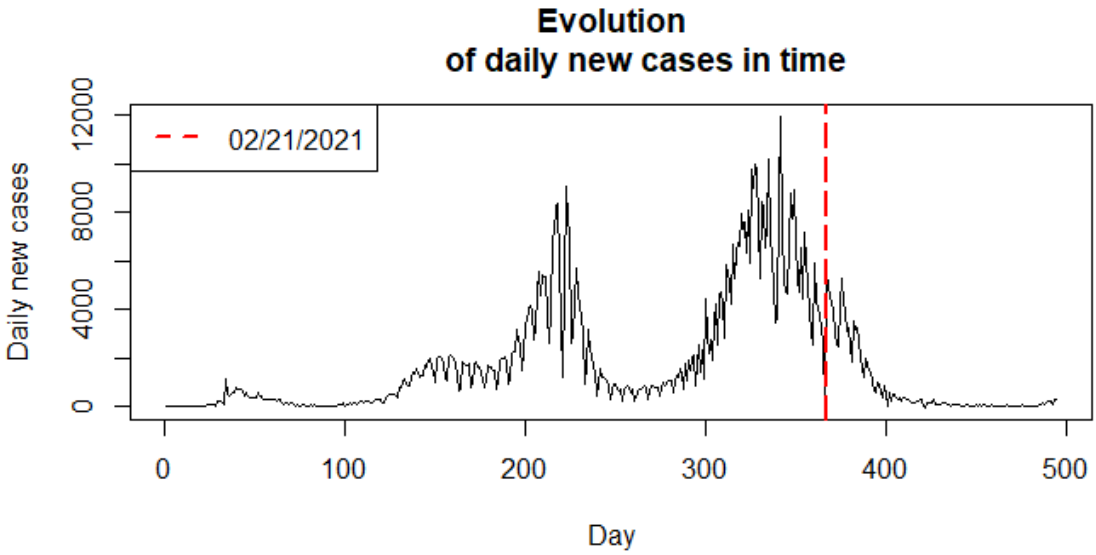
noted $R_0$).

Our purpose here is to statistically determine, through a hypothesis testing, if vaccination is effective or not. As it is widely known, Israel was the fastest country when vaccinating its population. Because of these, we will isolate data coming from Israel and we will be working on this dataset.

We select the daily new cases of this country from 02/21/2020 to 06/30/2021, studying the vaccination impact on this variable.



Furthermore, we will make a second filter by splitting the data into two groups. The first one is the data within 02/21/2020 to 02/26/2021, while the second group is the data collected between 02/27/2021 and 06/30/2021. The reason behind this particular date is that it is the day when 50% of Israelis had already received at least one COVID-19 vaccine dose. We now study the behaviour of new daily cases in both groups.

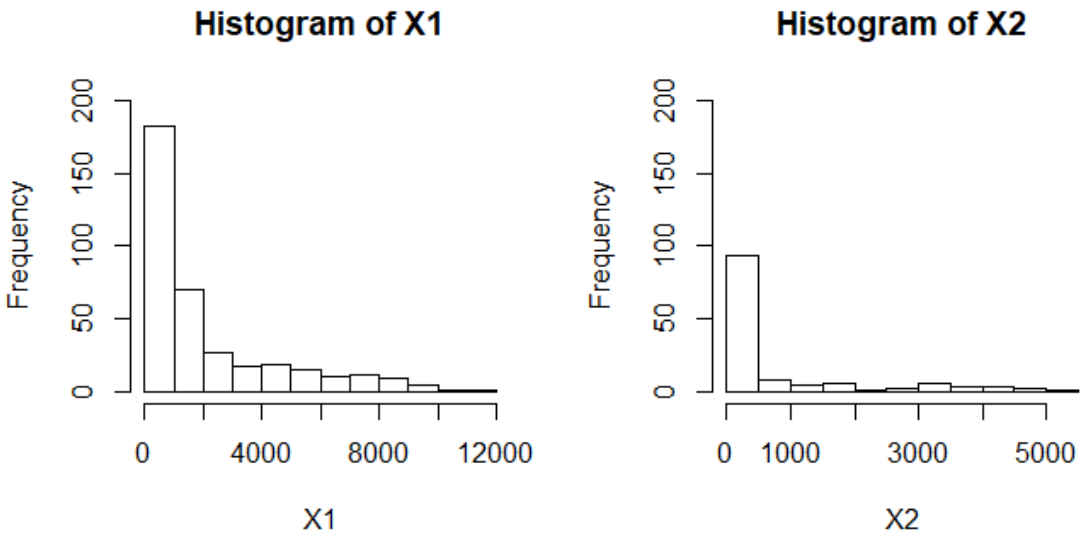



Let us denote $\mathbf{X}_1 = (X_{11}, X_{12}, ..., X_{1367})$ the observations coming from the first group and $\mathbf{X}_2 = (X_{21}, X_{12}, ..., X_{2128})$ the second group. Remark our sample size is $n = 495$, with $n_1 = 367$ and $n_2 = 128$. We have two populations, and we suspect the first one is affected by some kind of fixed effects due to the absence of vaccination. As there are less vaccinated inhabitants in the first group, the probability of being infected with COVID-19 could seem to be higher than the that in the second sample. This leads to our scheme $(\mathbf{X_1} + \delta, \mathbf{X_2})$ and the hypothesis testing

$$H_0 : \{\delta = 0\} \text{ vs. } H_1 : \{\delta > 0\}$$

First thing we should make is testing normality of both populations. Of course, if normality is not rejected, we could assume it and consequently apply traditional parametric methods such as an ANOVA or a $t$-student test. We run

the well-known Shapiro-Wilk test to this aim.




```
> shapiro.test(X1)

        Shapiro-Wilk normality test

data:  X1
W = 0.78685, p-value < 2.2e-16

> shapiro.test(X2)

        Shapiro-Wilk normality test

data:  X2
W = 0.82693, p-value = 3.705e-14
```

As we see, both $p$-values are extremely small, leading to the rejection of normal behaviour in both populations. We should then look for some statistical alternatives, which would clearly be the permutation approach.

The next step when testing is a pretty sensitive issue: we should determine what statistic $T$ should we select to base our test on. As we have expounded previously, the difference of averaged means seems a suitable statistic to our aim. We thus select $T = \sum_{i=1}^{n_1} X_{1i}/n_1 - \sum_{i=1}^{n_2} X_{2i}/n_2$. Observe that under the alternative $T$ would assume lower values when mixing observations between samples, as the first population is provided of not null side effects, whereas the second one

assumes lower values.

At this point, results explained in section 2.3 become extremely useful. We will work with the test statistic $\tilde{T} = \sum_{i=1}^{n_1} X_{1i}/n_1$ due to its permutational equivalence to $T$. This fact eases the computational aspect of permutation tests.

After setting the test up we should now compute the $p$-value. We go to the Monte Carlo algorithm presented in 2.2.2, fixing $B = 1000$.

```
Z = c(X1, X2)

B = 1000;

T0 = mean(X1)

T <- array(0, dim = c(B+1,1))

T[1] = T0

for (b in 1:B)
{
  Zstar = sample(Z)
  T0star = mean(Zstar[1:n1])
  T[b+1] = T0star
}

extremevalues = T[T>=T0]

pvalue = length(extremevalues)/B
```

1
2
3
4

Figure 4.1: MC Algorithm Implementation following 2.2.2

```
> pvalue
[1] 0.001
```

The resulting $p$-value allows us to reject the null hypothesis and strongly conclude that $\delta$ effects are such that $\delta > 0$.

Running the algorithm for $B = 10000$, we obtain

```
> pvalue
[1] 1e-04
```

We can appreciate that the $p$-value decreases as the number of iterations $B$ increases. Attending to the final comment made in 2.2.2, we can ensure through

Glivenko-Cantelli theorem that the $p$-value converges to its real value as $B$ goes to infinity. The fact that the $p$-value has decreased as $B$ has increased enhances the idea of the real value of $\lambda$ being sufficiently small so as to reject the null hypothesis.

### 4.1.1 Balancing data

As Taleb states in [8], we should go from problems arisen due to empiricism to books and not the other way. Here we illustrate a situation which suitably embodies this idea.

The first time the previous experiment was carried out was 04/03/2021. At this time the sample size was $n = 408$, with the first group size $n_1 = 372$ whereas the second one was $n_2 = 35$. There is a huge difference between sample sizes. Data is heavily unbalanced, being the majority of the observations in the first group. Let us see how this fact could massively affect to our results.

Recall the $T$ statistic we have proposed has the form $T(\mathbf{X}) = S_1(\mathbf{X}_1) - S_2(\mathbf{X}_2)$. Let us suppose now that the difference of group sizes is pretty considerable, with $n_1 > n_2$ (the case $n_2 > n_1$ is totally analogous). This would clearly lead to a problem when mixing observation between groups, as changes in the structure of the observed sample $\mathbf{X}_1$ (here the group which dominates in size terms) would be insignificant, and consequently values assumed by $T$ when permuting data would be very similar to that assumed in $T^0(\mathbf{X})$. As a consequence, the estimation of the $p$-value using the Monte Carlo Algorithm, which is

$$\hat{\lambda} = \sum_{i=1}^{B} \mathbb{I}[T_i^*(\mathbf{X}^*) \geq T^0]/B$$

could be disproportionately bigger than its real approximation, causing severe mistakes in our decisions.

Let us show this problem in a concrete example going back to the situation which gave rise to this section. The probability of obtaining an individual coming

from the second group on the first observation of the first group when permuting data is $n_2/n = 0.09$ whereas that of obtaining someone proceeding from the first group is 0.91.

Let us suppose we have been lucky and an individual from the second group has been set in the first place of the permuted sample of the first group. At this point, the probability of obtaining someone from the second group in the second place of $\mathbf{X}_1^*$ is $(n_2 - 1)/(n - 1) = 0.083$, while that of obtaining someone proceeding from the first group is now 0.917. It is clear that we are not going to be that lucky every time, and if we keep in mind we have to repeat this process $B$ times, all the more reason to realize this would not work (just the simple Law of Large Numbers argument reinforces this fact).

Computing the MC Algorithm with the data previously presented for $B = 1000$, the resulting estimated $p$-value was $\hat{\lambda} = 0.221$, which would categorically lead us to the non-rejection of the null hypothesis, concluding that we cannot refute $\delta = 0$. As it can be appreciated, the conclusion of this testing is vehemently opposed to that of the first experiment with balanced data.

In conclusion, we need to treat with balanced data (or balancing it if not the case) to avoid conclusion fallacies.

## 4.2 Testing Symmetry

Now that we have shown an effective application of the permutation strategy based on the theoretical framework, let us illustrate this technique in a hypothesis testing of different nature with respect to that previously studied.

In this case, we will study the symmetry of a population. Our aim is to determine if the followed distribution is symmetric, that is

$$H_0 : F(x) = 1 - F(-x)$$

versus the alternative

$$H_1 : F(x) \neq 1 - F(-x)$$

Remark this symmetry is studied with respect to zero.

Let us consider the sample data $\mathbf{X} = (X_1, ..., X_n)$. In order to approach this problem, we write $\mathbf{X} = \mathbf{Y}_1 - \mathbf{Y}_2$ as the difference of two possible underlying paired observations. This could be interpreted as $X_i$ the result of measuring some event in two different times, say 1 and 2, resulting in $X_i = Y_{1i} - Y_{2i}$. It is clear that $X_i$ is the difference of two paired observations (as the individual on which $Y_1$ and $Y_2$ have been measured is the same).

Attending to this, we can equivalently state the hypothesis testing as

$$H_0 : Y_1 = Y_2$$

versus the alternative hypothesis

$$H_1 : Y_1 \neq Y_2$$

We now focus on the latter.

Remark that under $H_0$ (that is, $Y_1 = Y_2$) we can exchange observations between variables without distinction as they follow the same distribution. So given the observations $y_1$ and $y_2$, it is verified that $F_{Y_1}(y_1) = F_{Y_2}(y_1)$ and $F_{Y_2}(y_2) = F_{Y_1}(y_2)$. Because of this, if we have an observation $x_i = y_{1i} - y_{2i}$, $F_X(x_i) = F_{Y_1}(y_{1i}) - F_{Y_2}(y_{2i}) = F_{Y_1}(y_{2i}) - F_{Y_2}(y_{1i})$, and we can thus infer that $x_i = y_{2i} - y_{1i}$ in probability terms.

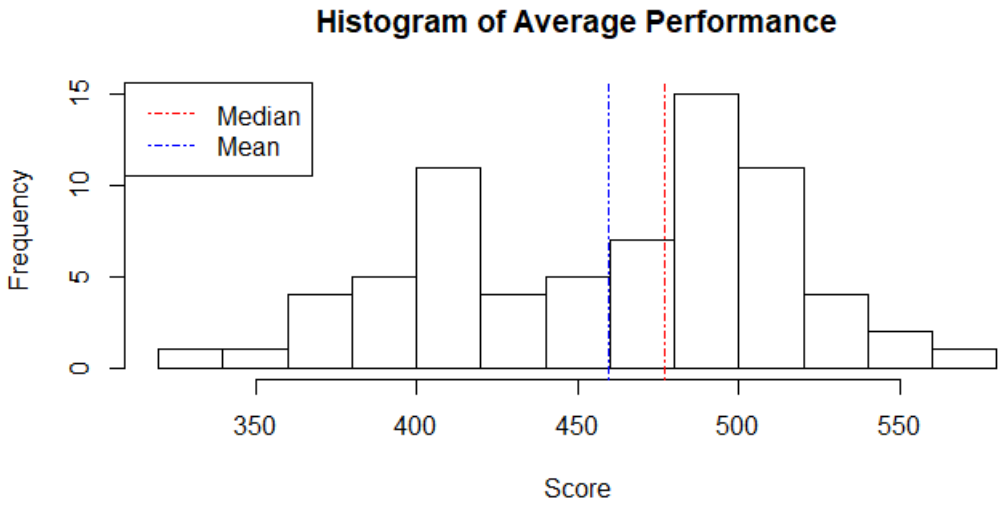
This fact is translated into random sign assignment of $X_i$, as $X_i = Y_{1i} - Y_{2i} = Y_{2i} - Y_{1i}$ in case of symmetry. According to this, our permutation sample space could be written as $\mathcal{X}_{|\mathbf{X}} = \{\bigcup_U \ X_i U_i \ , i = 1, ..., n\}$ with $U_i \sim 1 - 2\mathcal{B}(1, 1/2)$.

Finally, a suitable test statistic would be $T = Md(X)$ or $T = |\sum_i X_i/n|$. It is easy to observe that in $H_0$ these statistics would assume values close to 0 as

signs are basically randomly distributed, whereas in $H_1$ the mean would assume more dispersed values and the absolute mean would assume larger values.

An important observation is that if data is suspected to be symmetric with respect to a different value than the origin, we just have to make the transformation $\mathbf{X} - Md(\mathbf{X})$ if necessary.

Let us implement everything previously showed. The dataset we are going to analyse is about the Programme for International Student Assessment (commonly known as PISA study), which is a worldwide study to evaluate educational systems of different countries by measuring students' performance on Mathematics, science and reading. For this study, we have selected the mean performance of $n = 72$ countries in Mathematics of the 2015 report.



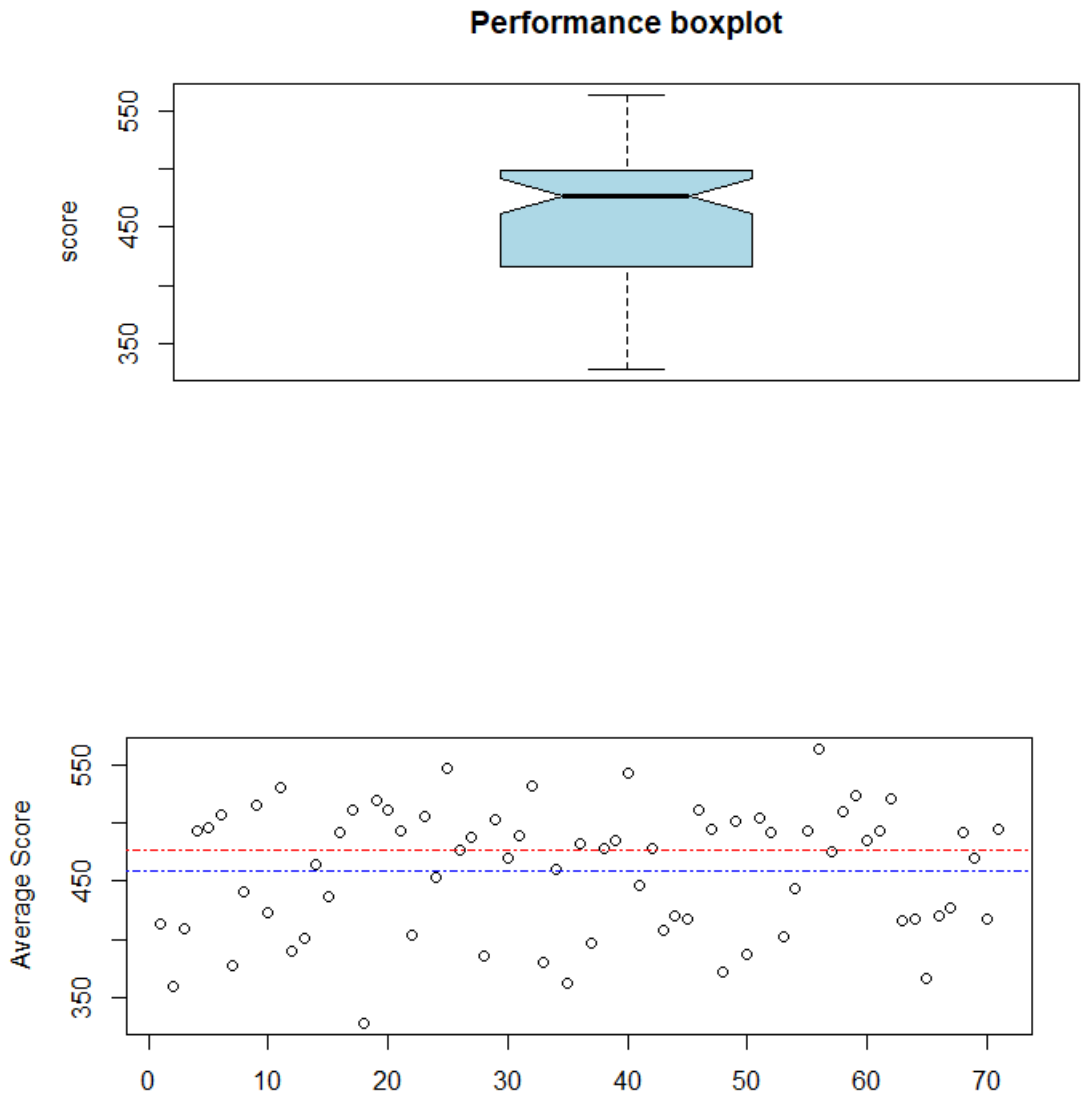



Figure 4.2: Population point cloud including the mean. It seems to be more dense on the upper side. Mean in blue and median in red.

Let us write $\mathbf{X} = (X_1, ..., X_{72})$ the observed sample. Similarly to the previous dataset, we study the normal behaviour of the population through the Shapiro-Wilk test

```
> shapiro.test(datos$Average)

        Shapiro-Wilk normality test

data:  datos$Average
W = 0.95648, p-value = 0.01502
```

The resulting $p$-value leads us to reject normality. We will approach this testing using the most obvious test statistic: the median. Furthermore, we would centre data with respect to the median to have a clearer graphical interpretation.

Let us first use $T = Md(X)$. It is extremely similar to the process attached in the first study case, but we now have to introduce the random sign assignment discussed before. We run the MC algorithm setting $B = 10000$

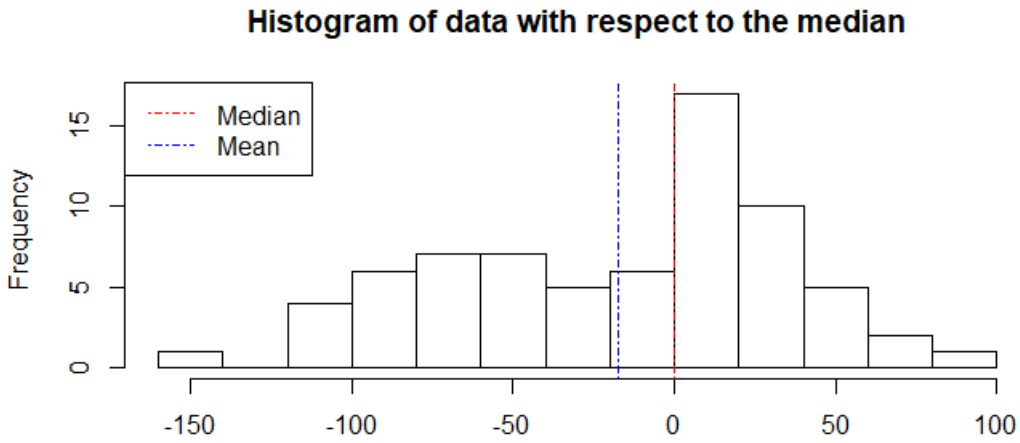
```
T0 = median(datos$Average)

median(datos$Average)

B = 10000

T = array(0, dim = B+1)

T[1] = T0

sign = c(1,-1)

for (b in 1:B)
{

  u = sample(sign, n, replace = TRUE)

  Tstar = median(u*datos$Average)

  T[b+1] = Tstar

}

extremedata = T[T >= T0]

pvalue = length(extremedata)/B
```

The $p$-value turns out to be

```
> pvalue
[1] 1e-04
```

so we should reject the null hypothesis, resulting in asymmetry.

Centring data with respect to the, we can appreciate data is slightly right-biased, so higher values seem more probable than little ones. If the distribution were symmetric, we should expect a head as heavy as the tail. This is highly consistent with the decision we have made about the hypothesis testing.




### 4.2.1 Asymptotic behaviour for symmetry statistics

Let us consider the test statistic

$$T = \sum_{i=1}^{n} X_i \cdot U_i / (\sum_{i=1}^{n} X_i^2)^{1/2}$$

The denominator is invariant at the permutation sample space. As $\mathbb{E}(T) = 0$ and $Var(T) = 1$, we can conclude $T$ is the sum of $n$ independent variables. Due to the PCLT, $T \sim \mathcal{N}(0, 1)$ in the limit.

## 4.3 A study for Ordered Categorical Variables

The third case study heavily differs from the two previous ones. Our aim now is to test different nature variables to those hitherto studied. We would be

treating with categorical variables which would intrinsically possess some order notion.

The current framework would be a categorical variable $\mathbf{X}$ partitioned into $h \geq 2$ classes, $\{A_k,\ k = 1, ..., h\}$. As we commented in advanced on the previous paragraph, the relationship $A_k < A_j$ for every pair where $1 \leq k < j \leq h$ would be some self-explanatory issue.

Given two independent random samples $\mathbf{X}_1$ and $\mathbf{X}_2$ our purpose here is to conclude whether categories are equally distributed in both samples or if distributions vary from one sample to another. In terms of hypotheses testing, this would be stated as

$$H_0 : F_1(k) = F_2(k) \ \forall \ k = 1, ..., h$$

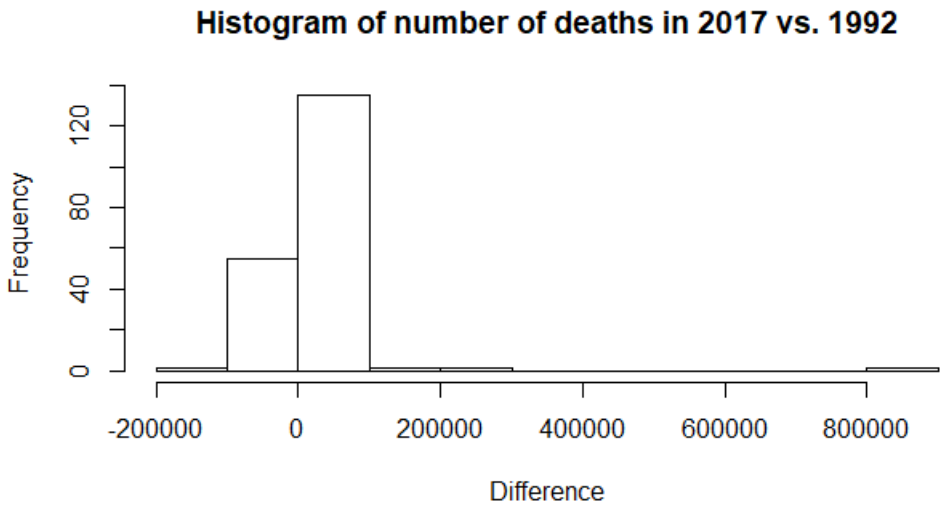
against the alternative hypothesis claiming some difference on the impact of at least one class, that is

$$H_1 : F_1(k) \neq F_2(k) \text{ for some } k = 1, ..., h$$

The notation $F_j(k) = \mathbb{P}\{X_j \leq A_k\}, j = 1, 2$ is clearly consistent due to the order notion associated to classes.

The dataset that we will be using in this study case is related to the number of deaths provoked by smoking in different countries periodically from 1990 to 2017 [10]. The socio-economic level of each country is also gathered in the dataset. This would be our target variable.

We then want to study if the socio-economic level is a meaningful factor in smoking deaths. For this task, we select the number of passings in each country in 1992 and we compare it to that in 2017. To make a distinction, we would split data into two groups, the first group being countries where the number of deaths have decreased whereas the second groups is conformed by those where deaths have increased. This way we have two independent random samples to test the hypothesis testing of interest.




This split results in two samples: the first one is $\mathbf{X}_1 = (X_{11}, ..., X_{156})$ whereas the second one is $\mathbf{X}_2 = (X_{21}, ..., X_{2138})$. As it can be appreciated on the upper histogram, the majority of death differences accumulate around 0.

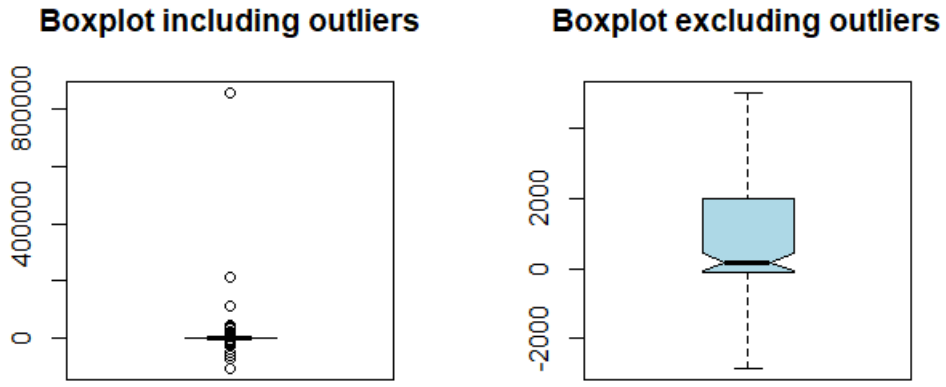



Figure 4.3: A remarkable outlier could be appreciated.

However, we can detect numerously outliers in the sample, including one which is specially far from the rest of observations. This case is China, a country where deaths has hugely gone up.

Once we have split and skimmed the data, let us analyse the variable of interest. The socio-economic level of each country is specified in four different categories: low, lower-middle, upper-middle and high. As we previously explained, classes respect some kind of order relationship.

| | ***Low*** | ***Lower-middle*** | ***Upper-middle*** | ***High*** |
|---|---|---|---|---|
| **X1** | 7 | 17 | 8 | 24 |
| **X2** | 48 | 54 | 23 | 13 |

Table 4.1: Distribution of countries attending to socio-economic level

As we can see, richer countries prevail among those where deaths have decreased (almost half of them are rich countries), whereas poorer countries tend to be in the second group, where deaths have gone up.

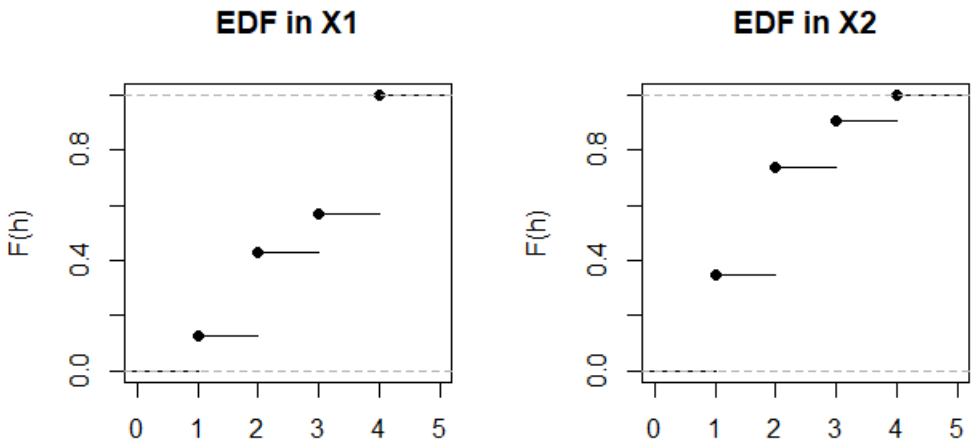
As well as in the previous cases, we need now to determine a suitable statistic test to carry out the experiment. Attending to our aim of determining if categories are equally distributed over both samples, we propose a statistic involving the respective EDFs. This statistic is

$$T = \sum_{k=1}^{h-1} [\hat{F}_1(k) - \hat{F}_2(k)]^2 \cdot (\hat{F}(h)[1 - \hat{F}(h)])^{-1}$$

where $\hat{F}_1$ is the EDF of the first sample, $\hat{F}_2$ the EDF of the second sample, and $\hat{F}$ the joint EDF. Observe that the denominator is permutationally invariant, so attending again to 2.3, we can dismiss this term and use the statistic

$$T = \sum_{k=1}^{h-1} [\hat{F}_1(k) - \hat{F}_2(k)]^2$$

which uniquely compares the squared differences of every class in each sample.




Remark that jumps are the probability of coming across with each class. In the first sample, the biggest gap is the one between the third and the fourth category, that is between upper-middle socio-economic level countries and rich countries. However, this changes in $\mathbf{X}_2$. The biggest jumps here are the one between 0 and the first category, poor countries, and the one between poor and lower-middle socio-economic level countries, meaning a higher presence of poorer countries in the second sample.

Let us compute the $p$-value. We make use of the Monte Carlo algorithm (2.2.2) again to estimate it. Setting $B = 1000$

```
B <- 1000

T <- array(0, dim = c(1,B+1))

acum <- array(0, dim = c(1, h-1))

for (k in 1:(h-1))
{
  acum[k] <- ((Fn1(k) - Fn2(k))^2)
}

T0 <- sum(acum)

T[1] = T0

Z <- rbind(X1, X2)

for (b in 2:B+1)
{
  Xstar <- sample(Z$Income, size = n)

  X1star <- Xstar[1:n1]

  X2star <- Xstar[(n1+1):n2]

  Fn1star <- ecdf(X1star)

  Fn2star <- ecdf(X2star)

  acumstar <- array(0, dim = c(1, h-1))

  for (k in 1:h-1)
  {
    acumstar[k] <- (Fn1star(k) - Fn2star(k))^2
  }

  T0star <- sum(acumstar)

  T[b] <- T0star

}

extremevalues <- T[T>=T0]

pvalue <- length(extremevalues)/B
```

the resulting estimation is

```
> pvalue
[1] 0.001
```

This p-value clearly leads to rejecting the null hypothesis. We can conclude that the socio-economic level of a country plays an meaningful role when it comes

to smoke habits, leading to more deaths in poorer countries.

# Bibliography


[1] Fisher, R.A. (1935) *The Design of Experiments* 27-41

[2] Pitman, E.J.G. (1937) *Significance Tests Which May be Applied to Samples From any Populations* Wiley for the Royal Statistical Society. Supplement to the Journal of the Royal Statistical Society

[3] Wasserman, L. (2006) *All of nonparametric statistics* Springer, Springer Texts in Statistics

[4] Gibbons, J.D. and Chakraborti, S. (2003) *Nonparametric statistical inference* Marcel Dekker

[5] Lehman, E.L. and Romano, J.P.(2008) *Springer, Springer Texts in Statistics*

[6] Hoeffding, W. (1952) *The large-sample power of tests based on permutations of observations* The Annals of Mathematical Statistics

[7] Pesarin, F. and Salmaso, L. (2010) *Permutation Tests for Complex Data: Theory, Applications and Software* John Wiley & Sons, Wiley Series in Probability and Statistics

[8] Taleb, N.N. (2005) *Fooled by randomness: the hidden role of chance in life and in the markets.* New York: Random House

[9] Ritchie, H. and more (2021). *Israel: Coronavirus Pandemic Country Profile* Our World In Data

[10] *PISA test score: Mean performance on the mathematics scale, 2000 to 2015* World Bank EdStats

[11] *Number of deaths from smoking in 1990 vs. 2017* Global Burden of Disease Collaborative Network