Statistical sampling on ordered structures [1]

[1] This work has been awarded the 2019 edition of the INE, Eduardo García España Prize.

Statistical sampling on ordered structures

Pedro García-Segador
National Statistics Institute
Paseo de la Castellana, 183
28071 Madrid (Spain)
pedro.garcia.segador@ine.es

> "Do not fear to be eccentric in opinion, for every opinion now accepted was once eccentric."
>
> — Bertrand Russell

Abstract

This work deals with the problem of using statistical sampling on spaces where no algebraic operations are defined in advance. Sometimes, the target variable is not numeric and finding a good representative value of this variable from a sample is not easy. Clearly, this complexity depends on the nature of the problem and the variables involved. We start discussing a general framework to cope with this kind of problems. For this purpose, we develop some techniques to use statistical sampling on metric spaces. Subsequently, we study the specific case of variables with partially ordered values. These variables represent the set of preferences of a collection of individuals regarding some topic. We benefit from order theory to build well-behaved metrics and provide a fully operational theory to work with partially ordered sets. Finally, we apply these results to some relevant areas of official statistics. Among these applications, we highlight a way of estimating the voting intention in respect of each political party in some election campaigns. Further applications may include improving surveys about habits and opinion, about user satisfaction or even estimating the control structure of enterprise groups.

Keywords: Statistical sampling; metric spaces; posets; official statistics.

Contents

1 Introduction

Any statistical survey is designed to obtain aggregated data from a target population. The variables measuring these data can be classified as qualitative (aka, categorical) or quantitative (aka, numeric). When we work with quantitative variables the task is much easier, since we can measure distances and perform basic arithmetic operations in order to build estimators in the classic way. We also have powerful tools to perform statistical inference with these data.

The qualitative case is more challenging because one should find a convenient way to perform basic algebra and compute distances on the categories associated. Imagine the next situation. A company that manufactures and sells cars, wants to build a new model of car. In order to design this new model, the company asks one hundred people about their favourite car. The interviewed people can choose any car model among history of the car industry. Once the survey is done and the preferences of people are collected, the company starts processing this information in order to find the car model which is most preferred by people. The objective is designing the new car model similar to the preferences of people. Since the sample is just one hundred and there are much more car models, it is highly likely that each car model will appear just once in the sample. Therefore, the statistical mode give us no information about the real preferences. Nor can we compute a mean value, since there is not a natural way of doing arithmetics with cars. Obviously, other approach is needed. Our first goal is to provide a general framework to use sampling theory with qualitative variables. For this purpose, we will endow the target population with a metric space structure.

Subsequently, we are going to work with a specific type of quantitative variable, which are partially ordered sets, or posets for short. Posets are normally understood as a way of giving particular preferences over a set of objects where one person can prefer object A to B or be indifferent. These mathematical structures model a lot of important situations in economy and social sciences, such as consumer preferences, voting intention in respect of different political parties, choossing top priority goals for a company, or any kind of comparison we can imagine. In mathematics, posets are the core structure studied in order theory. Order theory constitutes a powerful tool in many different fields, such as Decision Making, Game Theory, Evidence Theory, Utility theory and many others (see e.g. [17, 14, 16] and the references therein). For this reason, the study of the properties of these structures has attracted the attention of many researchers. Posets play also a fundamental role in the geometrical structure of fuzzy measures [4, 12, 13].

In this work, we are going to develop a wide framework to use statistical sampling with posets. We will start defining how to deal with the sampling problem. We will see from here that one of the key facts consists in defining a well-behaved distance. We will study several kinds of distances and analyze their properties. This is an interesting problem for several reasons. First, and most evident, it is an appealing problem from a mathematical point of view. Next, being able to use statistics on posets allows for estimating several interesting population characteristics in official statistics.

Next, we will apply these results to some relevant areas of official statistics. The examples elaborated in this work illustrate several possible applications of order theory to official statistics, however the presented theory can be applied in many more cases. Considering issues of confidentiality, we will not use real data. Instead, we will use computer simulations inspired on the publicly available data. Firstly, we focus on estimating the voting intention of society by asking people about political preferences. The Spanish Sociological Research Center (CIS) carries out several election surveys. The objective is always to determine the general public opinion about the different political parties. In many cases, this kind of surveys include questions about the order of preference among the different political parties. These preferences are modelled by posets. We will provide techniques to estimate the poset which best represents the opinion of the people. Then, we will explore some uses of this theory into social surveys. Most of social surveys, as for example the fertility survey conducted by the National Statistics Institute (INE), include a lot of multiple choice questions on their questionnaires. Sometimes, people are asked to order some choices according to their preferences. Once again, posets play an important role here. Later, we will focus on a quality question. We will discuss a way of estimating the preferences of official statistics users among European Statistical System dimensions of quality. Here the set of preferences is a poset and we ask about which poset explains the best the preferences of a given sample. Note that each user of official statistics has certain preferences in accordance with their needs. The objective is again to compute a poset that represent as best as possible the general opinion of the population. Moreover, we will provide a way of determining a global satisfaction index by using order theory.

Other interesting application is linked to quality of life indicators. In this case, we will use the theory presented here to build global indicators taking into account multidimensional aspects of quality of life dimensions. Finally, we will study some applications of these ideas to business registers and enterprise statistics. Specificaly, we will discuss a way of comparing the control structure of two groups. Sometimes, we get different information from different sources concerning the control structure of an enterprise group. The goal is computing a poset fitting in an optimal way the enterprise group structure. Moreover, we will develop inference methods to decide if two groups structures are sufficiently dissimilar not to be the same group.

The rest of the work goes as follows: we start reviewing the basic theory of partially ordered sets. In the next section, we deal with the problem of defining a theoretical framework to work with samples in a metric space. We also give inference techniques inside this framework. Next, we move to study different distances on posets that allow us to apply these results about metric spaces to partially ordered sets. Then we explain several important applications of this theory to official statistics. Finally, we conclude with conclusions and open problems. For the sake of being self-contained, we also include two appendices about polyhedra and polytopes (appendix A) and linear and integer programming (appendix B).

Throughout this work, we have used mathematical and statistical software in order to solve some problems or do some simulations. The software used are R Commander, SAS and GAMS.

2 Basic concepts and tools on posets

In this section we review the basics on order theory. For a general introduction on the theory of posets see [6, 18]. Let us consider a finite set P endowed with a binary relation $\preceq$.

Definition 1. *Let P be a set and $\preceq$ be a binary relation over P. We say that P is a* **partially order set** *(or* **poset** *for short) if:*

i) Reflexivity: $x \preceq x$, $\forall x \in P$,

ii) Antisymmetry: If $x \preceq y$ and $y \preceq x$, then $x = y$,

iii) Transitivity: If $x \preceq y$ and $y \preceq z$, then $x \preceq z$.

With some abuse of notation, we will usually omit $\preceq$ and write P instead of $(P, \preceq)$ when referring to posets. Elements of P are denoted x, y and so on, and also $a_1, a_2, ...$. If $|P| = m$, we will also use the notation $P = \{1, ..., m\}$. Subsets of P are usually denoted by capital letters $A, B, ...$. We say that y **covers** x, denoted $x \lessdot y$, if $x \preceq y$ and there is no $z \in P \setminus \{x, y\}$ satisfying $x \preceq z \preceq y$. A poset can be represented through *Hasse diagrams*. Suppose that $(P, \preceq)$ is a poset. We draw a Hasse diagram for this poset by representing each element of P by a distinct point so that whenever $x \preceq y$, the point representing y is situated higher than the point representing x. The converse of this statement is not true in general. Moreover, if y covers x, then we connect the points representing x and y by a straight line segment. Figure 1 shows a Hasse diagram for the poset $(P, \preceq)$ where $P = \{1, 2, 3, 4\}$ and $1 \preceq 4, 1 \preceq 3, 2 \preceq 4$ and $2 \preceq 3$.

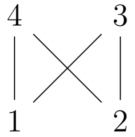

Figure 1: Poset P and its Hasse diagram.

An element x such that $x \not\prec y$, $\forall y \in P$ is called a **maximal element**. Similarly, if x is such that $y \not\prec x, \forall y \in P$, x is called a **minimal element**. For a poset P, we can define the **dual poset** $P^\partial = (P, \preceq_\partial)$ such that $x \preceq_\partial y \Leftrightarrow y \preceq x$.

A **chain** is a poset such that $\preceq$ is a total order, i.e. $x \preceq y$ or $y \preceq x$ for all x, y in the chain. We will denote the chain of m elements by $\mathbf{m}$. The **height** of a poset P, denoted by $h(P)$, is the maximum size of a chain in P. Similarly, an **antichain** is a poset where $\preceq$ is given by $x \preceq y \Leftrightarrow x = y$. We will denote the antichain of m elements by $\overline{m}$. The **width** of a poset P, denoted by $w(P)$, is the maximum size of an antichain in P.

Given an element x, we denote by $\downarrow x$ the subposet of P whose elements are $\{y : y \preceq x\}$; similarly, we denote by $\uparrow x$ the subposet of P whose elements are $\{y : x \preceq y\}$. These notions can be extended for a general subset A, thus obtaining $\downarrow A$ and $\uparrow A$. Finally, we will denote by $\updownarrow A$ the set of elements related to any element of A. An **ideal** or **downset** I of P is a subset of P such that if $x \in I$ then

$\downarrow x \subseteq I$. Symmetrically, a subset F of P is a **filter** or **upset** if for any $x \in F$ and any $y \in P$ such that $x \preceq y$, it follows that $y \in F$.

Two posets $(P, \preceq_P)$ and $(Q, \preceq_Q)$ are **isomorphic** if there is a bijection $f : P \to Q$ such that $x \preceq_P y \Leftrightarrow f(x) \preceq_Q f(y)$. If two posets are isomorphic, then their corresponding Hasse diagrams are the same up to differences in the names of the elements. Two elements $x, y \in P$ are said to be **interchangeable** if there is an automorphism $f : P \to P$ such that $f(x) = y$ and $f(y) = x$.

Given two posets, $(P, \preceq_P), (Q, \preceq_Q)$, their **direct sum**, denoted $P \oplus Q$, is a poset over the referential $P \cup Q$ (disjoint union) and whose partially order $\preceq_{P \oplus Q}$ is defined as follows: if $x, y \in P$ then $x \preceq_{P\oplus Q} y$ if and only if $x \preceq_P y$, if $x, y \in Q$ then $x \preceq_{P\oplus Q} y$ if and only if $x \preceq_Q y$, and if $x \in P, y \in Q$ then $x \preceq_{P\oplus Q} y$. A poset is **irreducible** by direct sum if it cannot be written as a direct sum of two posets. Similarly, the **disjoint union** of two posets $(P, \preceq_P), (Q, \preceq_Q)$, denoted $P \uplus Q$ is a poset $(P \cup Q, \preceq_{P \uplus Q})$ where $x \preceq_{P\uplus Q} y$ whenever $x, y \in P$ and $x \preceq_P y$, or $x, y \in Q$ and $x \preceq_Q y$. A poset which cannot be written as disjoint union of two posets is called **connected**. Obviously, the Hasse diagram of a connected poset is also a connected graph, see Figure 2.

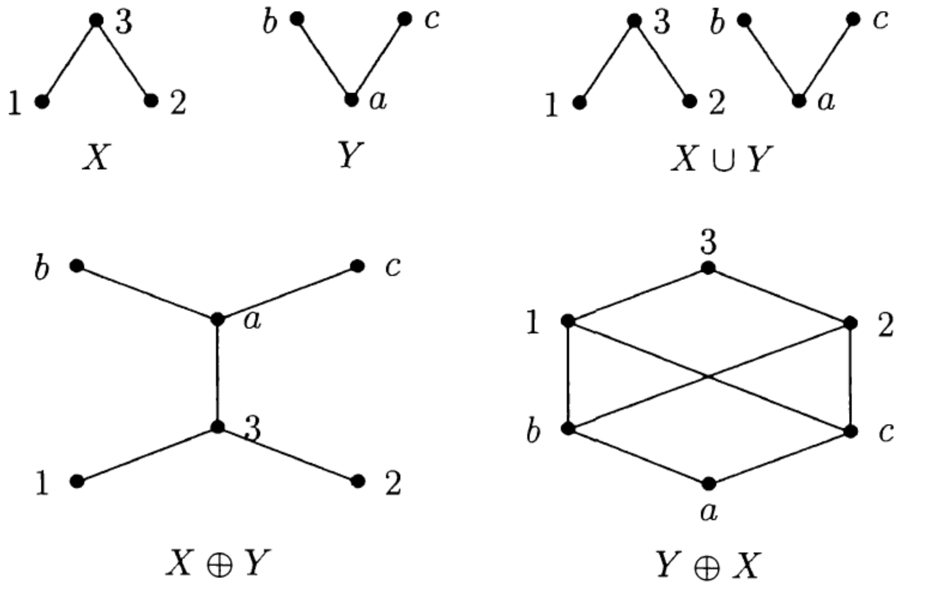

Figure 2: Direct sum and disjoint union of posets X and Y.

Other important operation is the **product order**. Given two posets, $(P, \preceq_P), (Q, \preceq_Q)$, their **product**, denoted by $P \times Q$, is a poset over the cartesian product $P \times Q$ and whose partially ordered $\preceq_{P\times Q}$ is defined as follows: if $(x_1, y_1) \preceq_{P\times Q} (x_2, y_2)$ if and only if $x_1 \preceq_P x_2$ and $y_1 \preceq_Q y_2$, see Figure 3.

A **linear extension** of $(P, \preceq)$ is a sorting of the elements of P that is compatible with $\preceq$, i.e. $x \preceq y$ implies that x is before y in the sorting. Linear extensions will be denoted ϵ_1, ϵ_2 and so on and the i-th element of ϵ is denoted $\epsilon(i)$. We will denote by $\mathcal{L}(P)$ the set of all linear extensions of poset $(P, \preceq)$ and by $e(P) = |\mathcal{L}(P)|$.

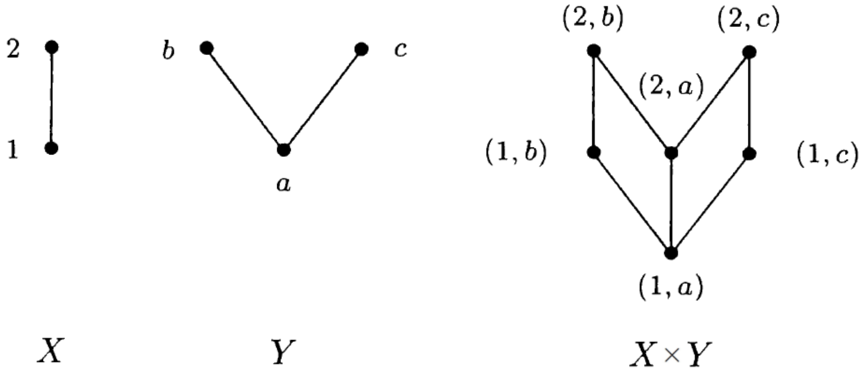

Figure 3: Product order of posets X and Y.

When working with finite posets, it is sometimes convenient to denote elements as natural numbers. A **labeling** is a bijective mapping $L : \{1, 2, \ldots, |P|\} \to P$ (see [18]). There are $m!$ ways to define a labeling. A labeling is **natural** if $x \preceq y$ implies $L^{-1}(x) \leq L^{-1}(y)$ with the natural numbers order. It is well-known that every finite poset admits a natural labeling. A poset endowed with a labeling is called a **labeled poset**. Observe that a labeled poset is simply a poset with different consecutive natural numbers assigned to its elements.

Example 1. *Consider the poset N, given by four elements $1, 2, 3, 4$ and whose corresponding Hasse diagram is given in Figure 4. The linear extensions of this poset N are*

$$(1, 2, 3, 4), (1, 2, 4, 3), (2, 1, 3, 4), (2, 1, 4, 3), (2, 3, 1, 4)$$

Note that we have used a natural labeling for N.

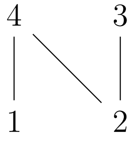

Figure 4: N poset and its Hasse diagram.

It is important to distinguish between posets and labeled posets. Let us show an example to clarify these issues. In addition, the next example exhibits how labeled posets model preferences.

Example 2. *Labeled posets are the most natural way of modelling preferences. Let A and B be two consumers and consider four consumer goods denoted by numbers $1, 2, 3$ and 4. Suppose that the consumer A and consumer B have preferences as the ones shown in Figure 5.*

Note that these two posets are isomorphic but they are not the same as labeled posets. In other words, they have the same shape but not the same numbers. So they are different as labeled posets. From the last Figure we can see how consumer A prefers 2 to 1 and is indifferent between 3 and 4. On the other hand, consumer B prefers 4 to 3 and is indefferent between 1 and 2. Despite the two consumers have isomorphic preference posets, their preferences are very different to each other.

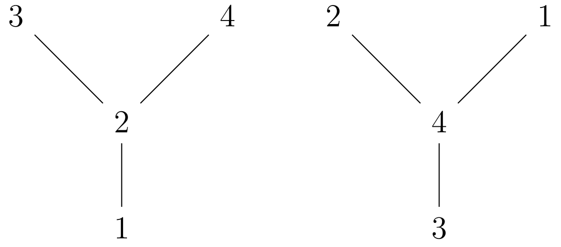

Figure 5: Labeled posets showing the preferences of consumers A and B.

In this paper, we will always work with labeled posets, that is, not only the shape but also the position of the numbers is important. For convenience, we will call posets to labeled posets from now on. The number of posets with m labeled elements, $p(m)$, grows exponentially. The following table reveals the first values.

m	1	2	3	4	5	6	7	8	9
$p(m)$	1	3	19	219	4231	130023	6129859	431723379	44511042511

Obtaining closed formulas counting the number of labeled posets with m elements is an open problem in mathematics. To learn more about this problem and enumerative combinatorics see [23].

In graph theory, a labeled poset is called a labeled acyclic transitive digraph. Indeed, the Hasse diagram of a poset gives its representation as a directed graph, with all arrows up. For this reason, we can define the adjacency matrix in the same way as it is done in graph theory.

Definition 2. *Let P be a labeled poset with m elements. The* **adjacency matrix** *of P, denoted by M_P, is defined to be the square $m \times m$ matrix such that*

$$M_P(i,j) = \begin{cases} 1, & \textit{if } i \prec j \\ 0, & \textit{otherwise.} \end{cases}$$

Example 3. *Consider the preference matrices of consumers A and B in Example 2. By abuse of notation, we use the same letters to name the two associated posets. Their adjacency matrices are*

$$M_A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad M_B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

3 Sampling theory on metric spaces

As stated above, we are going to work with posets. Posets as many other qualitative variables have not a natural way of performing arithmetic computations with them. The impossibility of using this algebraic properties disables the classical sampling theory. Moreover, endowing a set with an algebraic struture with good mathematical and practical attributes is usually complicated. It is much simpler to define a metric on the set of elements of our population. This is the path we are going to follow in this paper. This section is devoted to provide a powerful framework to use classical sampling theory on a set together with a distance map. For an overview of sampling theory see [22].

Definition 3. *A* **metric space** *is an ordered pair* $(\boldsymbol{X}, d)$ *where* $\boldsymbol{X}$ *is a set and* $d : \boldsymbol{X} \times \boldsymbol{X} \to \mathbb{R}^+$ *is a* **metric** *(or* **distance***) on* $\boldsymbol{X}$ *such that for any* $x, y, z \in \boldsymbol{X}$

1. *Identity of indiscernibles:* $d(x, y) = 0 \Leftrightarrow x = y$*,*

2. *Symmetry:* $d(x, y) = d(y, x)$*,*

3. *Triangle inequality:* $d(x, z) \leq d(x, y) + d(y, x)$*.*

When only 1. and 2. are verified the function d is called **semimetric** and the pair $(\boldsymbol{X}, d)$ is called a **semimetric space.** Most of the ideas we will see below hold for semimetric spaces. The **discrete metric**, where $d_D(x, y) = 0$ if $x = y$ and $d_D(x, y) = 1$ otherwise, is a simple but important example, since for any set $\boldsymbol{X}$ the pair $(\boldsymbol{X}, d_D)$ is always a metric space.

Unless otherwise specified, in this work the set $\boldsymbol{X}$ will be a finite set. Let us call $\boldsymbol{X}_P = \{X_1, \ldots, X_N\}$ to our target population of size N, where $X_i \in \boldsymbol{X}$. Note that $\boldsymbol{X}_P$ can have repeated values. Any ordered subset of $\boldsymbol{X}_P$, $S = \{X_{i_1}, X_{i_2}, \ldots X_{i_n}\}$, with possibly repeated values, will be called a **sample** of size n.

As seen in the introduction, the statistical mode does not always have good properties. Particularly, the statistical mode is not informative when all the values in the sample are different to each other. In these cases, other representative value of the target population is needed. Notwithstanding the foregoing, it is always good having alternative measures of position for the target population. Our first goal is to define the population mean $\overline{X}$ of our target population $\boldsymbol{X}_P$. The first attempt will be defining $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ as usual. However, we cannot compute the sum of two elements of $\boldsymbol{X}$. Then to define it we are going to use an elemental but important property of the standard mean.

Proposition 1. *Let* $X_1, \ldots, X_r$ *be real numbers. Then the function* $f : \mathbb{R} \to \mathbb{R}, y \to \sum_{i=1}^{r}(X_i - y)^2$*, reachs a global minimum in the mean value* $\frac{1}{N}\sum_{i=1}^{N} X_i$*.*

So, a way of defining the mean of a set of real values $X_1, \ldots, X_r$ is by stating that it is the value minimizing the sum of the squared distances. We use this philosophy to define the population mean $\overline{X}$ in a finite metric space.

Definition 4. *Let* $(\boldsymbol{X}, d)$ *be a finite metric space and* $\boldsymbol{X}_P = \{X_1, \ldots, X_N\}$ *be the target population then*

$$\overline{X} := \arg\min_{y \in \boldsymbol{X}} \sum_{i=1}^{N} d^2(X_i, y).$$

Note that the minimum of the last definition may not be unique. Therefore, the population mean $\overline{X}$ in a metric space will be a subset of $\boldsymbol{X}$. Defining the variance is easier. Since for numeric data we have $V(X) := \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$ we replace the squared euclidean distance by our metric d.

Definition 5. *Let $(\boldsymbol{X}, d)$ be a finite metric space and $\boldsymbol{X}_P = \{X_1, \dots, X_N\}$ be the target population then*

$$V(\boldsymbol{X}_P) := \frac{1}{N}\sum_{i=1}^{N} d^2(X_i, \overline{X}).$$

We are considering here that the distance between two sets is the minimum of the possible distances between the elements of the two sets, i.e. $d(A, B) = \min_{a\in A, b\in B} d(a, b),\ \forall A, B \subseteq \boldsymbol{X}$.

From now on, we will use that the distance between two sets is the minimum of the possible distances between the elements of the two sets. Observe that $V(\boldsymbol{X}_P) \in \mathbb{R}^+$. Once we have defined the most important population values, we must introduce the probability spaces. We use the standard definitions.

Definition 6. *Let Ω be a set. Then $\mathcal{F} \subseteq 2^\Omega$ is called* σ-**algebra** *if*

1. $\Omega \in \mathcal{F}$,
2. *If* $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$,
3. *If we have a countable collection* $\{A_i\}_{i=1}^{\infty}$ *such that* $A_i \in \mathcal{F}, \forall i \in \mathbb{N}$ *then* $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The elements of $\mathcal{F}$ are called events. The pair $(\Omega, \mathcal{F})$ is called a **measurable space**.

Definition 7. *Let $(\Omega, \mathcal{F})$ be a measurable space then $\boldsymbol{P} : \mathcal{F} \to [0, 1]$ is a* **probability measure** *if*

1. $\boldsymbol{P}(\Omega) = 1$,
2. *If we have a countable collection* $\{A_i\}_{i=1}^{\infty}$ *such that* $A_i \in \mathcal{F}, \forall i \in \mathbb{N}$ *and* $A_i \cap A_j = \emptyset$ *for* $i \neq j$, *then also* $\boldsymbol{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \boldsymbol{P}(A_i)$.

This set of conditions are usually known as Kolmogorov axioms.

Definition 8. *A* **probability space** *is a triple $(\Omega, \mathcal{F}, \boldsymbol{P})$ where Ω is the set of possible outcomes, $\mathcal{F}$ is a σ-algebra and $\boldsymbol{P}$ is a probability measure function.*

Let $\Omega = \boldsymbol{X}, \mathcal{F} = 2^{\boldsymbol{X}}$ and $\boldsymbol{P}$ any probability mass function. Then $(\boldsymbol{X}, 2^{\boldsymbol{X}}, \boldsymbol{P})$ is a probability space over the metric space $\boldsymbol{X}$. Note that the last construction is completely standard. In a similar way we can use random variables associated to $\boldsymbol{X}$.

Definition 9. *Let $(\Omega, \mathcal{F})$ and $(E, \mathcal{G})$ be two measurable spaces. A* **random variable** *is a function* $\mathrm{X} : \Omega \to E$ *such that* $\mathrm{X}^{-1}(G) \in \mathcal{F},\ \forall G \in \mathcal{G}$.

Making $\Omega = \boldsymbol{X}$ and $E = \mathbb{R}$ we get real random variables on $\boldsymbol{X}$. However, we can also consider $E = \boldsymbol{X}$ to get random variables with values on $\boldsymbol{X}$.

In statistical sampling theory, we work on the probability space $(\mathcal{S}, 2^{\mathcal{S}}, \boldsymbol{P})$ where $\mathcal{S}$ is the set of possible samples drawn from the population $\boldsymbol{X}_P$ and $\boldsymbol{P}$ is a probability measure on $\mathcal{S}$. Usually, $(\mathcal{S}, 2^{\mathcal{S}}, \boldsymbol{P})$ is called a **sampling design**. Now, we are in conditions to define an estimator.

Definition 10. *Let $(\boldsymbol{X}, d)$ be a metric space and $(\mathcal{S}, 2^{\mathcal{S}}, \boldsymbol{P})$ be a sampling design, an* **estimator** *in the metric space $\boldsymbol{X}$ is just a function $\widehat{\theta} : \mathcal{S} \to 2^{\boldsymbol{X}}$, such that $\widehat{\theta}(S) \neq \emptyset$, $\forall S \in \mathcal{S}$.*

The sampling distribution of an estimator $\widehat{\theta} : \mathcal{S} \to 2^{\boldsymbol{X}}$ is just

$$\boldsymbol{P}(\widehat{\theta} = \Theta) := \sum_{S \in \mathcal{S},\ \widehat{\theta}(S) = \Theta} \boldsymbol{P}(S),$$

for all $\Theta \in 2^{\boldsymbol{X}}$. Similarly,

$$\boldsymbol{P}(t \in \widehat{\theta}) := \sum_{S \in \mathcal{S},\ t \in \widehat{\theta}(S)} \boldsymbol{P}(S),$$

for all $t \in \boldsymbol{X}$. When a decision rule $R : 2^{\boldsymbol{X}} \to \boldsymbol{X}$ is available we can define point estimators as $R(\widehat{\theta}) : \mathcal{S} \to \boldsymbol{X}$.

The next sensible step would be defining the expected value and the variance of an estimator. As first attempt we could try $\mathbb{E}[\widehat{\theta}] = \sum_{S \in \mathcal{S}} \boldsymbol{P}(S)\widehat{\theta}(S)$. Nevertheless, we cannot sum elements in $\boldsymbol{X}$. To deal with this drawback, we focus on the next property of real numbers.

Proposition 2. *Let $X_1, \ldots, X_r$ be real numbers and $\boldsymbol{P}$ a probability on $\{1, 2, \ldots, r\}$. Then the function $f : \mathbb{R} \to \mathbb{R}, y \to \sum_{i=1}^{r} \boldsymbol{P}(i)(X_i - y)^2$, reachs a global minimum in the weighted mean value $\sum_{i=1}^{N} \boldsymbol{P}(i) X_i$.*

Following the same philosophy as above, we can define the expected value as the element minimizing the last weighted sum of distances.

Definition 11. *Let $(\boldsymbol{X}, d)$ be a metric space, $(\mathcal{S}, 2^{\mathcal{S}}, \boldsymbol{P})$ be a sampling design, and $\widehat{\theta} : \mathcal{S} \to 2^{\boldsymbol{X}}$ be an estimator. We define the expected value as*

$$\mathbb{E}[\widehat{\theta}] := \arg\min_{y \in \boldsymbol{X}} \sum_{S \in \mathcal{S}} \boldsymbol{P}(S) d^2(\widehat{\theta}(S), y).$$

By definition, $\mathbb{E}[\widehat{\theta}]$ is a subset of $\boldsymbol{X}$. So the expected value is not necessarily unique. Since the classical definitions for variance, bias and mean squared error (MSE) of $\widehat{\theta}$ are respectively

$$V(\widehat{\theta}) := \sum_{S \in \mathcal{S}} \boldsymbol{P}(S) \left(\widehat{\theta}(S) - \mathbb{E}[\widehat{\theta}]\right)^2,$$

$$B(\widehat{\theta}) := \mathbb{E}[\widehat{\theta}] - \overline{X},$$

$$MSE(\widehat{\theta}) := \sum_{S \in \mathcal{S}} \boldsymbol{P}(S) \left(\widehat{\theta}(S) - \overline{X}\right)^2.$$

We define the following adaptations to metric spaces.

Definition 12. *Let $(\boldsymbol{X}, d)$ be a metric space, $(\mathcal{S}, 2^{\mathcal{S}}, \boldsymbol{P})$ be a sampling design, and $\widehat{\theta} : \mathcal{S} \to 2^{\boldsymbol{X}}$ be an estimator. We define the variance, bias and mean squared error (MSE) of $\widehat{\theta}$ respectively as*

$$V(\widehat{\theta}) := \sum_{S \in \mathcal{S}} \boldsymbol{P}(S) d^2(\widehat{\theta}(S), \mathbb{E}[\widehat{\theta}]),$$

$$B(\widehat{\theta}) := d(\overline{X}, \mathbb{E}[\widehat{\theta}]),$$

$$MSE(\widehat{\theta}) := \sum_{S \in \mathcal{S}} \boldsymbol{P}(S) d^2(\widehat{\theta}(S), \overline{X}).$$

These three measures are always positive real numbers. In a similar way, we can speak about expressions like $\boldsymbol{P}\left(d(\widehat{\theta}_n, \overline{X}) \geq \epsilon\right)$. If $\widehat{\theta}_n : \mathcal{S}_n \to 2^{\boldsymbol{X}}$ is an estimator, the sample size is n and $\theta \subset \boldsymbol{X}$, we say

$$\boldsymbol{P}\left(d(\widehat{\theta}_n, \theta) \geq \epsilon\right) = \sum_{S \in \mathcal{A}} \boldsymbol{P}(S),$$

where $\mathcal{A} = \{S \in \mathcal{S}_n \mid d(\widehat{\theta}_n(S), \theta) \geq \epsilon\}$. With these concepts in mind we can give a definition for an estimator to be unbiased and consistent.

Definition 13. *Let $(\boldsymbol{X}, d)$ be a metric space, $(\mathcal{S}_n, 2^{\mathcal{S}_n}, \boldsymbol{P})$ be a sampling design, with sample size n, and $\widehat{\theta}_n : \mathcal{S}_n \to 2^{\boldsymbol{X}}$ be an estimator. We will say that $\widehat{\theta}_n : \mathcal{S}_n \to 2^{\boldsymbol{X}}$ is* **unbiased** *if $\mathbb{E}[\widehat{\theta}_n] \cap \overline{X} \neq \emptyset$. We say that $\widehat{\theta}_n$ is* **consistent** *if $\forall \epsilon > 0$*

$$\lim_{n \to \infty} \boldsymbol{P}\left(d(\widehat{\theta}_n, \overline{X}) \geq \epsilon\right) = 0.$$

Let $(\mathcal{S}_n, 2^{\mathcal{S}_n}, \boldsymbol{P})$ be a sampling design with sample size n. Consider the random variable $N_n^i : (\mathcal{S}_n, 2^{\mathcal{S}_n}) \longrightarrow (\{0, 1, \ldots, n\}, 2^{\{0,1,\ldots,n\}})$ counting the number of times that the element $i \in \boldsymbol{X}$ appears in a sample $S_n \in \mathcal{S}_n$. We can also count the number of times that the element $i \in \boldsymbol{X}$ appears in the target population $\boldsymbol{X}_P$, we call it $N^i := N_N^i(\boldsymbol{X}_P)$.

Definition 14. *Let $(\mathcal{S}_n, 2^{\mathcal{S}_n}, \boldsymbol{P})$ be a sampling design with sample size n and N_n^i be a map counting the number of repetitions of element $i \in \boldsymbol{X}$ in a sample $S_n \in \mathcal{S}_n$. We say that the sampling design is* **asymptotically consistent** *if $\forall i \in \boldsymbol{X}$ we have*

$$\frac{N_n^i}{n} \xrightarrow{\mathcal{L}} \frac{N^i}{N} L_i \Leftrightarrow \lim_{n \to \infty} \boldsymbol{P}\left(\frac{N_n^i}{n} = \frac{N^i}{N} L_i\right) = 1,$$

where $N^i := N_N^i(\boldsymbol{X}_P)$ and $L_i \in \mathbb{N}$. In other words, $\dfrac{N_n^i}{n}$ converges in distribution to a constant value $\dfrac{N^i}{N} L_i$, for every $i \in \boldsymbol{X}$. Besides, if $L_i = 1\ \forall i \in \boldsymbol{X}$ we say that the sampling design has **asymtotically equal probabilities of selection**.

A very common and important estimator in statistical sampling is the sample mean. To define it in a setting of metric spaces we can apply the same philosophy as above and define the sample mean estimator as the one minimizing the sum of squared distances. We will call it the Horvitz-Thompson estimator.

Definition 15. *Let $(\boldsymbol{X}, d)$ be a finite metric space and $(\mathcal{S}_n, 2^{\mathcal{S}_n}, \boldsymbol{P})$ be a sampling design with sample size n. Let $S = \{X_1, \dots, X_n\}$ be a sample of size n. Then the* **Horvitz-Thompson estimator** *for the population mean $\overline{X}$ is defined as*

$$\widehat{\overline{X}}_{HT}(S) := \arg\min_{y \in \boldsymbol{X}} \sum_{i=1}^{n} d^2(X_i, y).$$

It is not difficult to show that this estimator is consistent for sampling designs having asymptotically equal probabilities of selection. For this purpose, we will use the next technical Lemma.

Lemma 1. *Let $(\boldsymbol{X}, d)$ be a finite metric space and $f_n : \boldsymbol{X} \to \mathbb{R}$ such that for all $x \in \boldsymbol{X}$, $\lim_{n\to\infty} f_n(x) = \tilde{f}(x)$. Then the following finite time convergence holds*

$$d\left(\arg\min_{x\in\boldsymbol{X}} f_n(x), \arg\min_{x\in\boldsymbol{X}} \tilde{f}(x)\right) \xrightarrow[n\to\infty]{} 0.$$

Proof. For all $x \in \boldsymbol{X}$, $\lim_{n\to\infty} f_n(x) = \tilde{f}(x)$ then $\forall \epsilon_x > 0,\ \exists n_x \in \mathbb{N}$ such that $|f_m(x) - \tilde{f}(x)| < \epsilon_x$ for all $m > n_x$. Denote $\epsilon = \max_{x\in\boldsymbol{X}} \epsilon_x$ and

$$\delta = \min_{\substack{|\tilde{f}(x)-\tilde{f}(y)|>0 \\ x,y\in\boldsymbol{X}}} |\tilde{f}(x) - \tilde{f}(y)|.$$

Let $\epsilon = \frac{\delta}{2}$ and take m bigger than the maximum $n = \max_{x\in\boldsymbol{X}} n_x$. Now observe that if $\tilde{f}(x) < \tilde{f}(y)$ then $f_m(x) < f_m(y)$ for all $m > n$. Indeed,

$$f_m(x) < \tilde{f}(x) + \frac{\delta}{2} \le \tilde{f}(y) - \frac{\delta}{2} < f_m(y).$$

Therefore, if $\arg\min_{x\in\boldsymbol{X}} \tilde{f}(x) = \Theta \subseteq \boldsymbol{X}$ then $\forall m > n$, $\exists\, x_m^* \in \Theta$ such that $x_m^* \in \arg\min_{x\in\boldsymbol{X}} f_m(x)$. Thus

$$d\left(\arg\min_{x\in\boldsymbol{X}} f_n(x), \arg\min_{x\in\boldsymbol{X}} \tilde{f}(x)\right) \xrightarrow[n\to\infty]{} 0,$$

in finite time. □

Proposition 3. *Let $(\boldsymbol{X}, d)$ be a finite metric space and $(\mathcal{S}_n, 2^{\mathcal{S}_n}, \boldsymbol{P})$ be a sampling design with asymtotically equal probabilities of selection. Then $\widehat{\overline{X}}_{HT}$ is a consistent estimator for the population mean $\overline{X}$.*

Proof. Let $\mathcal{A}_n = \{S_n \in \mathcal{S}_n \mid d(\widehat{\overline{X}}_{HT}(S_n), \overline{X}) \geq \epsilon\}$. Observe that

$$\widehat{\overline{X}}_{HT}(S_n) = \arg\min_{y\in\boldsymbol{X}} \sum_{i\in\boldsymbol{X}} N_n^i(S_n)\cdot d^2(i,y) = \arg\min_{y\in\boldsymbol{X}} \sum_{i\in\boldsymbol{X}} \frac{N_n^i(S_n)}{n}\cdot d^2(i,y),$$

and

$$\overline{X} = \arg\min_{y\in\boldsymbol{X}} \sum_{i\in\boldsymbol{X}} N^i\cdot d^2(i,y) = \arg\min_{y\in\boldsymbol{X}} \sum_{i\in\boldsymbol{X}} \frac{N^i}{N}\cdot d^2(i,y).$$

By Lemma 1

$$f_n(y) = \sum_{i\in\boldsymbol{X}} \frac{N_n^i(S_n)}{n}\cdot d^2(i,y) \xrightarrow[n\to\infty]{} \sum_{i\in\boldsymbol{X}} \frac{N^i}{N}\cdot d^2(i,y) = \tilde{f}(y)$$

therefore $d(\widehat{\overline{X}}_{HT}(S_n), \overline{X}) \xrightarrow[n\to\infty]{} 0$ and $\mathcal{A}_n \xrightarrow[n\to\infty]{} \emptyset$. Hence,

$$\boldsymbol{P}\left(d(\widehat{\overline{X}}_{HT}(S_n), \overline{X}) \geq \epsilon\right) = \sum_{S_n\in\mathcal{A}_n} \boldsymbol{P}(S_n) \to 0.$$

□

The last proposition just works for asymtotically equal probabilities of selection. However, it could be modified to get consistent estimators when the probabilities of selection are not asymtotically equal. To this end, we can define the generalized Horvitz-Thompson estimator for the population mean $\overline{X}$.

Definition 16. *Let $(\boldsymbol{X}, d)$ be a finite metric space and $(\mathcal{S}_n, 2^{\mathcal{S}_n}, \boldsymbol{P})$ be an asymptotically consistent sampling design with $\frac{N_n^i}{n} \xrightarrow{\mathcal{L}} \frac{N^i}{N}L_i$. Let $S = \{X_1, \ldots, X_n\}$ be a sample of size n. The* **Generalized Horvitz-Thompson estimator** *for the population mean $\overline{X}$ is defined as*

$$\widehat{\overline{X}}_{GHT}(S) := \arg\min_{y\in\boldsymbol{X}} \sum_{i=1}^{n} \frac{d^2(X_i,y)}{L_i}.$$

The Generalized Horvitz-Thompson estimator is consistent for asymptotically consistent sampling designs.

Proposition 4. *Let $(\boldsymbol{X}, d)$ be a finite metric space and $(\mathcal{S}_n, 2^{\mathcal{S}_n}, \boldsymbol{P})$ be a consistent sampling design with $\frac{N_n^i}{n} \xrightarrow{\mathcal{L}} \frac{N^i}{N}L_i$. Then $\widehat{\overline{X}}_{GHT}$ is a consistent estimator for the population mean $\overline{X}$.*

Proof. Let $\mathcal{A}_n = \{S_n \in \mathcal{S}_n \mid d(\widehat{\overline{X}}_{GHT}(S_n), \overline{X}) \geq \epsilon\}$. Observe that

$$\widehat{\overline{X}}_{GHT}(S_n) = \arg\min_{y\in\boldsymbol{X}} \sum_{i\in\boldsymbol{X}} N_n^i(S_n)\cdot \frac{d^2(i,y)}{L_i} = \arg\min_{y\in\boldsymbol{X}} \sum_{i\in\boldsymbol{X}} \frac{N_n^i(S_n)}{n}\cdot \frac{d^2(i,y)}{L_i},$$

and

$$\overline{X} = \arg\min_{y\in\boldsymbol{X}} \sum_{i\in\boldsymbol{X}} N^i \cdot d^2(i,y) = \arg\min_{y\in\boldsymbol{X}} \sum_{i\in\boldsymbol{X}} \frac{N^i}{N} \cdot d^2(i,y).$$

By Lemma 1

$$f_n(y) = \sum_{i\in\boldsymbol{X}} \frac{N_n^i(S_n)}{n} \cdot \frac{d^2(i,y)}{L_i} \xrightarrow[n\to\infty]{} \sum_{i\in\boldsymbol{X}} \frac{N^i}{N} \cdot d^2(i,y) = \tilde{f}(y)$$

therefore $d(\widehat{\overline{X}}_{GHT}(S_n), \overline{X}) \xrightarrow[n\to\infty]{} 0$ and $\mathcal{A}_n \xrightarrow[n\to\infty]{} \emptyset$. Hence,

$$\boldsymbol{P}\left(d(\widehat{\overline{X}}_{GHT}(S_n), \overline{X}) \geq \epsilon\right) = \sum_{S_n\in\mathcal{A}_n} \boldsymbol{P}(S_n) \to 0.$$

□

In the classic schemes of sampling with replacement and without replacement and different probabilities we can also modify $\widehat{\overline{X}}_{HT}$ to get a consistent estimator.

Proposition 5. *Let $(\boldsymbol{X}, d)$ be a finite metric space, $\boldsymbol{X}_P = \{X_1, \ldots, X_N\}$ be the target population, and $(\mathcal{S}_n, 2^{\mathcal{S}_n}, \boldsymbol{P})$ be a sampling design. Then,*

i) In a sampling with replacement scheme such that the probability of selecting element k is p_k, $\forall k \in \{1, \ldots . N\}$, the **weighted Hansen-Horvitz estimator** *for the population mean*

$$\widehat{\overline{X}}_{WHH}(S) := \arg\min_{y\in\boldsymbol{X}} \sum_{i=1}^{n} \frac{d^2(X_i, y)}{p_i},$$

is consistent.

ii) In a sampling without replacement scheme such that the probability that element k is drawn at least once is given by π_k, $\forall k \in \{1, \ldots . N\}$, the **weighted Horvitz-Thompsom estimator** *for the population mean*

$$\widehat{\overline{X}}_{WHT}(S) := \arg\min_{y\in\boldsymbol{X}} \sum_{i=1}^{n} \frac{d^2(X_i, y)}{\pi_i},$$

is consistent.

Proof. i) Note that when n tends to infinity the proportion of elements X_i in the sample tends to p_i. Let $N_n^{X_i}$ be the number of times that the individual X_i is in the sample. By Lemma 1 since

$$f_n(y) = \sum_{X_i\in\boldsymbol{X}_P} \frac{N_n^{X_i}}{n} \cdot \frac{d^2(X_i, y)}{p_i} \xrightarrow[n\to\infty]{} \sum_{X_i\in\boldsymbol{X}_P} p_i \cdot \frac{d^2(X_i, y)}{p_i} = \tilde{f}(y),$$

$\overline{X} = \arg\min_{y\in\boldsymbol{X}} \sum_{X_i\in\boldsymbol{X}_P} d^2(X_i,y)$ and $\widehat{\overline{X}}_{WHH}(S) = \arg\min_{y\in\boldsymbol{X}} \sum_{X_i\in\boldsymbol{X}_P} \dfrac{N_n^{X_i}}{n}\cdot\dfrac{d^2(X_i,y)}{p_i}$ the consistency holds.

ii) Since the sampling scheme is without replacement, n tends to N and π_i tends to 1. Therefore by Lemma 1

$$f_n(y) = \sum_{i=1}^{n} \frac{d^2(X_i,y)}{\pi_i} \xrightarrow[n\to\infty]{} \sum_{i=1}^{N} \frac{d^2(X_i,y)}{1} = \tilde{f}(y),$$

then the consistency holds.

□

Note the similarity between the last result and the classic estimators: $\sum_{i=1}^n \frac{X_i}{np_i}$ and $\sum_{i=1}^n \frac{X_i}{\pi_i}$. It is not difficult to see that $\widehat{\overline{X}}_{HT}$ is not unbiased in general. It does depend on the sampling design, the probabilities and the metric. However, some partial results hold.

Proposition 6. *Let $(\boldsymbol{X},d)$ be a finite metric space and $(\mathcal{S}_n, 2^{\mathcal{S}_n}, \boldsymbol{P})$ be a sampling design, with sample size n. If the probabilities are the same for each sample, i.e. $\boldsymbol{P}(S) = \boldsymbol{P}(S')$, $\forall S,S'\in\mathcal{S}_n$, and the sample size is $n=1$ then $\widehat{\overline{X}}_{HT}$ is unbiased.*

Proof. If the sample is $\{X_i\} \Rightarrow \widehat{\overline{X}}_{HT} = \arg\min_{y\in\boldsymbol{X}} d^2(X_i,y) = X_i$. Now,

$$\mathbb{E}[\widehat{\overline{X}}_{HT}] = \arg\min_{y\in\boldsymbol{X}} \sum_{i=1}^{N} \boldsymbol{P}(X_i) d^2(\widehat{\overline{X}}_{HT}(X_i),y) = \arg\min_{y\in\boldsymbol{X}} \sum_{i=1}^{N} \boldsymbol{P}(X_i) d^2(X_i,y) = \arg\min_{y\in\boldsymbol{X}} \sum_{i=1}^{N} d^2(X_i,y) = \overline{X}.$$

In the last equality we have used that $\boldsymbol{P}(X_i)$ is constant.

□

Proposition 7. *Let $(\boldsymbol{X},d)$ be a finite metric space and $(\mathcal{S}, 2^{\mathcal{S}}, \boldsymbol{P})$ be a sampling design. We say that $(\boldsymbol{X},d)$ is starry at $z\in\boldsymbol{X}$ if $d(x,z)\le d(x,y)$ $\forall x,y\in\boldsymbol{X}\setminus\{z\}$. If $(\boldsymbol{X},d)$ is starry at z, then $\widehat{\overline{X}}_{HT}$ is unbiased and $z\in\mathbb{E}[\widehat{\overline{X}}_{HT}]$.*

Proof. By definition, $\mathbb{E}[\widehat{\overline{X}}_{HT}] = \arg\min_{y\in\boldsymbol{X}} \sum_{S\in\mathcal{S}} \boldsymbol{P}(S) d^2(\widehat{\overline{X}}_{HT}(S),y)$. Since $(\boldsymbol{X},d)$ is starry at z, we get for all $y\in\boldsymbol{X}$

$$\sum_{S\in\mathcal{S}} \boldsymbol{P}(S) d^2(\widehat{\overline{X}}_{HT}(S),y) \ge \sum_{S\in\mathcal{S}} \boldsymbol{P}(S) d^2(\widehat{\overline{X}}_{HT}(S),z).$$

Then $z\in\mathbb{E}[\widehat{\overline{X}}_{HT}]$. Moreover, $\overline{X} = \arg\min_{y\in\boldsymbol{X}} \sum_{i=1}^N d^2(X_i,y)$ and $z\in\overline{X}$. Finally, $z\in\overline{X}\cap\mathbb{E}[\widehat{\overline{X}}_{HT}]\neq\emptyset$ and $\widehat{\overline{X}}_{HT}$ is unbiased. □

In metric spaces we also have a law of large numbers.

Theorem 1 (**Law of Large Numbers**)**.** *Let $(\boldsymbol{X}, d)$ be a finite metric space and $(\mathcal{S}_n, 2^{\mathcal{S}_n}, \boldsymbol{P})$ be a sampling design with sample size n. Now, take $S_1, \ldots, S_k$ independent trials of $(\mathcal{S}_n, \boldsymbol{P})$. If $\widehat{\theta} : \mathcal{S}_n \to 2^{\boldsymbol{X}}$ is an estimator, then for any $\epsilon > 0$,*

$$\boldsymbol{P}\left(d(\widehat{M_k}, \mathbb{E}[\widehat{\theta}]) \geq \epsilon\right) \xrightarrow[k\to\infty]{} 0,$$

where

$$\widehat{M_k} = \arg\min_{y\in\boldsymbol{X}} \sum_{i=1}^{k} d^2(\widehat{\theta}(S_i), y).$$

Proof. Since k tends to infinity and there is a finite number of different samples, eventually a time will come when every sample will appear in proportion to its probability $\boldsymbol{P}(S)$. Let N_k^S be the number of times that the sample S is drawn among the k samples. Observe that $\widehat{M_k} = \arg\min_{y\in\boldsymbol{X}} \sum_{S\in\mathcal{S}_n} \frac{N_k^S}{k} d^2(\widehat{\theta}(S), y)$. Applying Lemma 1

$$f_k(y) = \sum_{S\in\mathcal{S}_n} \frac{N_k^S}{k} d^2(\widehat{\theta}(S), y) \xrightarrow[k\to\infty]{} \sum_{S\in\mathcal{S}_n} \boldsymbol{P}(S) d^2(\widehat{\theta}(S), y) = \tilde{f}(y).$$

Finally, $d(\widehat{M_k}, \mathbb{E}[\widehat{\theta}]) \xrightarrow[k\to\infty]{} 0$ and the result follows.

□

Finally, we are going to provide some methodology to estimate the variance and the bias of an estimator. We will use a bootstrapping resampling technique. From our original sample S we resample B bootstrap samples of size n with replacement. We denote these bootstrap samples by BS_i, $i \in \{1, \ldots, B\}$. The more bootstrap samples we take, the better the estimation becomes. For a numeric estimator $\widehat{\theta}$, the classic bootstrap use the following estimations:

$$\widehat{B}(\widehat{\theta}) = \frac{1}{B}\sum_{i=1}^{B} \widehat{\theta}(BS_i) - \widehat{\theta}(S),$$

$$\widehat{V}(\widehat{\theta}) = \frac{1}{B}\sum_{i=1}^{B} (\widehat{\theta}(BS_i) - \widehat{\theta}_B(S))^2,$$

where $\widehat{\theta}_B(S) = \frac{1}{B}\sum_{i=1}^{B} \widehat{\theta}(BS_i)$. The classic Law of Large Numbers tell us that, if our simulation times B is large enough, the bootstrap variance estimation is a good approximation for the variance. The idea of the bias approximation is that $\widehat{\theta}_B(S) \approx \mathbb{E}[\widehat{\theta}]$ by the law of large numbers and $\widehat{\theta}(S) \approx \overline{X}$. For a good performance of bootstrap method, it is also necessary the convergence of the empirical distribution to the true distribution function when sample size is large, see [8].

For metric spaces, let $\widehat{\theta} : \mathcal{S} \to 2^{\boldsymbol{X}}$ be an estimator, the bootstrap estimations are as follows:

$$\widehat{B}(\widehat{\theta}) = d(\widehat{\theta}_B(S), \widehat{\theta}(S)),$$

$$\widehat{V}(\widehat{\theta}) = \frac{1}{B}\sum_{i=1}^{B} d^2(\widehat{\theta}(BS_i), \widehat{\theta}_B(S)),$$

where $\widehat{\theta}_B(S) = \arg\min_{y\in \boldsymbol{X}} \sum_{i=1}^{B} d^2(\widehat{\theta}(BS_i), y)$. Note that by the metric version of the law of large numbers we get $\widehat{\theta}_B(S) \approx \mathbb{E}[\widehat{\theta}]$ and we can use in general the same ideas used for numerical variables.

Some examples illustrating the last concepts are outlined below.

Example 4. *The next example is very interesting for a Bussines Register department. Let us consider, for instance, an enterprise group formed by 50 companies. Each company has an economic activity. We use the NACE activity code for determining the economic activity of each company. NACE code is the four digit "statistical classification of economic activities in the European Community" see [11]. The objective is to assign an activity code to the whole group based on the activities of its companies. Furthermore, suppose that we are limited to use just information about the activity code. In order to keep a low administrative burden on the entreprise group a statistics office ask just the activity code of* $n = 5$ *companies from the total group* $N = 50$*. Their NACE codes are, respectively, 4213 ("Construction of bridges and tunnels"), 4211 ("Construction of roads and motorways"), 7111 ("Architectural activities"), 7112 ("Engineering activities and related technical consultancy") and 6910 ("Legal activities"). Similar NACE codes have similar beginning numbers, then we can define the following distance on NACE codes:*

$$d((x_1,x_2,x_3,x_4),(y_1,y_2,y_3,y_4)) = \begin{cases} 6, & \textit{if } x_1 \neq y_1 \\ 5, & \textit{if } x_1 = y_1 \ \wedge \ x_2 \neq y_2 \\ 4, & \textit{if } x_1 = y_1 \ \wedge \ x_2 = y_2 \ \wedge \ |x_3 - y_3| > 1 \\ 3, & \textit{if } x_1 = y_1 \ \wedge \ x_2 = y_2 \ \wedge \ x_3 = y_3 \pm 1 \\ 2, & \textit{if } x_1 = y_1 \ \wedge \ x_2 = y_2 \ \wedge \ x_3 = y_3 \ \wedge \ |x_4 - y_4| > 1 \\ 1, & \textit{if } x_1 = y_1 \ \wedge \ x_2 = y_2 \ \wedge \ x_3 = y_3 \ \wedge \ x_4 = y_4 \pm 1 \\ 0, & \textit{if } x_1 = y_1 \ \wedge \ x_2 = y_2 \ \wedge \ x_3 = y_3 \ \wedge \ x_4 = y_4. \end{cases}$$

Needless to say, there are better and more complex ways of defining distances between NACE codes, however there are out of the scope of this work. With the help of a computer we get that:

$$\widehat{\overline{X}}_{HT} := \arg\min_{y\in \boldsymbol{X}} \sum_{i=1}^{n} d^2(X_i, y) = 7111 \ \textit{NACE}.$$

So we can take the NACE 7111 as the economic activity of this enterprise group. By using bootstrap estimation with $B = 1000$*, we get* $\widehat{B}(\widehat{\overline{X}}_{HT}) = 0$ *and* $\widehat{V}(\widehat{\overline{X}}_{HT}) = 19.41$.

3 .1 Statistical inference on metric spaces

So far, we have achieved an excellent framework to work over metric spaces. We have been able to give the most important definitions appearing in statistical sampling and probability. However, it would be interesting to go further and elaborate a way of using statistical inference for metric spaces. In this subsection we will see how to do this. We need a previous result.

Theorem 2 (Chebyshev's inequality on metric spaces). *Let $(\boldsymbol{X}, d)$ be a finite metric space and $\widehat{\theta} : \mathcal{S} \to 2^{\boldsymbol{X}}$ an estimator. Consider also a positive number $\epsilon > 0$, then*

$$\boldsymbol{P}\left(d(\widehat{\theta}, \mathbb{E}[\widehat{\theta}]) \geq \epsilon\right) \leq \frac{V(\widehat{\theta})}{\epsilon^2}.$$

Proof. Using the definitions given previously, we get

$$V(\widehat{\theta}) = \sum_{S \in \mathcal{S}} \boldsymbol{P}(S) d^2(\widehat{\theta}(S), \mathbb{E}[\widehat{\theta}]) = \sum_{S \in \mathcal{A}} \boldsymbol{P}(S) d^2(\widehat{\theta}(S), \mathbb{E}[\widehat{\theta}]) + \sum_{S \notin \mathcal{A}} \boldsymbol{P}(S) d^2(\widehat{\theta}(S), \mathbb{E}[\widehat{\theta}]) \geq$$

$$\geq \sum_{S \in \mathcal{A}} \boldsymbol{P}(S) d^2(\widehat{\theta}(S), \mathbb{E}[\widehat{\theta}]) + 0 \geq \epsilon^2 \sum_{S \in \mathcal{A}} \boldsymbol{P}(S) = \epsilon^2 \boldsymbol{P}\left(d(\widehat{\theta}, \mathbb{E}[\widehat{\theta}]) \geq \epsilon\right),$$

where $\mathcal{A} = \{S \in \mathcal{S} \mid d^2(\widehat{\theta}(S), \mathbb{E}[\widehat{\theta}]) \geq \epsilon^2\}$.

□

Corollary 1. *Let $(\boldsymbol{X}, d)$ be a finite metric space and $\widehat{\theta} : \mathcal{S} \to 2^{\boldsymbol{X}}$ an estimator. Consider also a positive number $\epsilon > 0$, then*

$$\boldsymbol{P}\left(d(\widehat{\theta}, \mathbb{E}[\widehat{\theta}]) < \epsilon\right) \geq 1 - \frac{V(\widehat{\theta})}{\epsilon^2}.$$

Corollary 2. *Let $(\boldsymbol{X}, d)$ be a finite metric space and $\widehat{\theta} : \mathcal{S} \to 2^{\boldsymbol{X}}$ an estimator. Consider also a positive number $\epsilon > 0$, then*

$$\boldsymbol{P}\left(d(\widehat{\theta}, \overline{X}) < \epsilon\right) \geq 1 - \frac{MSE(\widehat{\theta})}{\epsilon^2}.$$

Proof. The proof is the same as the one for Chebyshev's inequality, but replacing $\mathbb{E}[\widehat{\theta}]$ with $\overline{X}$. □

Now we can give a definition for confidence intervals in metric spaces.

Definition 17. *Let $(\boldsymbol{X}, d)$ be a finite metric space and $\Theta \subset \boldsymbol{X}$. We say that the estimator $\widehat{\theta} : \mathcal{S} \to 2^{\boldsymbol{X}}$ gives a* **symmetric confidence interval** *for Θ of significance level α and radius r if*

$$\boldsymbol{P}\left(d(\widehat{\theta}, \Theta) < r\right) \geq 1 - \alpha.$$

Therefore, we can use the Chebyshev's inequality to build symmetric confidence intervals for the population mean $\overline{X}$ by making $r = \epsilon = \sqrt{\dfrac{MSE(\widehat{\theta})}{\alpha}}$,

$$\boldsymbol{P}\left(d(\widehat{\theta},\overline{X}) < \sqrt{\frac{MSE(\widehat{\theta})}{\alpha}}\right) \geq 1-\alpha.$$

Fixing α, the lower the radius the better the estimator. However, we have not a method to estimate $MSE(\widehat{\theta})$. Fortunately, we can get an upper bound for $MSE(\widehat{\theta})$.

Proposition 8. *Let $(\boldsymbol{X}, d)$ be a finite metric space and $\widehat{\theta} : \mathcal{S} \to 2^{\boldsymbol{X}}$ be an estimator. Then*

$$MSE(\widehat{\theta}) \leq \left(\sqrt{V(\widehat{\theta})} + B(\widehat{\theta})\right)^2.$$

Proof. By the triangle inequality,

$$MSE(\widehat{\theta}) := \sum_{S\in\mathcal{S}} \boldsymbol{P}(S) d^2(\widehat{\theta}(S),\overline{X}) \leq \sum_{S\in\mathcal{S}} \boldsymbol{P}(S)\left(d(\widehat{\theta}(S),\mathbb{E}[\widehat{\theta}]) + d(\mathbb{E}[\widehat{\theta}],\overline{X})\right)^2 =$$

$$= V(\widehat{\theta}) + B^2(\widehat{\theta}) + 2B(\widehat{\theta})\sum_{S\in\mathcal{S}} \boldsymbol{P}(S) d(\widehat{\theta}(S),\mathbb{E}[\widehat{\theta}]).$$

Recall that a function $\varphi : \mathbb{R} \to \mathbb{R}$ is concave if for any $\alpha \in [0,1]$ and $x, y \in \mathbb{R}$, it holds

$$\varphi\left((1-\alpha)x + \alpha y\right) \geq (1-\alpha)\varphi(x) + \alpha\varphi(y).$$

By the Jensen's inequality, if φ is a concave function, x_i real numbers in the domain of φ, and $\alpha_i \geq 0$ for all $i \in \{1, 2, \ldots, L\}$ are positive weights then

$$\varphi\left(\frac{\sum_{i=1}^L \alpha_i x_i}{\sum_{i=1}^L \alpha_i}\right) \geq \frac{\sum_{i=1}^L \alpha_i \varphi(x_i)}{\sum_{i=1}^L \alpha_i}.$$

Now applying the Jensen's inequality for the concave function $\varphi(x) = \sqrt{x}$ we get

$$\sum_{S\in\mathcal{S}} \boldsymbol{P}(S)\sqrt{d^2(\widehat{\theta}(S),\mathbb{E}[\widehat{\theta}])} \leq \sqrt{\sum_{S\in\mathcal{S}} \boldsymbol{P}(S) d^2(\widehat{\theta}(S),\mathbb{E}[\widehat{\theta}])} = \sqrt{V(\widehat{\theta})}.$$

Therefore,

$$MSE(\widehat{\theta}) \leq V(\widehat{\theta}) + B^2(\widehat{\theta}) + 2B(\widehat{\theta})\sqrt{V(\widehat{\theta})} = \left(\sqrt{V(\widehat{\theta})} + B(\widehat{\theta})\right)^2.$$

□

Finally we can use the bootstrap method to get an operative symmetric confidence interval. Let $\widehat{\theta} : \mathcal{S} \to 2^{\boldsymbol{X}}$ be an estimator, then

$$\boldsymbol{P}\left(d(\widehat{\theta},\overline{X}) < \frac{\sqrt{\widehat{V}(\widehat{\theta})} + \widehat{B}(\widehat{\theta})}{\sqrt{\alpha}}\right) \geq 1-\alpha,$$

is a symmetric confidence interval for the population mean $\overline{X}$ of significance level α.

Now let us see some examples.

Example 5. *Imagine that some company is going to launch a new product onto the market. The company wonders which color consumers like best. In this case the target population is the set of all potential consumers. This way the company expects to maximise their sales. With this idea in mind, the company conducts a survey asking about which color will be the best for its new product. They ask 100 people about a RGB color code. The sampling is done without replacement and same selection probabilities. The color is expressed as an RGB triplet* (r, g, b)*, each component of which can vary from zero to a maximum value 255. If all the components are at zero the result is black; if all are at maximum, the result is the brightest representable white. For instance,* $(255, 0, 0)$ *is the red color. Since a lot of possibilities are allowed, it is highly likely that each person will choose a different color. Therefore, we cannot use the statistical mode to get a good representative value of the population. In this case, following the ideas defined above, we need a distance. The euclidean distance is an easy one, that is*

$$d\left((r_1, g_1, b_1), (r_2, g_2, b_2)\right) = \sqrt{(r_2 - r_1)^2 + (g_2 - g_1)^2 + (b_2 - b_1)^2}.$$

Let us suppose that the sample is given by $X_i = (r_i, g_i, b_i) = (i, 2i, i/3)$ *where* $i = 1, 2, \dots, 100$. *The Horvitz-Thompson estimator is*

$$\widehat{\overline{X}}_{HT} := \arg\min_{y \in \boldsymbol{X}} \sum_{i=1}^{n} d^2(X_i, y) = \arg\min_{(r,g,b)} \sum_{i=1}^{n} \left[(r - r_i)^2 + (g - g_i)^2 + (b - b_i)^2\right].$$

Since we have chosen the euclidean distance the last expression reachs a minimum for

$$\left(\frac{\sum_{i=1}^{n} r_i}{n}, \frac{\sum_{i=1}^{n} g_i}{n}, \frac{\sum_{i=1}^{n} b_i}{n}\right) = \left(\frac{n+1}{2}, n+1, \frac{n+1}{6}\right) = (50.5, 101, 16.8).$$

However, the vector $(50.5, 101, 16.8)$ *is not a valid rgb color since the coordinates must be natural numbers. Simply take the nearest integer number of the mean, the selected rule is to round half-integers to the nearest even integer:*

$$\left(\lfloor\frac{\sum_{i=1}^{n} r_i}{n}\rceil, \lfloor\frac{\sum_{i=1}^{n} g_i}{n}\rceil, \lfloor\frac{\sum_{i=1}^{n} b_i}{n}\rceil\right) \in \widehat{\overline{X}}_{HT}$$

In fact, when one mean is a half-integer then the two nearest integer numbers give elements in $\widehat{\overline{X}}_{HT}$ *and these are all the possible solutions. In our case* $\widehat{\overline{X}}_{HT} = \{(50, 101, 17), (51, 101, 17)\}$.

Let us take $(50, 101, 17)$ *as an estimation of the population mean* $\overline{X}$. *Just out of curiosity, this color is shown in Figure 6.*

By using bootstrap estimation with $B = 1000$, *we get* $\widehat{B}(\widehat{\overline{X}}_{HT}) = 0$ *and* $\widehat{V}(\widehat{\overline{X}}_{HT}) = 44.85594$. *By using the last proposition for* $\alpha = 0.05$, *since* $d(x, y) \geq 1$, *we obtain*

$$\boldsymbol{P}\left(d(\widehat{\overline{X}}_{HT}, \overline{X}) < \frac{\sqrt{\widehat{V}(\widehat{\overline{X}}_{HT}) + \widehat{B}(\widehat{\overline{X}}_{HT})}}{\sqrt{\alpha}}\right) \geq 1 - \alpha \Leftrightarrow \boldsymbol{P}\left(d(\widehat{\overline{X}}_{HT}, \overline{X}) < 29.94\right) \geq 0.95.$$

Figure 6: RGB(50,101,17) color.

Then the distance between the estimator $RGB(50, 101, 17)$ *and the population mean* $\overline{X}$ *is lower than 29.94 with a probability of 95%.*

In some cases the mean is very sensitive to outliers. The last example is a good example for that. A good solution is performing the mean without outliers, this is called the **truncated mean or trimmed mean.** In metric spaces, we also can define robust estimators like the trimmed mean. It is enough to define what is an outlier. We say that some element z in a metric space $(\boldsymbol{X}, d)$ is an M-outlier if $\{x \in \boldsymbol{X} \mid d(x, z) \leq M\} = \{z\}$. This way we can define the **Trimmed Horvitz-Thompson estimator** on metric spaces,

$$\widehat{\overline{X}}_{THT} := \arg\min_{y \in \boldsymbol{X}'} \sum_{i=1}^{n} d^2(X_i, y),$$

where $\boldsymbol{X}'$ is the set of non M-outliers of $\boldsymbol{X}$.

Example 6. *In the Economically Active Population Survey one important variable is the attained level of education. This is the typical example of quantitative variable with several categories. The first step is selecting a metric. A possible choice would be the discrete metric* d_D. *With this metric the Horvitz-Thompson estimator is the mode. This is a reasonable option, however we can change the metric to add information about the categories. Usually, the distance between categories is not uniform and using non trivial distances enhance the quality of the model. For defining a non trivial distance we assign a number to each of the categories. These loads have been discussed with experts on the Economically Active Population Survey, see the next table.*

Levels of education	*Loads*	*Sample*
Illiterates	*0*	*4*
Incomplete primary education	*2*	*10*
Primary education	*6*	*11*
First stage of secondary education	*9*	*13*
Second stage of secondary education, with general guidance	*11*	*20*
Second stage of secondary education, with career guidance	*12*	*17*
Higher education	*16*	*25*

The distance between two categories is given by the difference of its loads $d(A, B) = |l_A - l_B|$. *Note that the bigger the distance between two categories, the more complicated it is to move from the first category to the second one. Let us take a sample without replacement of* $n = 100$. *The number* *of individuals on each category is given in the table. Our objective is to give a representative value for the population. Start by computing the Horvitz-Thompson estimator:*

$$\widehat{\overline{X}}_{HT} := \arg\min_{y\in \boldsymbol{X}} \sum_{i=1}^{n} d^2(X_i, y) = \arg\min_{y\in \boldsymbol{X}} \sum_{i=1}^{n} (l_{X_i} - l_y)^2.$$

Since we have chosen the euclidean distance the last expression reachs a minimum for

$$\frac{1}{n}\sum_{i=1}^{n} l_{X_i} = \frac{4\cdot 0 + 10\cdot 2 + 11\cdot 6 + 13\cdot 9 + 20\cdot 11 + 17\cdot 12 + 25\cdot 16}{100} = 10.27.$$

However, the result is not the load of any category. For this reason we should search for the nearest integer being a load, in this case 11. This way,

$$\widehat{\overline{X}}_{HT} = \text{“Second stage of secondary education, with general guidance”}.$$

Observe, that this value is very different from the mode “Higher education”. By using bootstrap estimation with $B = 1000$, we get $\widehat{B}(\widehat{\overline{X}}_{HT}) = 0$ and $\widehat{V}(\widehat{\overline{X}}_{HT}) = 1.261$. By using the last proposition for $\alpha = 0.05$, since $d(x, y) \geq 1$, we obtain

$$\boldsymbol{P}\left(d(\widehat{\overline{X}}_{HT}, \overline{X}) < \frac{\sqrt{\widehat{V}(\widehat{\overline{X}}_{HT}) + \widehat{B}(\widehat{\overline{X}}_{HT})}}{\sqrt{\alpha}}\right) \geq 1 - \alpha \Leftrightarrow \boldsymbol{P}\left(d(\widehat{\overline{X}}_{HT}, \overline{X}) < 5.02\right) \geq 0.95.$$

Then the distance between the estimator $\widehat{\overline{X}}_{HT}$ and the population mean $\overline{X}$ is lower than 5.02 with a probability of 95%. Therefore, we can state that the population mean $\overline{X}$ is not the category “Illiterates” nor “Incomplete primary education” with a probability of 95%.

4 Distances between posets

In the section above, we could see how we can derive a way of doing statistics on a set where the only mathematical structure is given by a metric. In this section, we analyse convenient ways of defining distances between posets. For this purpose, we start working with the adjacency matrix associated to a poset P, M_P. Then we can define the matricial distance between posets as follows.

Definition 18. *Let P, Q be two labeled posets with the same number of elements, $|P| = |Q| = m$. Then we define the* **matricial distance** *between P and Q as:*

$$d_M(P, Q) = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}(M_P(i,j) - M_Q(i,j))^2}.$$

We should check that this function is indeed a metric. The symmetry (property 2.) holds trivially. For 1., note that $\sum_{i=1}^{m}\sum_{j=1}^{m}(M_P(i,j)-M_Q(i,j))^2=0 \;\Leftrightarrow\; M_P(i,j)=M_Q(i,j), \forall i,j \Leftrightarrow P$ is equal to Q. Finally, the triangle inequality is a direct consequence of the classic Minkowski inequality for real numbers x_i, y_i,

$$\sqrt{\sum_{i=1}^{N}(x_i+y_i)^2} \leq \sqrt{\sum_{i=1}^{N}x_i^2}+\sqrt{\sum_{i=1}^{N}y_i^2},$$

adding on each pair (i,j). Therefore, the matricial distance between posets is, in fact, a distance. Under this metric, we can consider the following metric space $(\mathcal{P}_m, d_M)$ where $\mathcal{P}_m$ is the set of labeled posets with m elements and d_M is the matricial distance. Consider here a population of size N, where each individual has an associated poset $P_k \in \mathcal{P}_m$. In a sample of size n, remember that the Horvitz-Thompson estimator for the population mean $\overline{X}$ is

$$\widehat{\overline{X}}_{HT} := \arg\min_{Q\in\mathcal{P}_m}\sum_{k=1}^{n} d_M^2(P_k,Q).$$

By section 3, this estimator is consistent when we consider a sampling scheme without replacement and we have techniques to estimate bias and variance. However, if n is big enough this minimum could be very difficult to compute because the cardinal of $\mathcal{P}_m$ grows amazingly fast. For this reason, we need a practical way of computing $\widehat{\overline{X}}_{HT}$ for this metric.

Lemma 2. *Consider the metric space $(\mathcal{P}_m, d_M)$ and the Horvitz-Thompson estimator $\widehat{\overline{X}}_{HT}$ for a sample of size n, then*

$$\widehat{\overline{X}}_{HT} = \arg\min_{Q\in\mathcal{P}_m} d_E^2(Q,\overline{M}),$$

where $\overline{M}(i,j)=\dfrac{1}{n}\sum_{k=1}^{n} M_{P_k}(i,j)$ and d_E is the euclidean distance between matrices, i.e.

$$d_E^2(Q,\overline{M})=\sum_{i=1}^{m}\sum_{j=1}^{m}\left(M_Q(i,j)-\overline{M}(i,j)\right)^2.$$

Proof. By definition,

$$\widehat{\overline{X}}_{HT} = \arg\min_{Q\in\mathcal{P}_m}\sum_{k=1}^{n} d_M^2(P_k,Q) = \arg\min_{Q\in\mathcal{P}_m}\sum_{k=1}^{n}\sum_{i=1}^{m}\sum_{j=1}^{m}(M_{P_k}(i,j)-M_Q(i,j))^2 =$$

$$= \arg\min_{Q\in\mathcal{P}_m}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{n}(M_{P_k}(i,j)-M_Q(i,j))^2.$$

Now, observe that

$$\sum_{k=1}^{n}(M_{P_k}(i,j)-M_Q(i,j))^2 = \sum_{k=1}^{n}\left(\left(M_{P_k}(i,j)-\overline{M}(i,j)\right)+\left(\overline{M}(i,j)-M_Q(i,j)\right)\right)^2 =$$

$$= \sum_{k=1}^{n} \left(M_{P_k}(i,j) - \overline{M}(i,j)\right)^2 + \sum_{k=1}^{n} \left(\overline{M}(i,j) - M_Q(i,j)\right)^2 .$$

In the last line we have used that $\sum_{k=1}^{n} \left(M_{P_k}(i,j) - \overline{M}(i,j)\right) \left(\overline{M}(i,j) - M_Q(i,j)\right) = 0$. Therefore,

$$\widehat{\overline{X}}_{HT} = \arg \min_{Q \in \mathcal{P}_m} \sum_{i=1}^{m} \sum_{j=1}^{m} \left(\sum_{k=1}^{n} \left(M_{P_k}(i,j) - \overline{M}(i,j)\right)^2 + \sum_{k=1}^{n} \left(\overline{M}(i,j) - M_Q(i,j)\right)^2 \right) =$$

$$= \arg \min_{Q \in \mathcal{P}_m} \left(\sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{n} \left(M_{P_k}(i,j) - \overline{M}(i,j)\right)^2 + \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{n} \left(\overline{M}(i,j) - M_Q(i,j)\right)^2 \right) =$$

$$= \arg \min_{Q \in \mathcal{P}_m} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{n} \left(M_Q(i,j) - \overline{M}(i,j)\right)^2 ,$$

since the first summand remains constant for Q. Finally,

$$\arg \min_{Q \in \mathcal{P}_m} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{n} \left(M_Q(i,j) - \overline{M}(i,j)\right)^2 = \arg \min_{Q \in \mathcal{P}_m} \sum_{i=1}^{m} \sum_{j=1}^{m} \left(M_Q(i,j) - \overline{M}(i,j)\right)^2 .$$

□

Corollary 3. *Consider the metric space $(\mathcal{P}_m, d_M)$ and the Horvitz-Thompson estimator $\widehat{\overline{X}}_{HT}$ for a sample of size n. If the matrix $\lfloor \overline{M} \rceil$, whose coordinates are the coordinates of $\overline{M}$ rounded to the nearest integer, is the adjacency matrix of some poset P, then $\widehat{\overline{X}}_{HT} = P$.*

Observe that the last lemma has a geometric interpretation, see Figure 7. $\widehat{\overline{X}}_{HT}$ is the minimum of a paraboloid in $m \times m$ dimensions. If we worked with real numbers, the minimum would be achieved in the mean value $\overline{M}$. It is just the usual situation for the mean. Typically, $\overline{M}$ is not the adjacency matrix of any poset. Then, we should look for the adjacency matrix which is closest to the general minimum $\overline{M}$. In other words, we are looking for posets in the level curves of a paraboloid, so the closer to the mean the better. This is $\widehat{\overline{X}}_{HT}$ are the posets minimizing its euclidean distance to $\overline{M}$. Before we state the next theorem, we should remember that the adjacency matrix just takes 0 and 1 values.

Theorem 3. *Consider the metric space $(\mathcal{P}_m, d_M)$ and the Horvitz-Thompson estimator $\widehat{\overline{X}}_{HT}$ for a sample of size n, then $\widehat{\overline{X}}_{HT}$ is the solution to the next integer linear programming problem:*

$$\begin{array}{lll} \textit{minimize} & \sum_{i=1}^{m} \sum_{j=1}^{m} \left(1 - 2\overline{M}(i,j)\right) M_Q(i,j) & \\ & & \\ \textit{subject to} & M_Q(i,j) + M_Q(j,i) \leq 1, & i,j = 1,...,m \\ & M_Q(i,j) + M_Q(j,k) \leq 1 + M_Q(i,k), & i,j,k = 1,...,m \\ & M_Q(i,j) \in \{0,1\}, & i,j = 1,...,m \end{array}$$

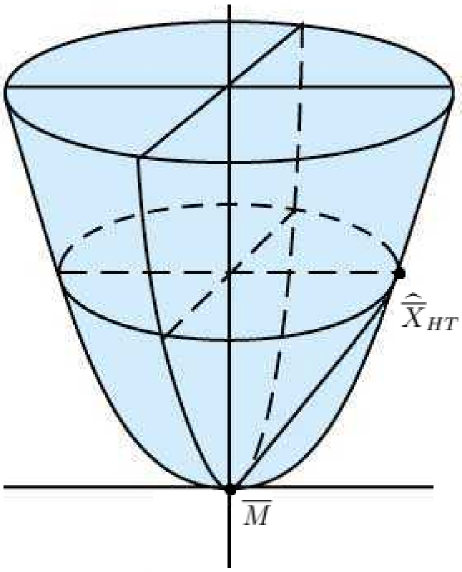

Figure 7: Geometric interpretation, $\widehat{\widehat{X}}_{HT}$ are the posets minimizing its euclidean distance to $\overline{M}$.

Proof. By the last lemma, we have

$$\widehat{\widehat{X}}_{HT} = \arg\min_{Q\in\mathcal{P}_m} \sum_{i=1}^{m}\sum_{j=1}^{m} \left(M_Q(i,j) - \overline{M}(i,j)\right)^2 =$$

$$= \arg\min_{Q\in\mathcal{P}_m} \sum_{i=1}^{m}\sum_{j=1}^{m} \left[M_Q^2(i,j) + \overline{M}^2(i,j) - 2M_Q(i,j)\overline{M}(i,j)\right].$$

Since the adjacency matrix is 0/1-valued we get $M_Q^2(i,j)+\overline{M}^2(i,j)-2M_Q(i,j)\overline{M}(i,j) = M_Q(i,j)+\overline{M}^2(i,j) - 2M_Q(i,j)\overline{M}(i,j)$. Moreover, the second summand is constant for Q then

$$\widehat{\widehat{X}}_{HT} = \arg\min_{Q\in\mathcal{P}_m} \sum_{i=1}^{m}\sum_{j=1}^{m} \left[M_Q(i,j) - 2M_Q(i,j)\overline{M}(i,j)\right] =$$

$$= \arg\min_{Q\in\mathcal{P}_m} \sum_{i=1}^{m}\sum_{j=1}^{m} \left(1 - 2\overline{M}(i,j)\right) M_Q(i,j).$$

Now, we are going to state this problem in a mathematical optimization setting. Before that, we should realize two things. For a matrix M_Q to be the adjacency matrix of some poset Q, we need that M_Q verifies antisymmetry and transitivity conditions. Symmetry can be translated as

$M_Q(i,j) + M_Q(j,i) \leq 1$, $\forall i,j$. Transitivity translates as $M_Q(i,j) + M_Q(j,k) \leq 1 + M_Q(i,k)$, $\forall i,j,k$. Then $\widehat{\overline{X}}_{HT}$ is the solution to the next integer linear programming problem.

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{m} \sum_{j=1}^{m} \left(1 - 2\overline{M}(i,j)\right) M_Q(i,j) \\ & \\ \text{subject to} & \begin{array}{ll} M_Q(i,j) + M_Q(j,i) \leq 1, & i,j = 1,...,m \\ M_Q(i,j) + M_Q(j,k) \leq 1 + M_Q(i,k), & i,j,k = 1,...,m \\ M_Q(i,j) \in \{0,1\}, & i,j = 1,...,m \end{array} \end{array}$$

□

There are many algorithms to solve integer linear programming (ILP) problems, see appendix B. Most of them are implemented in common programming languages as R, SAS or Matlab. However, ILP problems are NP-complete, so they are in general difficult to solve. Nevertheless, we always will be able to solve this problem for a fairly small m. The polytope determined by the inequalities,

$$\begin{array}{ll} M_Q(i,j) + M_Q(j,i) \leq 1, & i,j = 1,...,m \\ M_Q(i,j) + M_Q(j,k) \leq 1 + M_Q(i,k), & i,j,k = 1,...,m \\ M_Q(i,j) \geq 0, & i,j = 1,...,m \end{array}$$

is the polytope associated to this linear programming problem. Let us denote this polytope by $\mathcal{PL}_m$. Since the objective function is linear the optimum is achieved in a vertex of this polytope. It is not very surprising that every labeled poset is a vertex of this polytope. Although there are other vertices apart from the ones defining labeled posets. These other vertices are not 0/1 valued, then $\mathcal{PL}_m$ is not an integer polytope.

Theorem 4. *Let $P \in \mathcal{P}_m$ be a labeled poset with m elements. Then the adjacent matrix of P, M_P is a vertex of $\mathcal{PL}_m$.*

Proof. We are going to prove that every labeled poset is an extreme point of $\mathcal{PL}_m$. Observe that any labeled poset is represented as a 0/1-valued adjacency matrix M_P. Suppose that M_P is not a vertex of $\mathcal{PL}_m$, then exist two different points $X, Y \in \mathcal{PL}_m$ such that $M_P = \alpha X + (1-\alpha)Y$ for some $\alpha \in (0,1)$. If $M_P(i,j) = 0 = \alpha X(i,j) + (1-\alpha)Y(i,j)$ then $X(i,j) = Y(i,j) = 0$ because $0 \leq X, Y \leq 1$. Symmetrically, if $M_P(i,j) = 1 = \alpha X(i,j) + (1-\alpha)Y(i,j)$ then $X(i,j) = Y(i,j) = 1$. Therefore, $X = Y = M_P$ and we attain a contradiction. We conclude that M_P is a vertex of $\mathcal{PL}_m$. □

In practice, we can avoid dealing with integer linear programming since all the labeled posets are vertices of $\mathcal{PL}_m$. Instead we should work with the next linear programming problem:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{m} \sum_{j=1}^{m} \left(1 - 2\overline{M}(i,j)\right) M_Q(i,j) \\ & \\ \text{subject to} & \begin{array}{ll} M_Q(i,j) + M_Q(j,i) \leq 1, & i,j = 1,...,m \\ M_Q(i,j) + M_Q(j,k) \leq 1 + M_Q(i,k), & i,j,k = 1,...,m \\ M_Q(i,j) \geq 0. & i,j = 1,...,m \end{array} \end{array}$$

Of course, this problem has a much lower complexity than the last integer linear problem. Indeed, this problem has polynomial complexity and it could be solved with the famous simplex method. If the optimum is achieved in a 0/1 vertex then we have already concluded. If that is not the case, we should search for the second best vertex using simplex algorithm, and so on. This way, we would look for the optimum in the vertices which are adjacent to the linear programming solution. Usually, simplex algorithm does that for us. The simplex method is an algorithm easy to find in most mathematical or statistical software.

Let us see now an example illustrating the last ideas.

Example 7. *Imagine the next situation. The goverment wants to conduct a survey asking about the major concerns for citizens. The target population are all the people of legal age of that country. They can choose among five different issues of concern: "Unemployment" (U), "Education system" (E), "Environment" (N), "Public health" (H) and "Pensions" (P). The sample is designed without replacement and having size* $n = 100$. *Note that each person reports about their preference poset, for instance see Figure 8 showing the preference posets of two persons. Note that the first person is more concerned about pensions than about education while the second one is indifferent between these two issues. The rest of the sample is simulated by computer.*

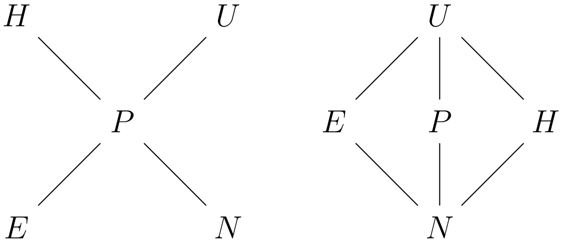

Figure 8: Two preference posets.

It is a good and easy exercise to compute the matricial distance between the last two posets, $d_M(P, Q) = 5$. *Now, we are going to compute the Horvitz-Thompson estimator:*

$$\widehat{\overline{X}}_{HT} := \arg\min_{Q \in \mathcal{P}_m} \sum_{k=1}^{n} d_M^2(P_k, Q).$$

In order to get the value of $\widehat{\overline{X}}_{HT}$, *we should solve the next ILP problem.*

$$\begin{array}{ll} \textit{minimize} & \sum_{i=1}^{5} \sum_{j=1}^{5} \left(1 - 2\overline{M}(i,j)\right) M_Q(i,j) \\ & \\ \textit{subject to} & M_Q(i,j) + M_Q(j,i) \leq 1, \qquad\qquad\quad i,j = 1,...,5 \\ & M_Q(i,j) + M_Q(j,k) \leq 1 + M_Q(i,k), \;\; i,j,k = 1,...,5 \\ & M_Q(i,j) \in \{0,1\}, \qquad\qquad\qquad\quad\, i,j = 1,...,5 \end{array}$$

Using a mathematical software we get that $\widehat{\overline{X}}_{HT}$ *is just the poset displayed in the next Figure 9.*

Therefore, in this example, the unemployment is the major concern for citizens, followed by public health and pensions. The enviroment and the education are of much lower concern. With this

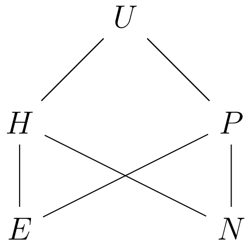

Figure 9: $\widehat{\widehat{X}}_{HT}$ obtained by solving the associated ILP.

estimation, the goverment wants to know how to spend the money according to the preferences of people. As it is explained in appendix A, we can give a solution to this by using the center of gravity of the order polytope associated to $\widehat{\widehat{X}}_{HT}$. Let us call P to the poset $\widehat{\widehat{X}}_{HT}$. The center of gravity c_P can be computed as follows:

$$c_P = \frac{1}{m+1}H(P),$$

where $H(P)(a) := \frac{1}{e(P)}\sum_{\epsilon\in\mathcal{L}(P)} h_\epsilon(a)$ being $h_\epsilon(a) = |\{x \in P : x \preceq_\epsilon a\}|$. Poset P has 4 linear extensions: (E, N, H, P, U), (N, E, H, P, U), (E, N, P, H, U), (N, E, P, H, U). Now, using the order (E, N, H, P, U), we sum the vectors having the positions of each element on each linear extension:

$$H(P) = \frac{1}{4}\left[(1, 2, 3, 4, 5) + (2, 1, 3, 4, 5) + (1, 2, 4, 3, 5) + (2, 1, 4, 3, 5)\right] = \frac{1}{4}(6, 6, 14, 14, 20).$$

Finally, $c_P = \frac{1}{6}H(P) = (3/12, 3/12, 7/12, 7/12, 10/12)$.

We just divide c_P by its sum to get a vector which add up to 1, i.e. $\overline{c_P} = (3/30, 3/30, 7/30, 7/30, 10/30)$. This vector gives us the percentage that must be invested by the goverment on each issue.

Finally we are going to introduce other interesting distance between posets, see [1].

Definition 19. *Let P, Q be two labeled posets with the same number of elements, $|P| = |Q| = m$. Then we define the internal distance between P and Q as:*

$$d_I(P, Q) = m - \eta(P, Q),$$

where $\eta(P, Q)$ is the cardinal of biggest poset which is isomorphic simultaneously to P and Q as labeled posets.

Proposition 9. *$(\mathcal{P}_m, d_I)$ is a metric space.*

Proof. We have to show that d_I is a distance. Properties 1. and 2. hold trivially. Let us see the triangle inequality. Let P_1, P_2 and P_3 be three labeled posets. Let us call Q_{ij} to the biggest poset which is isomorphic simultaneously to P_i and P_j. We want to prove that $d_I(P_1, P_2) \leq d_I(P_1, P_3) + d_I(P_3, P_2)$. We consider three different cases:

- If $|Q_{13} \cap Q_{32}| = 0 \Rightarrow \eta(P_1, P_3) + \eta(P_3, P_2) \leq m \leq m + \eta(P_1, P_2) \Rightarrow d_I(P_1, P_2) \leq d_I(P_1, P_3) + d_I(P_3, P_2)$.
- If $|Q_{13} \cap Q_{32}| = 1 \Rightarrow \eta(P_1, P_3) + \eta(P_3, P_2) \leq m + 1 \leq m + \eta(P_1, P_2) \Rightarrow d_I(P_1, P_2) \leq d_I(P_1, P_3) + d_I(P_3, P_2)$.
- If $|Q_{13} \cap Q_{32}| \geq 2$. Take a pair of elements $(x, y) \in Q_{13} \cap Q_{32}$, then $x \leq_{P_1} y \Leftrightarrow x \leq_{P_3} y \Leftrightarrow x \leq_{P_2} y$ then $(x, y) \in Q_{12} \Rightarrow Q_{13} \cap Q_{32} \subseteq Q_{12} \Rightarrow Q_{12}^c \subseteq Q_{13}^c \cup Q_{32}^c \Rightarrow d_I(P_1, P_2) \leq d_I(P_1, P_3) + d_I(P_3, P_2)$.

□

This distance perfectly captures the idea of poset distance at the expense of a higher complexity. Finding a way of making operative this internal distance between posets remains as an open problem.

4 .1 Distances between chains

Sometimes, it is difficult to collect data with a partially ordered structure. Particularly, when m is big enough, asking people to choose between so many alternatives could not be the best option. In this case, using chains is an easier alternative. Each individual reports about a total order being their preferences on some topic. In a total order every two elements are related to each other. We are going to work with the following metric space $(\mathcal{C}_m, d_M)$ where $\mathcal{C}_m$ is the set of labeled chains with m elements and d_M is the matricial distance. Let $C \in \mathcal{C}_m$ be a chain and M_C its associated adjacency matrix, note that $M_C(i, j) + M_C(j, i) = 1 \ \forall i \neq j$. Then the ILP problem associated to the Horvitz-Thompson estimator changes.

Theorem 5. *Consider the metric space $(\mathcal{C}_m, d_M)$ and the Horvitz-Thompson estimator $\widehat{\overline{X}}_{HT}$ for a sample of size n, then $\widehat{\overline{X}}_{HT}$ is the solution to the next integer linear programming problem:*

$$\begin{array}{ll} \textit{minimize} & \sum_{i=1}^m \sum_{j=1}^m \left(1 - 2\overline{M}(i,j)\right) M_Q(i,j) \\ & \\ \textit{subject to} & \begin{array}{ll} M_Q(i,j) + M_Q(j,i) = 1, & i,j = 1, ..., m \\ M_Q(i,j) + M_Q(j,k) \leq 1 + M_Q(i,k), & i,j,k = 1, ..., m \\ M_Q(i,j) \in \{0,1\}, & i,j = 1, ..., m \end{array} \end{array}$$

Proof. The proof is the same as the one above done for $(\mathcal{P}_m, d_M)$. This is because a chain is a specific type of poset. □

However, we can get another interesting metric space for chains by changing a little bit the distance. To determine a chain we just need to know the position of each element in the chain. Let us define another matrix associated to a chain, called the position matrix and denoted by T_C. It is defined as the square $m \times m$ matrix such that

$$T_C(i,j) = \begin{cases} 1, & \text{if } i \text{ is in position } j \text{ at } C \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, we can define a new distance between chains by

$$d_T(C,D) = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}(T_C(i,j) - T_D(i,j))^2}.$$

The following result is proved in the same way as in the last subsection.

Lemma 3. *Consider the metric space $(\mathcal{C}_m, d_T)$ and the Horvitz-Thompson estimator $\widehat{\overline{X}}_{HT}$ for a sample of size n, namely $\{C_1, \ldots, C_n\}$, then*

$$\widehat{\overline{X}}_{HT} = \arg\min_{D\in\mathcal{C}_m} d_E^2(D, \overline{T}),$$

where $\overline{T}(i,j) = \dfrac{1}{n}\sum_{k=1}^{n} T_{C_k}(i,j)$ and d_E is the euclidean distance between the matrices, i.e.

$$d_E^2(D,\overline{T}) = \sum_{i=1}^{m}\sum_{j=1}^{m}\left(T_D(i,j) - \overline{T}(i,j)\right)^2.$$

Now we can get a linear programming problem for this estimator.

Theorem 6. *Consider the metric space $(\mathcal{C}_n, d_T)$ and the Horvitz-Thompson estimator $\widehat{\overline{X}}_{HT}$ for a sample of size n, then $\widehat{\overline{X}}_{HT}$ is the solution to the next linear programming problem:*

$$\begin{array}{lll}
\textit{minimize} & \sum_{i=1}^{m}\sum_{j=1}^{m}\left(1 - 2\overline{T}(i,j)\right)T_D(i,j) & \\
 & & \\
\textit{subject to} & \displaystyle\sum_{i=1}^{m} T_D(i,j) = 1, & j = 1,...,m \\
 & \displaystyle\sum_{j=1}^{m} T_D(i,j) = 1, & i = 1,...,m \\
 & T_D(i,j) \geq 0, & i,j = 1,...,m
\end{array}$$

Proof. By the last lemma, we have

$$\widehat{\overline{X}}_{HT} = \arg\min_{D\in\mathcal{C}_m}\sum_{i=1}^{m}\sum_{j=1}^{m}\left(T_D(i,j) - \overline{T}(i,j)\right)^2 =$$

$$= \arg\min_{D\in\mathcal{C}_m}\sum_{i=1}^{m}\sum_{j=1}^{m}\left[T_D^2(i,j) + \overline{T}^2(i,j) - 2T_D(i,j)\overline{T}(i,j)\right].$$

Since the position matrix is 0/1-valued we get $T_D^2(i,j) + \overline{T}^2(i,j) - 2T_D(i,j)\overline{T}(i,j) = T_D(i,j) + \overline{T}^2(i,j) - 2T_D(i,j)\overline{T}(i,j)$. Moreover, the second summand is constant for D then

$$\widehat{\overline{X}}_{HT} = \arg\min_{D\in\mathcal{C}_m}\sum_{i=1}^{m}\sum_{j=1}^{m}\left[T_D(i,j) - 2T_D(i,j)\overline{T}(i,j)\right] =$$

$$= \arg\min_{D\in\mathcal{C}_m} \sum_{i=1}^{m}\sum_{j=1}^{m} \left(1-2\overline{T}(i,j)\right) T_D(i,j).$$

Now, we are going to state this problem in a mathematical optimization setting. Before that, we should realize two things. For a matrix T_D to be the adjacency matrix of some chain D, we need that each element is placed at just one position, $\sum_{i=1}^{m} T_D(i,j) = 1 \ \forall j$. Also, we can only have one element on each position, $\sum_{i=j}^{m} T_D(i,j) = 1 \ \forall i$. Then the linear problem associated looks like

$$\begin{array}{lll} \text{minimize} & \sum_{i=1}^{m}\sum_{j=1}^{m} \left(1-2\overline{T}(i,j)\right) T_D(i,j) & \\ \text{subject to} & \displaystyle\sum_{i=1}^{m} T_D(i,j) = 1, & j=1,...,m \\ & \displaystyle\sum_{j=1}^{m} T_D(i,j) = 1, & i=1,...,m \\ & T_D(i,j) \in \{0,1\}, & i,j=1,...,m \end{array}$$

Now, we are going to prove that the polytope defined by the constraints

$$\begin{array}{lll} \text{subject to} & \displaystyle\sum_{i=1}^{m} T_D(i,j) = 1, & j=1,...,m \\ & \displaystyle\sum_{j=1}^{m} T_D(i,j) = 1, & i=1,...,m \\ & T_D(i,j) \geq 0, & i,j=1,...,m \end{array}$$

is an integer polytope. It is enough to show that the matrix associated to the equalities is totally unimodular. By using the Hoffman-Gale theorem (see appendix A) with sets of rows R_1 for the first set of equalities and R_2 for the second set of equalities the result follows. Since the polytope is integer, the integer problem and the linear problem are equivalent to each other and the theorem holds.

□

Note that the last problem is a classic assignement problem. This way working with chains implies a lower complexity than working with general posets.

Example 8. *We continue with the last example. However, in this case we work in the restricted case of chains. The sample is designed without replacement and having size $n = 100$. In this case each person reports about their preference chain, for instance see Figure 10 showing the preference chains of two persons. Fixing the order (E, N, H, P, U), their position matrices are*

$$T_C = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad T_D = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The rest of the sample is simulated by computer.

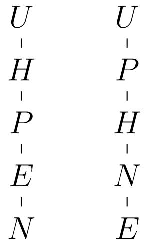

Figure 10: Two preference chains C and D.

It is a good and easy exercise to compute the matricial distance between the last two posets, $d_T(C, D) = 8$. *Now, we are going to compute the Horvitz-Thompson estimator:*

$$\widehat{\widehat{X}}_{HT} := \arg \min_{D \in \mathcal{C}_m} \sum_{k=1}^{n} d_T^2(C_k, D).$$

In order to get the value of $\widehat{\widehat{X}}_{HT}$, *we should solve the next linear programming problem.*

$$\begin{array}{lll} \textit{minimize} & \sum_{i=1}^m \sum_{j=1}^m \left(1 - 2\overline{T}(i,j)\right) T_D(i,j) & \\ & & \\ \textit{subject to} & \displaystyle\sum_{i=1}^{m} T_D(i,j) = 1, & j = 1, ..., m \\ & \displaystyle\sum_{j=1}^{m} T_D(i,j) = 1, & i = 1, ..., m \\ & T_D(i,j) \geq 0, & i, j = 1, ..., m \end{array}$$

Using a mathematical software we get that $\widehat{\widehat{X}}_{HT}$ *is just the poset displayed in the next Figure 11. Therefore, in this example, the unemployment is also the major concern for citizens.*

As in the last example, now the goverment wants to know how to spend the money according to the preferences of people. We can also give a solution to this by using the center of gravity of the order polytope associated to $\widehat{\widehat{X}}_{HT}$. *Let us call* C *to the poset* $\widehat{\widehat{X}}_{HT}$. *The center of gravity* c_C *can be computed as follows:*

$$c_C = \frac{1}{m+1} H(C),$$

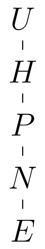

Figure 11: $\widehat{\widehat{X}}_{HT}$ obtained by solving the associated LP problem.

Now, using the order (E, N, H, P, U), *since there is just one linear extension:*

$$H(P) = (1, 2, 4, 3, 5).$$

Finally, $c_C = \frac{1}{6}H(P) = (1/6, 2/6, 4/6, 3/6, 5/6)$.

We just divide c_C *by its sum to get a vector which add up to 1, i.e.* $\overline{c_C} = (1/15, 2/15, 4/15, 3/15, 5/15)$. *This vector gives us the percentage that must be invested by the goverment on each issue.*

5 Applications to official statistics

In this section we present several possible uses of the last theoretical framework. We explain how to apply sampling theory on ordered structures to problems in official statistics. The goal of this section is not to use real data for giving a completely finished methodology, but explaing the main ideas, while leaving details for future work. Considering issues of confidentiality, we will not use real data. Instead, we will use computer simulations inspired on the publicly available data. In order to obtain these data we would need to include a simple and concrete question in the associated questionnaire asking about the poset of preferences. The examples elaborated in this setion illustrate several possible applications of order theory to official statistics, however the presented theory can be applied in many more cases.

5 .1 Election surveys and voting intention

One of the most common uses of statistics is estimating the voting intention by asking people about their political preferences. The spanish Sociological Research Center (CIS) carries out regular surveys covering these topics, see [3]. The objective is always to determine the general public opinion about the different political parties. In many cases, this kind of surveys include questions about the order of preference among the different political parties. This order is always supposed to be a chain. This information is very interesting but it can be improved by allowing people to be indifferent between several political groups. That means allowing people to inform about its political preferences poset. In this case, the target population could be the population of voting age of a whole country or some specific region. Let N be the total number of individuals in the target population and n the sample size. For this example, the sample size will be $n = 100$. The elements of this sample are denoted by

$P_1, P_2, \ldots, P_n$. Once again, since we do not have data with this information we simulate by computer the preference of each individual. Let us suppose that there are five political parties A, B, C, D and E. Obviously we will use the distance d_M and our goal is to estimate $\overline{X}$. As explained in last sections, the first step is computing the Horvitz-Thompson estimator:

$$\widehat{\overline{X}}_{HT} := \arg \min_{Q \in \mathcal{P}_m} \sum_{k=1}^{n} d_M^2(P_k, Q).$$

In order to get the value of $\widehat{\overline{X}}_{HT}$, we can solve the next ILP problem:

$$\begin{array}{lll}
\text{minimize} & \sum_{i=1}^{5} \sum_{j=1}^{5} \left(1 - 2\overline{M}(i,j)\right) M_Q(i,j) & \\
 & & \\
\text{subject to} & M_Q(i,j) + M_Q(j,i) \leq 1, & i,j = 1, ..., 5 \\
 & M_Q(i,j) + M_Q(j,k) \leq 1 + M_Q(i,k), & i,j,k = 1, ..., 5 \\
 & M_Q(i,j) \in \{0,1\}, & i,j = 1, ..., 5
\end{array}$$

Using a mathematical software we get that $\widehat{\overline{X}}_{HT}$ is just the poset displayed in the next Figure 12

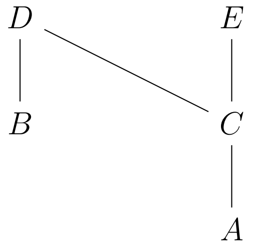

Figure 12: $\widehat{\overline{X}}_{HT}$ obtained by solving the associated ILP.

Therefore, in this example, the political parties D and E are the ones preferred by most citizens. Voting intention for A and B is much lower.

Now, we give a more precise estimation of the voting intention. As it is explained in appendix A, we can give a solution to this by using the center of gravity of the order polytope associated to $\widehat{\overline{X}}_{HT}$. Let P be the poset $\widehat{\overline{X}}_{HT}$. The center of gravity c_P can be computed as follows:

$$c_P = \frac{1}{m+1} H(P),$$

where $H(P)(a) := \frac{1}{e(P)} \sum_{\epsilon \in \mathcal{L}(P)} h_\epsilon(a)$ being $h_\epsilon(a) = |\{x \in P : x \preceq_\epsilon a\}|$. Poset P has 7 linear extensions: (B, A, C, D, E), (B, A, C, E, D), (A, B, C, D, E), (A, B, C, E, D), (A, C, B, D, E), (A, C, B, E, D), (A, C, E, B, D). Now, using the order (A, B, C, D, E), we sum the vectors having the positions of each element on each linear extension:

$$H(P) = \frac{1}{7}[(2,1,3,4,5) + (2,1,3,5,4) + (1,2,3,4,5) + (1,2,3,5,4) +$$

$$+(1,3,2,4,5)+(1,3,2,5,4)+(1,4,2,5,3)]=\frac{1}{7}(9,16,18,32,30).$$

Finally, $c_P=\frac{1}{6}H(P)=(9/42,16/42,18/42,32/42,30/42)$.

We just divide c_P by its sum to get a vector which add up to 1, i.e.

$$\overline{c_P}=(9/105,16/105,18/105,32/105,30/105).$$

This vector gives us an estimation of the voting intention in respect of each political party.

5 .2 Social Surveys: fertility, habits and opinion

Conducting surveys about habits and opinion is the best way of kwnowing the way of thinking and doing of a society. There are many public institutions dedicated to design and conduct this kind of surveys, among which we can find the National Statistics Institute (INE) or the Sociological Research Center (CIS). We can regularly find questions in these questionnaires asking about ordering by importance several choices. In these cases, we could improve the accuracy of the answer by letting people be indifferent to different choices. Once again, posets improve the model. The fertility survey published by the National Statistics Institute (INE), see [10], is perfect to implement our new techniques. This survey has a lot of multiple choice questions where a ranking order is needed. Some examples of these variables are: “Most valued aspects of a job for women”, “Main reasons why women do not intend to have children” or “Main reasons for the delay in motherhood with respect to the ideal time”. In this example, we are going to work with the variable “Most valued aspects of a job for women”. This variable can take 6 possible values: “Good economic conditions” (E), “Long-term labor stability” (S), “Family reconciliation measures” (F) (teleworking, flexibility in the work schedule, in the holidays ...), “An interesting job that satisfies me personally and professionally” (I), “Good work schedule” (SC) and “Good Location” (L). In this case, women are asked about the three most valued aspects of a job. However, it would be possible to improve the accuracy of the estimation by asking about a complete ranking order of these six categories. Obtaining this information is as easy as incorporaring this question in the questionnaire. Let us suppose that we already have these data. Then, for this example we have allowed people to report about their chain of preferences. This way the answers are chains in $\mathcal{C}_6$. We will use the distance d_T. The target population in this case are the women aged 18 to 55, both ages included, who reside in main family dwellings throughout the national territory, see [10]. Let N be the total number of individuals in the target population and n the sample size. In this case, the sample size will be $n=100$. The elements of this sample are denoted by $C_1,C_2,\ldots,C_n$. In this illustrating example, we simulate by computer the chain of each individual based on the publicly available information. The objective is to estimate the mean population chain $\overline{X}$. The first step is, as always, computing the Horvitz-Thompson estimator:

$$\widehat{\overline{X}}_{HT}:=\arg\min_{D\in\mathcal{C}_6}\sum_{k=1}^{n}d_T^2(C_k,D).$$

In order to get the value of $\widehat{\overline{X}}_{HT}$, we should solve the next linear programming problem. Note that the problem does not involve integer variables and is very easy to solve with any mathematical software.

$$
\begin{array}{lll}
\text{minimize} & \sum_{i=1}^{m}\sum_{j=1}^{m}\left(1-2\overline{T}(i,j)\right)T_D(i,j) & \\
 & & \\
\text{subject to} & \displaystyle\sum_{i=1}^{m} T_D(i,j)=1, & j=1,...,m \\
 & \displaystyle\sum_{j=1}^{m} T_D(i,j)=1, & i=1,...,m \\
 & T_D(i,j)\geq 0, & i,j=1,...,m
\end{array}
$$

Using a mathematical software we get that $\widehat{\overline{X}}_{HT}$ is just the poset displayed in the next Figure 13.

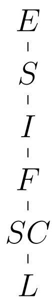

Figure 13: $\widehat{\overline{X}}_{HT}$ obtained by solving the associated LP.

Therefore, in this case, the "Good economic conditions" are the most valued aspect of a job in this chain and the "Good location" would be the least valued.

It is also interesting to obtain a symmetric confidence interval for the population mean chain. Now by using bootstrap estimation with $B = 200$, we get $\widehat{B}(\widehat{\overline{X}}_{HT}) = 1$ and $\widehat{V}(\widehat{\overline{X}}_{HT}) = 1.71$. For $\alpha = 0.05$, since $d_T(x, y) \geq 1$, we obtain

$$
\boldsymbol{P}\left(d_T(\widehat{\overline{X}}_{HT},\overline{X}) < \frac{\sqrt{\widehat{V}(\widehat{\overline{X}}_{HT})+\widehat{B}(\widehat{\overline{X}}_{HT})}}{\sqrt{\alpha}}\right) \geq 1-\alpha \Leftrightarrow \boldsymbol{P}\left(d(\widehat{\overline{X}}_{HT},\overline{X}) < 10.32\right) \geq 0.95.
$$

Then the distance between the estimator $\widehat{\overline{X}}_{HT}$ and the population mean $\overline{X}$ is lower than 10.32 with a probability of 95%. For instance, we can say that the population mean $\overline{X}$ is different from the chain $C' = (SC, L, I, F, E, S)$ with a probability of 95%, since $d_T(\widehat{\overline{X}}_{HT}, C') = 12$.

Other important surveys about habits and opinion can be studied with this technique. For instance, analysing the major concerns of citizens are one of the most common opinion surveys, see [3]. In Examples 7 and 8 we have explained how to use posets to deal with these problems. The main advantage here is that we benefit from the information of two choices to be equally important.

Moreover, the resulting posets are an excellent measure of how the society places importance on different social, ethical and legal issues.

5 .3 User Satisfaction Surveys

Over the last few decades, the National Statistic Institute (INE) has made many efforts both in terms of management and quality control of its products, with the aim of maintaining the high degree of confidence from which the official statistical information is currently profiting see [19].

The INE conducts user surveys in order to know their opinion and degree of satisfaction with statistics, as well as to detect new information needs. For instance, the 2016 User Satisfaction Survey (USS2016) collected the opinions of 272 qualified users who, on a voluntary basis, had collaborated in the research by expressing their opinions on quality and confidence in statistics, and by making suggestions for improving the statistical system, see [20]. Selected users represented the main institutions of experienced and qualified users in the statistical area: researchers and professors from university centres, public administration bodies, international experts, the media and other institutions. One of the questions appearing in this questionnaire is about ordering by importance the six different dimensions of quality. These dimensions are explained below. The six dimensions of quality are: "Relevance" (R),"Accuracy" (A), "Timeliness" (T), "Coherence" (C), "Geographical comparability" (GC) and "Temporal comparability" (TC). Relevance is understood as the degree to which INE statistics satisfy the needs of users. The accuracy is defined as how well INE statistics reflect reality. The development of statistical and mathematical methods as well as the progress of technology improve significantly the statistics accuracy. Timeliness is the lapse of time between the publication of the information and the period to which it refers. Geographical and temporal comparability attempt to measure to what extent the different sources are compatible with each other [20]. Asking users about ordering the diffent dimensions of quality is a core question which allows to compute an overall index of user satisfaction with the statistics. Currently, only chains are included as possible answer to this questions. A way of improving accuracy is to allow users to be indifferent between some dimensions of quality. This way the answer are posets in $\mathcal{P}_6$. Here we make an example of how one can use posets for this purpose. The target population in this case are the users of a specific statistical product. Let N be the total number of individuals in the target population and n the sample size. In this case, the sample size will be $n = 100$. The elements of this sample are denoted by $P_1, P_2, \ldots, P_n$. So far, this kind of data are not collected, but it would be as simple as including a new question in the questionnaire asking about the preference poset of each person. In this illustrating example, we simulate by computer the preference of each individual. We will use the distance d_M and our goal is to estimate $\overline{X}$. The first step is, as explained previously, computing the Horvitz-Thompson estimator:

$$\widehat{\overline{X}}_{HT} := \arg \min_{Q \in \mathcal{P}_m} \sum_{k=1}^{n} d_M^2(P_k, Q).$$

In order to get the value of $\widehat{\overline{X}}_{HT}$, we can solve the next ILP problem:

$$\begin{array}{lll} \text{minimize} & \sum_{i=1}^{6}\sum_{j=1}^{6}\left(1-2\overline{M}(i,j)\right)M_Q(i,j) & \\ & & \\ \text{subject to} & M_Q(i,j)+M_Q(j,i)\leq 1, & i,j=1,...,6 \\ & M_Q(i,j)+M_Q(j,k)\leq 1+M_Q(i,k), & i,j,k=1,...,6 \\ & M_Q(i,j)\in\{0,1\}, & i,j=1,...,6 \end{array}$$

Using a mathematical software we get that $\widehat{\overline{X}}_{HT}$ is just the poset displayed in the next Figure 14

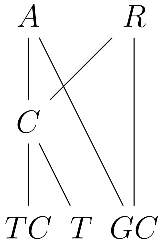

Figure 14: $\widehat{\overline{X}}_{HT}$ obtained by solving the associated ILP.

Therefore, in this example, both "Relevance" and "Accuracy" are the most important dimensions for most people. "Coherence" is the next most important dimension. Finally, "Timeliness", "Geographical comparability" and "Temporal comparability" are less important to users in accordance with the simulated data. From the last result, we can give an interesting approximation to a global satisfaction index. There are different methodological approaches for doing this, see [7]. This index is just a way of weighing the different dimensions of quality. As it is explained in appendix A, we can construct this index by using the center of gravity of the order polytope associated to $\widehat{\overline{X}}_{HT}$. Let us call P to the poset $\widehat{\overline{X}}_{HT}$. The center of gravity c_P can be computed as follows:

$$c_P=\frac{1}{m+1}H(P),$$

where $H(P)(a):=\dfrac{1}{e(P)}\sum_{\epsilon\in\mathcal{L}(P)}h_\epsilon(a)$ being $h_\epsilon(a)=|\{x\in P: x\preceq_\epsilon a\}|$. Poset P has 16 linear extensions. Now, using the order (TC,GC,C,T,A,R), we sum the vectors having the positions of each element on each linear extension and we get:

$$H(P)=\frac{1}{16}[(30,40,52,30,88,88)].$$

Finally, $c_P=\dfrac{1}{7}H(P)=(30/112,40/112,52/112,30/112,88/112,88/112)$.

We just divide c_P by its sum to get a vector which add up to 1, i.e.

$$\overline{c_P}=(30/328,40/328,52/328,30/328,88/328,88/328).$$

This vector gives us coefficients of each dimension in the satisfaction index. That is,

$$I=\frac{30}{328}TC+\frac{40}{328}GC+\frac{52}{328}C+\frac{30}{328}T+\frac{88}{328}A+\frac{88}{328}R.$$

5 .4 Quality of Life Indicators

Quality of life indicators measure the progress and well-being of individuals. This involves broadening the framework of economic development indicators traditionally used as measures of growth and well-being (GDP, other aggregated National Accounts indicators). The development of these indicators has been promoted by a large number of initiatives from both the scientific and the academic world, as well as from European and international organisations (United Nations, OECD, European Commission, Eurostat). The National Statistics Institute (INE) regularly releases statistics about quality of life, including these kinds of indicators. This publication contains information about the multidimensional measurement of the quality of life, including a set of indicators grouped in 9 dimensions and disaggregated by individual characteristics (sex, age, type of household, type of income, level of education, nationality, level of urbanisation) and geographical scope (Autonomous Communities, EU), see [25].

The selected indicators describe the quality of life organised in nine dimensions: "Material living conditions" (I), "Work" (II), "Health" (III), "Education" (IV), "Leisure and social relations" (V), "Physical and personal security" (VI), "Governance and basic rights" (VII), "Environment" (VIII) and "Subjective well-being" (IX). The Stiglitz-Sen-Fitoussi report [25] mentions among its recommendations the necessity of considering together the effect of all the quality of life dimensions. We can use the last methodology to build an index indicator. The idea is asking the experts about the ranking order of these dimensions allowing them to be indifferent between several dimensions. To illustrate this, let us look at an example with simulated data.

In this case the answer are posets in $\mathcal{P}_9$ and we will use the distance d_M. Here we make an example of how one can use posets for this purpose. The sample in this case would be the opinion of a committee of experts. For the sake of simplicity, suppose that we have only the opinion of three experts. That is, the elements of the sample are P_1, P_2 and P_3. For generating these posets we have taken inspiration from the publicly available data about the major concerns for citizens, see [3]. These three posets are given in the next Figure 15.

Now we compute the Horvitz-Thompson estimator:

$$\widehat{\widehat{X}}_{HT} := \arg\min_{Q\in\mathcal{P}_9} \sum_{k=1}^{3} d_M^2(P_k, Q).$$

In order to get the value of $\widehat{\widehat{X}}_{HT}$, we can solve the next ILP problem:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{9}\sum_{j=1}^{9}\left(1-2\overline{M}(i,j)\right) M_Q(i,j) \\ & \\ \text{subject to} & M_Q(i,j)+M_Q(j,i)\leq 1, \qquad\qquad\quad\; i,j=1,...,9 \\ & M_Q(i,j)+M_Q(j,k)\leq 1+M_Q(i,k), \;\; i,j,k=1,...,9 \\ & M_Q(i,j)\in\{0,1\}, \qquad\qquad\qquad\quad\;\; i,j=1,...,9 \end{array}$$

Using a mathematical software we get that $\widehat{\widehat{X}}_{HT}$ is just the poset displayed in the next Figure 16.

Therefore, in this example, "Work" would be the most relevant dimension followed by "Education" and "Enviroment". From the last result, we can construct a global quality of life indicator. There

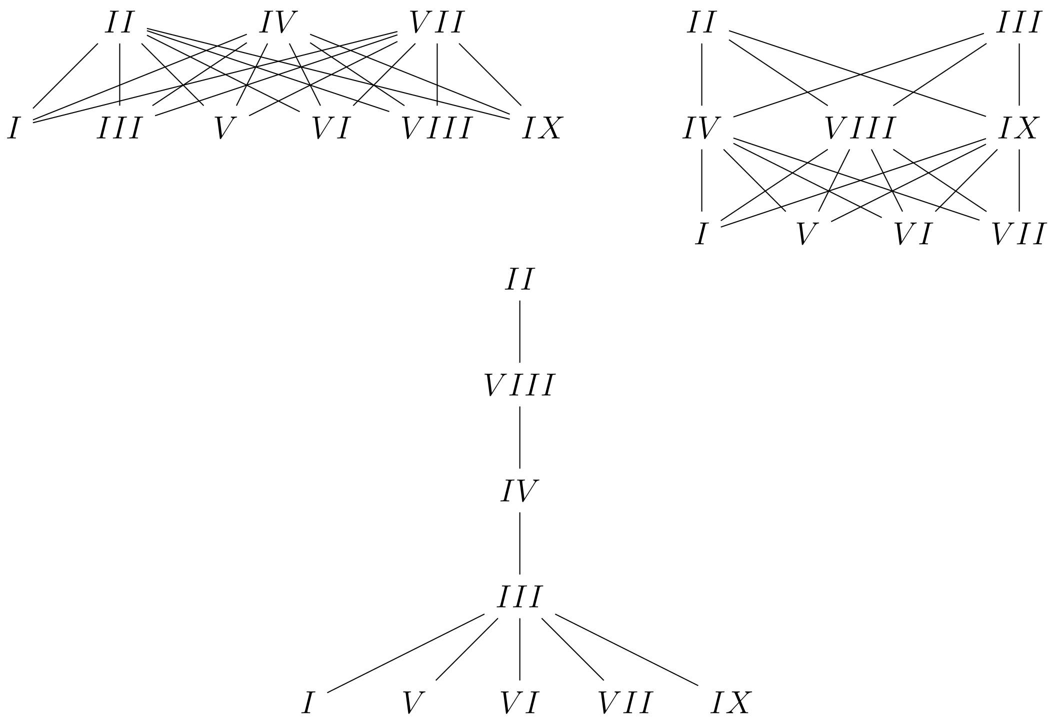

Figure 15: P_1, P_2 and P_3.

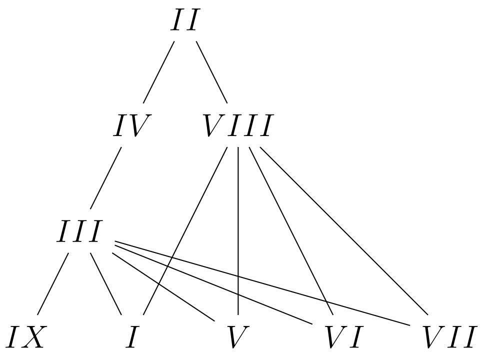

Figure 16: $\widehat{\overline{X}}_{HT}$ obtained by solving the associated ILP.

are different methodological approaches for doing this, see [25]. This index is just a way of weighing the different dimensions of quality. As it is explained in appendix A, we can construct this index by using the center of gravity of the order polytope associated to $\widehat{\overline{X}}_{HT}$. Let us call P to the poset $\widehat{\overline{X}}_{HT}$. The center of gravity c_P can be computed as follows:

$$c_P = \frac{1}{m+1} H(P),$$

where $H(P)(a) := \dfrac{1}{e(P)} \sum_{\epsilon \in \mathcal{L}(P)} h_\epsilon(a)$ being $h_\epsilon(a) = |\{x \in P : x \preceq_\epsilon a\}|$. Poset P has 384 linear extensions. Now, using the obvious order $(I, II, III, IV, V, VI, VII, VIII, IX)$, we sum the vectors having the positions of each element on each linear extension and we get:

$$H(P) = \frac{1}{384}[(1140, 3456, 2448, 2952, 1140, 1140, 1140, 2640, 1224)].$$

Finally, $c_P = \frac{1}{10}H(P)$.

We just divide c_P by its sum to get a vector which add up to 1, i.e.

$$\overline{c_P} = \frac{1}{17280}[(1140, 3456, 2448, 2952, 1140, 1140, 1140, 2640, 1224)].$$

This vector gives us coefficients of each dimension in the satisfaction index. That is,

$$I = \frac{1140}{17280}I + \frac{3456}{17280}II + \frac{2448}{17280}III + \frac{2952}{17280}IV + \frac{1140}{17280}V + \frac{1140}{17280}VI + \frac{1140}{17280}VII + \frac{2640}{17280}VIII + \frac{1224}{17280}IX \approx$$

$$\approx 0.06 \cdot I + 0.20 \cdot II + 0.14 \cdot III + 0.17 \cdot IV + 0.06 \cdot V + 0.06 \cdot VI + 0.06 \cdot VII + 0.15 \cdot VIII + 0.07 \cdot IX.$$

5 .5 Control structures in companies

In a Business Register the enterprise group is one of the fundamental statistical units which are researched and studied. An enterprise group is usually composed of many companies related to each other by control relationships. Let us call A and B to two companies of the same group. We say that A controls B directly if A has more than the 50% of the assets of B. We will denote it by $B \preceq A$. Note that there are not two companies A_1 and A_2 such that $B \preceq A_1$ and $B \preceq A_2$, otherwise A_1 and A_2 would have more than the 50% of the assets of B but this would add up to more than 100%. Considering the transitive closure of this binary relation we get a poset. Normally, there is always a company controlling the rest of the group, this is usually called the head of the group. Linking these considerations we find that the poset representing the control structure of a group is always a rooted tree, from a graph theoretic point of view. That is, this poset is connected, with one maximum and without cycles, see Figure 17. We call this poset the tree structure of a group. Two important problems arise from here. Firstly, the control structure of a group can change between two consecutive years. In some cases, these changes are so profound that led us to rethink if the current group should be considered as a completely new group with a new identifier. We can use our distance d_M to fix some threshold K and consider two groups $G_1, G_2 \in \mathcal{P}_m$ with different control structure separated groups if $d_M(G_1, G_2) > K$.

Secondly, we usually collect data from different sources and each source give us a different tree structure for a specific group. From now on, consider an example with a group with 9 companies, see again Figure 17.

In this example, we can see how the head A remains the same, however some remarkable changes have ocurred. In a more general setting we will have as many tree groups as different sources. Working with metric spaces we can define an intermediate version of tree group searching for the tree which

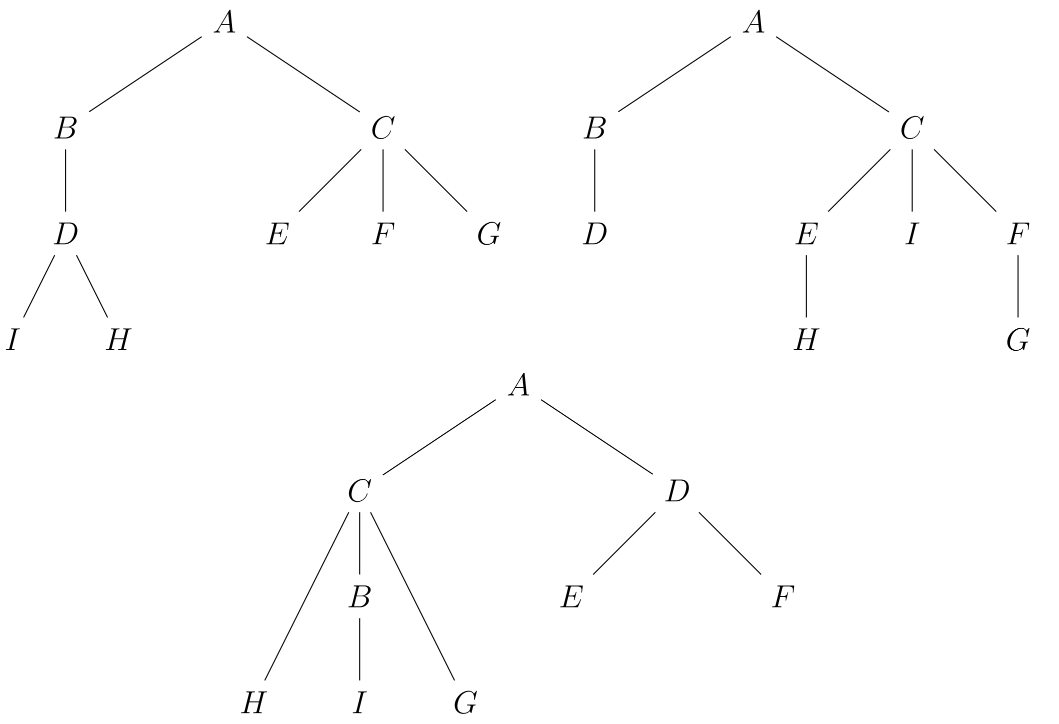

Figure 17: Three tree structures for the same group.

minimizes the distances to the given groups. Suppose that we have only the three tree structures given in Figure 17, then we can compute, for the metric space $(\mathcal{P}_9, d_M)$, the Horvitz-Thompson estimator:

$$\widehat{\overline{X}}_{HT} = \arg\min_{Q\in\mathcal{P}_9} \sum_{i=1}^{9}\sum_{j=1}^{9} \left(M_Q(i,j) - \overline{M}(i,j)\right)^2 .$$

The last minimum is reached in the next poset, see Figure 18 left. Indeed, $\lfloor \overline{M} \rceil$ is the adjacency matrix of some poset, see Corollary 3, so it is not necessary to solve the last ILP problem.

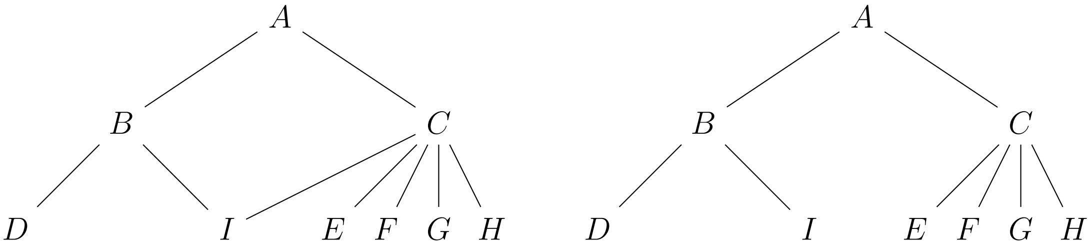

Figure 18: $\widehat{\overline{X}}_{HT}$ poset (left) and tree solution $\widehat{\overline{X}}_{HTT}$ for the problem (right).

This poset is not a tree since it has a cycle $A - B - I - C$. To obtain a tree solution we can just break the relation $I \prec C$ or $I \prec B$ in $\widehat{\overline{X}}_{HT}$, see Figure 18 right. Let us call $\widehat{\overline{X}}_{HTT}$ to this tree.

Both of these choices are tree structures with minimum objective function. In general, if we restrict ourselves to work with trees, we can modify the ILP problem associated to the Horvitz-Thompson estimator in order to get always trees. We should add the constraint $\sum_{j=1}^{m} M_Q(i,j) \leq 1$ for each i. This constraint breaks all the possible cycles in the resulting poset. Then the associated ILP is:

$$\begin{array}{lll} \text{minimize} & \sum_{i=1}^{m}\sum_{j=1}^{m}\left(1-2\overline{M}(i,j)\right)M_Q(i,j) & \\ & & \\ \text{subject to} & M_Q(i,j)+M_Q(j,i)\leq 1, & i,j=1,...,m \\ & M_Q(i,j)+M_Q(j,k)\leq 1+M_Q(i,k), & i,j,k=1,...,m \\ & \displaystyle\sum_{j=1}^{m} M_Q(i,j)\leq 1, & i=1,...,m \\ & M_Q(i,j)\in\{0,1\}, & i,j=1,...,m \end{array}$$

Now by using bootstrap estimation with $n=3, B=200$, we get $\widehat{B}(\widehat{\overline{X}}_{HT}) = 1.73$ and $\widehat{V}(\widehat{\overline{X}}_{HT}) = 4.07$. For $\alpha = 0.05$, since $d_M(x,y) \geq 1$, we obtain for trees

$$\boldsymbol{P}\left(d_M(\widehat{\overline{X}}_{HTT},\overline{X}) < \frac{\sqrt{\widehat{V}(\widehat{\overline{X}}_{HTT})+\widehat{B}(\widehat{\overline{X}}_{HTT})}}{\sqrt{\alpha}}\right) \geq 1-\alpha \Leftrightarrow \boldsymbol{P}\left(d(\widehat{\overline{X}}_{HTT},\overline{X}) < 16.75\right) \geq 0.95.$$

Then the distance between the estimator $\widehat{\overline{X}}_{HTT}$ and the population mean, seen as a tree, $\overline{X}$ is lower than 16.75 with a probability of 95%. Furthermore, $\widehat{\overline{X}}_{HTT}$ or any of the three initial trees structures are likely to be the population mean.

6 Conclusions and open problems

This work deals with the problem of using statistical sampling on spaces where no algebraic operations are defined in advance. We have seen that most of the usual statistical properties hold for metric spaces. Moreover, we are able to use statistical inference techniques on these sets. In addition, we have studied the specific case of variables with partially ordered values. We have analysed different ways of defining metrics over posets. Of course, other new distances could be defined and analysed. We have also studied the complexities to compute estimators associated to these distances. We have seen plenty of examples illustrating different applications of this theory to official statistics. Of course, many others interesting and useful applications could be studied. In the future, it would be interesting to use the ideas of this paper in a specific area of official statistics. This way we could study in depth the advantages and disadvantages of this theory for a specific example and develop a full methodology using the presented theory. For this, we would need in some cases a concrete question in the questionnaire asking about the poset of preferences.

From a mathematical point of view, there are many interesting open problems that could be studied further. About the sampling theory on metric spaces, it would be interesting to discover general conditions for the Horvitz-Thompson estimator to be unbiased on metric spaces. In the same way, replicating some of these results for the general case with different probabilities is a very interesting task. Also, developing other bias and variance estimation techniques in metric spaces would be engaging. About the distances between posets, finding a way of making operative the internal distance between posets remains as an open problem. This distance perfectly captures the idea of poset distance at the expense of a higher complexity. From a geometrical point of view, computing all the vertices of $\mathcal{PL}_m$ polytope is an open problem with a lot of interesting applications in order theory.

Acknowledgements

This work would not have been possible without the help of Sara Carrascosa (Economically Active Population Survey, INE), Alicia Fernández (Dissemination Department, INE) and Esteban Barbado (Bussines Register Department, INE). Thank you for your disinterested assistance and expert knowledge.

A Polyhedra and polytopes

In this appendix, we summarize the fundamentals of polyhedral combinatorics. For a more detailed treatment see [26, 15]. A **halfspace** in $\mathbb{R}^m$ is a set of the form $\{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{\alpha}^T \cdot \boldsymbol{x} \leq c\}$ for some vector $\boldsymbol{\alpha} \in \mathbb{R}^m$ and $c \in \mathbb{R}$. A **polyhedron** is the intersection of finitely many halfspaces: $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{A} \cdot \boldsymbol{x} \leq \boldsymbol{b}\}$. A **polytope** is a bounded polyhedron. We can define a polytope in another way. The convex hull of a finite set of points S, denoted by $Conv(S)$, is the set of all convex combinations of its points, i.e.

$$Conv(S) = \left\{ \sum_{i=1}^{|S|} \alpha_i x_i \mid \alpha_i \geq 0 \ \wedge \ \sum_{i=1}^{|S|} \alpha_i = 1 \right\}.$$

A polytope can always be written as the convex hull of a finite set of points. These points are called the **vertices** of the polytope. Remember that a **face** of a polytope $\mathcal{P}$ is defined as a subset $\mathcal{F} \subseteq \mathcal{P}$ such that there exists a vector $\boldsymbol{\alpha}$ and a constant $c \in \mathbb{R}$ such that

$$\boldsymbol{\alpha}^T \cdot \boldsymbol{x} \leq c, \ \forall \boldsymbol{x} \in \mathcal{P} \quad \text{and} \quad \mathcal{F} = \mathcal{P} \cap \{\boldsymbol{\alpha}^T \cdot \boldsymbol{x} = c\}.$$

We will denote the face defined via $\boldsymbol{\alpha}$ and c by $\mathcal{F}_{\boldsymbol{\alpha},c}$. Equivalently, a face $\mathcal{F}$ is also characterized by the vertices of $\mathcal{P}$ in $\mathcal{F}$. Obviously, the face of a polytope is a polytope. Recall that the **dimension** of a polytope $\mathcal{P}$ is defined as the dimension of the smallest affine subspace containing its vertices, denoted aff($\mathcal{P}$) (for more details, see [26]). One should consult [26, 15], for pictures of polyhedra and polytopes. Figure 19 shows the picture of a polytope whose faces are all pentagons. This polytope is called a dodecahedron. The dodecahedron has 12 faces, 30 edges and 20 vertices.

We say that a polytope $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{A} \cdot \boldsymbol{x} \leq \boldsymbol{b}\}$ is integer if it has integer vertices. Integer vertices are closely related to **totally unimodular matrices**. A matrix $\boldsymbol{A}$ is totally unimodular if every square submatrix has determinant $0, +1$, or -1. In particular, this implies that all entries are

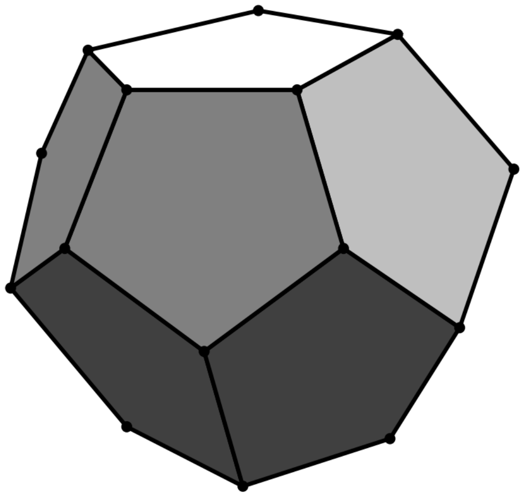

Figure 19: Example of a polytope (a dodecahedron).

0 or ±1. Totally unimodular matrices are very well behaved, because they always define polytopes with integer vertices, as long as the right-hand side is integer-valued.

Theorem 7. *If $\boldsymbol{A}$ is totally unimodular and $\boldsymbol{b}$ is an integer vector, then $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{A} \cdot \boldsymbol{x} \leq \boldsymbol{b}\}$ has integer vertices.*

Proposition 10. *Let $\boldsymbol{A}$ be a totally unimodular matrix. Multiplying any row or column of $\boldsymbol{A}$ by -1 results in a totally unimodular matrix.*

Proposition 11. *Let $\boldsymbol{A}$ be a totally unimodular matrix. Then the following matrices are all totally unimodular:*

$$-\boldsymbol{A},\ \boldsymbol{A}^T,\ [\boldsymbol{A}, \boldsymbol{I}],\ [\boldsymbol{A}, -\boldsymbol{A}].$$

Proposition 12. *Let $\boldsymbol{A}$ be a totally unimodular matrix. Then the following polyhedrons are all integral for any vectors $\boldsymbol{b}$ and $\boldsymbol{u}$ of integers:*

$$\{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{A} \cdot \boldsymbol{x} \leq \boldsymbol{b}\}$$

$$\{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{A} \cdot \boldsymbol{x} \geq \boldsymbol{b}\}$$

$$\{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{A} \cdot \boldsymbol{x} \leq \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0}\}$$

$$\{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{A} \cdot \boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0}\}$$

$$\{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{A} \cdot \boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{0} \leq \boldsymbol{x} \leq \boldsymbol{u}\}$$

Finally, we state some key theorems about totally unimodular matrices [2].

Theorem 8. ***Ghouila-Houri's Characterization.*** *An $m \times n$ integral matrix $\boldsymbol{A}$ is totally unimodular if and only if for each set $R \subseteq \{1, 2, \cdots, m\}$ can be divided into two disjoint sets R_1 and R_2 such that*

$$\sum_{i \in R_1} a_{ij} - \sum_{i \in R_2} a_{ij} \in \{-1, 0, 1\},\ j = 1, 2, \cdots, n.$$

Applying the Ghouila-Houri's Characterization to $\boldsymbol{A}^T$ we get the same results for columns.

Corollary 4. *An $m \times n$ integral matrix $\boldsymbol{A}$ is totally unimodular if and only if for each set $C \subseteq \{1, 2, \cdots, n\}$ can be divided into two disjoint sets C_1 and C_2 such that*

$$\sum_{j \in C_1} a_{ij} - \sum_{j \in C_2} a_{ij} \in \{-1, 0, 1\},\ i = 1, 2, \cdots, m.$$

Theorem 9. ***Hoffman-Gale sufficient conditions.***
An $(0, +1, -1)$ integral matrix $\boldsymbol{A}$ is totally unimodular if both of the following conditions are satisfied:

- *Each column contains at most two nonzero elements.*
- *The rows of $\boldsymbol{A}$ can be patitioned into two sets R_1 and R_2 such that two nonzero entries in a column are in the same set of rows if they have different signs and in different sets of rows if they have the same sign.*

There is a special kind of polytope that is very useful for our purposes [24]. Given a poset P with m elements, it is possible to associate to P a polytope $\mathcal{O}(P)$ over $\mathbb{R}^m$, called the **order polytope** of P, formed by the m-uples f of real numbers indexed by the elements of P satisfying

- $0 \leq f(a) \leq 1$ for every a in P,
- $f(a) \leq f(b)$ whenever $a \preceq b$ in P.

Our interest in order polytopes comes from the fact that there are known results characterizing the vertices and many other properties of the polytope $\mathcal{O}(P)$ in terms of the structure of the subjacent poset P. So we can learn interesting things about P from $\mathcal{O}(P)$ and vice versa. It can be seen [24] that the vertices of $\mathcal{O}(P)$ are the characteristic functions of filters of P. Similarly, it can be seen [5] that two vertices of an order polytope whose associated filters are F_1 and F_2 are adjacent vertices if and only if the symmetric difference $F_1 \Delta F_2 := (F_1 \setminus F_2) \uplus (F_2 \setminus F_1)$ is a connected subposet of P. Figure 20 shows an example of order polytope.

The volume of an order polytope $\mathcal{O}(P)$ depends on the number of linear extension of P in the next way: $Vol(\mathcal{O}(P)) = \dfrac{e(P)}{m!}$, where $|P| = m$. The center of gravity c_P of this polytope has appealing applications. It can be computed as follows:

$$c_P = \frac{1}{m+1} H(P),$$

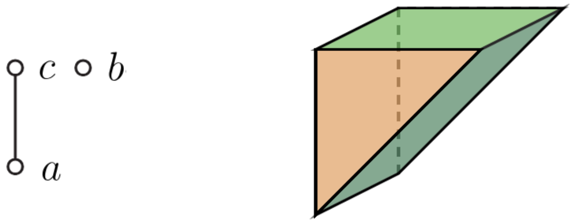

Figure 20: A poset (left) and its associated order polytope (right).

where $H(P)(a) := \dfrac{1}{e(P)} \sum_{\epsilon \in \mathcal{L}(P)} h_\epsilon(a)$ being $h_\epsilon(a) = |\{x \in P : x \preceq_\epsilon a\}|$. In other words, $h_\epsilon(a)$ is the height of a in the linear extension ϵ [21]. The next example explains how to compute the center of gravity of an order polytope and some applications.

Example 9. *Consider the labeled poset V of Figure 21 with three elements. $\mathcal{O}(V)$ is a square pyramid. To compute c_V we first should compute $h_\epsilon(a)$ for each linear extension ϵ. Poset V has just 2 linear extensions $\epsilon_1 = (1, 2, 3)$ and $\epsilon_2 = (1, 3, 2)$. Then, the position of element 1 in each extension is $h_{\epsilon_1}(1) = 1$ and $h_{\epsilon_2}(1) = 1$. The rest of values are $h_{\epsilon_1}(2) = 2, h_{\epsilon_1}(3) = 3, h_{\epsilon_2}(2) = 3$ and $h_{\epsilon_2}(3) = 2$. Therefore,*

$$H(V) = \frac{1}{2}\left[(1, 2, 3) + (1, 3, 2)\right] = \frac{1}{2}(2, 5, 5).$$

Finally, $c_V = \dfrac{1}{4} H(V) = (2/8, 5/8, 5/8)$.

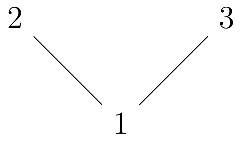

Figure 21: Labeled poset V.

Now imagine that we want to share a total quantity of resources among the elements of a poset according to the set of preferences given by the poset, for that we can use c_V. To put it in concrete terms, suppose that a trader is investing money in three financial products 1, 2 and 3. The trader prefers 2 and 3 to 1 and is indifferent between 2 and 3 as poset V models. He wants to know how much money should be invested in each product according to these preferences. Then the point c_V tells us an excellent way of sharing the total amount of money. We just divide c_V by its sum to get a vector which add up to 1, i.e. $\overline{c_V} = (2/12, 5/12, 5/12)$. This point is the center of gravity of the preferences seen as an order polytope. Finally, the trader should invest $2/12$ of the money in 1, and $5/12$ in 2 and 3.

B Linear and integer programming

Operations Research is the discipline that uses a scientific approach to decision making. It seeks to determine how best to design and operate a system, usually under conditions requiring the allocation of scarce resources, by means of quantitative methods. Some example within mathematical optimization are production planning, inventory control, packing problems, and assignment problems. To solve this kind of problems we should first give a mathematical model of the situation. In order to do this, the decisions are identified and decision variables are defined. For each decision it is defined a decision variable of suitable type (continuous, integer valued, binary) according to the particular needs. The next step is formulating the objective function to compute the benefit/cost in terms of decision variables and parameters. Finally, the constraints indicating the interplay between the different variables must be expressed in mathematical terms. When the objective function and the constraints are linear, the problem is called a linear programming problem (LP problem). Linear programs are problems that can be expressed in canonical form as

$$\begin{array}{ll} \text{maximize} & \boldsymbol{c}^T\boldsymbol{x} \\ \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0} \end{array}$$

where $\boldsymbol{x}$ represents the vector of variables (to be determined), $\boldsymbol{c}$ and $\boldsymbol{b}$ are vectors of (known) coefficients, $\boldsymbol{A}$ is a (known) matrix of coefficients.

Theorem 10. ***(Fundamental Theorem of Linear Programming.)***
Given a linear programming problem:

$$\begin{array}{ll} \textit{maximize} & \boldsymbol{c}^T\boldsymbol{x} \\ \textit{subject to} & \boldsymbol{x} \in \mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{A} \cdot \boldsymbol{x} \leq \boldsymbol{b}\}. \end{array}$$

If $\mathcal{P}$ is a bounded polyhedron and not empty and $\boldsymbol{x}^*$ is an optimal solution to the problem, then:

- $\boldsymbol{x}^*$ *is an extreme point (vertex) of $\mathcal{P}$, or*
- $\boldsymbol{x}^*$ *lies on a face $\mathcal{F} \subset \mathcal{P}$ of optimal solutions.*

The **simplex method** or (**simplex algorithm**) was the first algorithm to solve linear programming problems and was proposed in 1947 by George Dantzig. This algorithm proceeds by iterating through feasible solutions that are vertices of the polyhedron that represents the feasibility region. Finally, it will use an optimality condition to terminate. A few exceptions may occur, they determine initial infeasibility, unboundedness, more than one solution and cycling in case of degenerancies. The simplex method is remarkably efficient in practice and was a great improvement over earlier methods such as Fourier-Motzkin elimination.

However, in 1972, Klee and Minty gave an example, the Klee-Minty cube, showing that the worst-case complexity of simplex method as formulated by Dantzig is exponential time. The simplex algorithm has polynomial-time average-case complexity under various probability distributions, with the precise average-case performance of the simplex algorithm depending on the choice of a probability

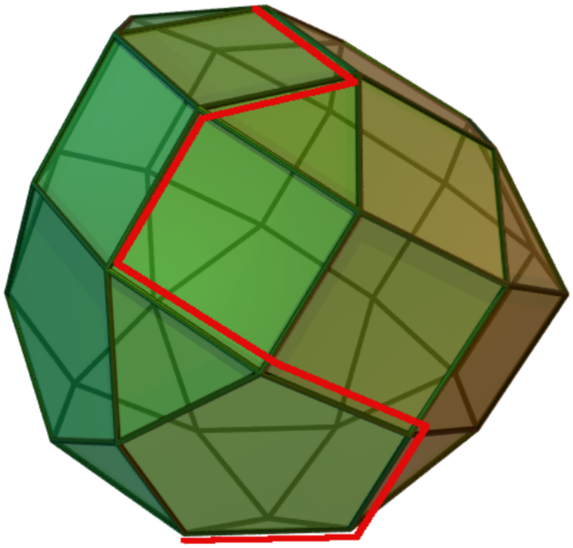

Figure 22: Simplex search in a generic polytope.

distribution for the random matrices. In Figure 22 we can see an example of searching process in a generic polytope.

An **integer programming problem** is a mathematical optimization problem in which some or all of the variables are restricted to be integers. In many settings the term refers to integer linear programming (ILP), in which the objective function and the constraints (other than the integer constraints) are linear. Integer programming is NP-complete. For this reason, it is preferable to work with linear programing problems than ILP problems.

References

[1] L. Bohdan. Distances between partially ordered sets. *Mathematica Bohemica*, 2:167–170, 1991.

[2] P. Camion. Characterization of Totally Unimodular Matrices. *Proceedings of the American Mathematical Society*, 16:1068–1073, 1965.

[3] Barómetro de enero 2019. Estudio n° 3238. *Centro de Investigaciones Sociológicas, CIS*, 2019.

[4] E. F. Combarro, I. Díaz, and P. Miranda. On random generation of fuzzy measures. *Fuzzy Sets and Systems*, 228:64–77, 2013.

[5] E. F. Combarro and P. Miranda. On some theoretical results relating random generation of fuzzy measures. In *Proceedings of 11th Int. Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'06)*, pages 1678–1675, Paris (France), 2006.

[6] B.A. Davey and H.A. Priestley. *Introduction to lattices and order*. Cambridge University Press, 2002.

[7] B. Bouyssou, T. Marchant and P. Perny. *Social Choice Theory and Multicriteria Decision Aiding*. Decision-making Process: Concepts and Methods, Chapter 19, Wiley Online Library, 2009.

[8] B. Efron and J. Tibshirani. *An Introduction to the Bootstrap.* Monographs on Statistics and Applied Probability, 1993.

[9] M. Fattore and R. Bruggemann. *Partial Order Concepts in Applied Sciences.* Springer, 2017.

[10] *Encuesta de Fecundidad 2018. Metodología.*. Instituto Nacional de Estadística, 2018.

[11] Eurostat. Methodologies and Working papers. *NACE Rev. 2 Statistical classification of economic activities in the European Community.* Eurostat, European comission, 2008.

[12] P. García-Segador and P. Miranda. Bottom-Up: a New Algorithm to Generate Random Linear Extensions of a Poset. *Order, Springer*, 1–26, 2018.

[13] P. García-Segador and P. Miranda. Applying Young diagrams to 2-symmetric fuzzy measures with an application to general fuzzy measures. *Fuzzy Sets and Systems*, 2019.

[14] M. Grabisch. *Set functions, games and capacities in Decision Making*, volume 46 of *Theory and Decision Library.* Springer, 2016.

[15] B. Grümbaum. *Convex Polytopes.* Springer-Verlag, 2003.

[16] M. Grabisch and J.L. Marichal and R. Mesiar. *Aggregation functions.* Cambridge University Press, 2009.

[17] M. Grabisch, T. Murofushi, and M. Sugeno. *Fuzzy Measures and Integrals- Theory and Applications.* Number 40 in Studies in Fuzziness and Soft Computing. Physica–Verlag, Heidelberg (Germany), 2000.

[18] J. Neggers and H. S. Kim. *Basic posets.* World Scienctific, 1998.

[19] Quality guidelines of the INE. *Instituto Nacional de Estadística, (INE)*, 2015.

[20] INE User Satisfaction Survey. Year 2016 *Instituto Nacional de Estadística, (INE)*, 2017.

[21] J. Matoušek. Lectures on Discrete Geometry. *Springer*, 2002.

[22] E. Särndal, B. Swensson, and J. Wretman. Model Assisted Survey Sampling. *Springer Series in Statistics*, 2003.

[23] R. Stanley. *Enumerative Combinatorics.* Cambridge University Press, Cambridge (UK), 2012.

[24] R. Stanley. Two poset polytopes. *Discrete Comput. Geom.*, 1(1):9–23, 1986.

[25] J.E. Stiglitz, A. Sen and J.P. Fitoussi Report by the Commission on the Measurement of Economic Performance and Social Progress *European Commission*, 2008.

[26] G. Ziegler. *Lectures on Polytopes.* Springer-Verlag, 1995.