

Article

# Optimizing Allocation Rules in Discrete and Continuous Discriminant Analysis

Dário Ferreira \*  and Sandra S. Ferreira 

Department of Mathematics, Center of Mathematics, University of Beira Interior, 6201-001 Covilhã, Portugal; sandraf@ubi.pt

\* Correspondence: dario@ubi.pt

**Abstract:** This paper presents an approach for the study of probabilistic outcomes in experiments with multiple possible results. An approach to obtain confidence ellipsoids for the vector of probabilities, which represents the likelihood of specific results, for both discrete and continuous discriminant analysis, is presented. The obtention of optimal allocation rules, in order to reduce the allocation costs is investigated. In the context of discrete discriminant analysis, the approach focuses on assigning elements to specific groups in an optimal way. Whereas in the continuous case, the approach involves determining the regions where each action is the optimal choice. The effectiveness of the proposed approach is examined with two numerical applications. One of them uses real data, while the other one uses simulated data.

**Keywords:** condence ellipsoids; discriminant analysis; inference; limit distributions

**MSC:** 62H11; 62H30; 62F15; 62E20



**Citation:** Ferreira, D.; Ferreira, S.S. Optimizing Allocation Rules in Discrete and Continuous Discriminant Analysis. *Mathematics* **2024**, *12*, 1173. <https://doi.org/10.3390/math12081173>

Academic Editor: Andrea De Gaetano

Received: 18 March 2024

Revised: 5 April 2024

Accepted: 10 April 2024

Published: 13 April 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The process of allocating mixed data with several possible outcomes begins with the identification of the data. This includes identifying the type of data, the source of the data, and any relevant characteristics. Once the data have been identified, the next step is to determine the type of allocation that will be used. There are several different types of allocation, including random, manual, and automated allocation. Moreover, there are multiple methods for examining data with categorical outcomes. These methods include descriptive statistics, bivariate analyses, log-linear regression, multinomial logistic regression, discriminant analyses and quadratic discriminant analysis.

Descriptive statistics are used to summarize the characteristics of a data set.

Bivariate analysis is used to examine the relationship between two variables. In the case of categorical data, bivariate analysis consists of calculating the frequency or percentage of one variable for each category of the other variable.

The purpose of log-linear models is to study the relationship between an outcome variable and one or more explanatory variables, where the outcome variable is expressed as the logarithm of a linear combination of the explanatory variables. The log-linear model is a frequently used and simple structure for a contingency table, see Hand and Christen [1]. It is based on the same principle as the analysis of variance models, as outlined in Birch [2], Fienberg and Rinaldo [3] or Goodman [4], and is formed by calculating the logarithms of the cell probabilities. Examples of log-linear models with frequency tables can be found in Haberman [5].

On the other hand, Multinomial logistic regression is a generalized linear model that models the log-odds of the categorical response being true as a linear combination of predictor variables. It is a significant tool in statistical modelling and prediction, extending the capabilities of the binary logistic regression model to situations with more than two outcomes.

The goal of discriminant analysis is to classify individuals into distinct groups, maximizing the difference between groups and minimizing the variation within groups. Discriminant analysis has many advantages over the other mentioned methods. For example, it is able to identify and measure the effects of one or more independent variables on a dependent variable, making it a more powerful tool for predicting the outcome of a certain event. This method was first introduced by Sir Ronald Fisher in 1936, in Fisher [6], and is a popular research topic. Just to give a few examples over time, Rao [7] applied discriminant analysis to two types of problems confronted in biological research, while Friedman [8] proposed alternatives to the usual maximum likelihood estimates for the covariance matrices in linear and quadratic discriminant analysis, in small sample high-dimensional settings. McFarland et al. [9] used the theory of Bessel functions to derive stochastic representations for the exact distributions of the “plug-in” quadratic discriminant functions for classifying a newly obtained observation. Discriminant analysis was used in Perriere and Thioulouse [10] to separate Gram negative bacteria proteins according to their subcellular location. The problem of classifying an individual into one of several populations based on mixed nominal, continuous, and ordinal data was studied in Flury, Bourkai and Flury [11]. Modern research continues to find discriminant analysis helpful in problem-solving. For example, using dynamic feature extraction along with quadratic discriminant analysis classifier can significantly enhance fault classification and diagnosis in dynamic nonlinear processes, as demonstrated by Li, Jia, and Mao [12]. Furthermore, Tong et al. [13] demonstrated the effectiveness of discriminant analysis in their work on bearing fault diagnosis, representing a crucial development in fault diagnosis for dynamic nonlinear processes. For more information on discriminant analysis, see McLachlan [14].

Discriminant analysis can be divided into two categories: continuous and discrete. In the continuous case, the predictor variables are continuous and the output of the analysis is a continuous function. In the discrete case, the predictor variables are discrete and the output of the analysis is a classification. In both cases, discriminant analysis is used to find the optimal allocation rules. In the continuous case, the optimal allocation rule is determined by finding the optimal separating hyperplane which maximizes the difference between the groups. In the discrete case, the optimal allocation rule is determined by finding the optimal decision tree, which maximizes the accuracy of the classification.

A unified approach, providing optimal allocation rules to minimize expected costs for both continuous and mixed cases, is presented in Ferreira et al. [15]. We will follow this last approach.

Our goal is to obtain confidence ellipsoids for the vector of probabilities,  $\mathbf{p}$ , of getting particular results in an experiment with  $m$  possible outcomes. Through duality, we will also test hypotheses. In addition, we will use support planes for ellipsoids, see Schott [16], to obtain simultaneous confidence intervals for these probabilities. Moreover, we will generalize these results to vectors

$$\Psi = \mathbf{G}\mathbf{p}, \quad (1)$$

where  $\mathbf{G}$  is a known matrix. Furthermore, we will show how to obtain optimal allocation rules, in order to reduce the allocation costs, first for the discrete case and then for the continuous case.

The rest of this paper is organized as follows. The next section, for preliminary results, is divided in two subsections. In the first one, we provide a simplified form of the continuous mapping theorem, CMT, based on Kallenberg [17]. This theorem will grant us the ability to derive confidence ellipsoids from the  $N(\mathbf{0}, W(\mathbf{p}))$  distribution in the second subsection, where we will follow the approach of Scheffé [18]. Section 3 begins with the study of individual samples, followed by the study of pairs and structured families of independent samples. These samples refer to the treatments of a fixed effects experiment. The optimal allocation rules will be obtained in Section 4. Section 5 contains two numerical applications. The first, referring to the discrete case and using real data; the second, referring to the continuous case and using simulated data. We end the paper with some conclusions.

## 2. Preliminary Results

### 2.1. Asymptotic Distributions

The Central Limit Theorem, CLT, serves as a fundamental result in probability theory and statistics, providing insights into the behavior of sample averages. Let  $Y_n$  be a sequence of random variables that converges in distribution to a limiting random variable  $Y$ , denoted as  $Y_n \xrightarrow{d} Y$ . Additionally, consider continuous functions  $g(Y_n)$  and  $g(Y)$ . The CLT, as established by Kallenberg [17], asserts that under these conditions,  $g(Y_n)$  converges in distribution to  $g(Y)$ , i.e.,  $g(Y_n) \xrightarrow{d} g(Y)$ .

Moreover, if the convergence in distribution of  $Y_n$  to  $Y$ , given by  $F_{Y_n}(\mathbf{y}) \rightarrow F_Y(\mathbf{y})$ , is uniform on a compact set  $L$ , then the convergence in distribution of  $g(Y_n)$  to  $g(Y)$  is also uniform on the same set  $L$ . This uniform convergence property is valuable in understanding the stability of the distributional characteristics of the transformed random variables.

Further insights can be gained by considering the continuity of the function  $g$ . If  $g$  is a continuous function, then for any given  $\epsilon > 0$ , there exists a compact set  $K(\epsilon)$  such that the probability measure associated with  $F_{Y_n}(\mathbf{y})$ , denoted as  $\phi_n(\cdot)$ , satisfies  $\phi_n(K(\epsilon)) \geq 1 - \epsilon$ . Notably, the Cartesian product

$$\dot{K}(\epsilon) = K(\epsilon) \times L \tag{2}$$

is also a compact set. Consequently, the convergence in distribution of  $g(Y_n)$  to  $g(Y)$  is uniform on the set  $\dot{K}(\epsilon)$ .

In summary, the asymptotic distribution of  $g(Y_n)$  can be expressed as  $F_{g(Y)}(\mathbf{x})$ , provided that  $\mathbf{x}$  is a point of continuity for  $F_{g(Y)}$ . This convergence is established through the concept of weak convergence of probability measures, highlighting the broader applicability of the CLT in understanding the behavior of transformed random variables in a distributional sense.

### 2.2. Confidence Ellipsoids

Confidence ellipsoids serve as valuable tools for visualizing and quantifying uncertainty in multivariate data, providing insights into the variability of observations. These ellipsoids are particularly useful for testing hypotheses, constructing simultaneous confidence intervals, and making predictions in statistical analysis.

Consider a set of observations modeled by the linear equation  $Y = X\beta + \epsilon$ , where  $X$  is a design matrix,  $\beta$  is a vector of fixed effects, and  $\epsilon$  is a vector of errors. The associated confidence ellipsoid, denoted as  $E_q$ , can be expressed as

$$E_q = \{Y - \mu \in \Omega : (Y - \mu)^\top C^{-1}(Y - \mu) \leq x_{h,1-q}\}, \tag{3}$$

where  $\mu = X\beta$  and  $C = X^\top X$ . Here,  $\Omega$  represents the feasible region of the ellipsoid, and  $x_{h,1-q}$  is a critical value from a chi-square distribution with  $h$  degrees of freedom, ensuring a desired confidence level  $1 - q$ .

The confidence ellipsoid  $E_q$  possesses notable properties. Firstly, it is a convex set, implying that any point within the ellipsoid is a convex combination of points on its boundary. This property facilitates a comprehensive understanding of the distribution of observations.

Secondly, the confidence ellipsoid aids in constructing simultaneous confidence intervals. For a vector of interest,  $\tau$ , the simultaneous confidence interval for  $\mu^\top Y$  is given by

$$\tau^\top (Y - \mu) \leq \sqrt{x_{h,1-q} \tau^\top C \tau}. \tag{4}$$

This interval provides a range for the linear combination  $\tau^\top Y$  with a specified level of confidence.

Finally, confidence ellipsoids are instrumental in hypothesis testing. Consider the following null hypothesis:

$$H_0 : \tau^\top \mu = a_0. \tag{5}$$

If the hypothesis  $H_0$  is true, then  $E_q$  will contain the point  $(Y - \mu)_{a_0} = (\tau^\top Y - \tau^\top \mu)$ ; otherwise, the hypothesis is rejected.

### 3. Inference

In statistical inference, we often encounter experiments with multiple possible outcomes, each associated with a certain probability. To quantify the likelihood of obtaining specific results in such experiments, given the probabilities  $p_1, \dots, p_m$  of each outcome, and the observed frequencies  $n_1, \dots, n_m$  in  $n$  trials, we can employ the following probability function:

$$pr\left(\bigcap_{i=1}^m (N_i = n_i)\right) = \frac{n!}{\prod_{i=1}^m n_i!} \prod_{i=1}^m p_i^{n_i} \xrightarrow{d} M(|n, p), \tag{6}$$

where  $M$  represents a singular multivariate distribution.

Consider the probability vector  $\mathbf{p} = (p_1, \dots, p_m)$ , and define the vector of estimators:

$$\tilde{\mathbf{p}}_n = (\tilde{p}_{n,1}, \dots, \tilde{p}_{n,m}), \tag{7}$$

with  $\tilde{p}_{n,i} = \frac{n_i}{n}$  for  $i = 1, \dots, m$ . These estimators represent the observed probabilities of each outcome. As the number of trials  $n$  approaches infinity, Wilks' theorem, in Wilks [19], provides insight into the asymptotic behavior of the estimators:

$$\sqrt{n}(\tilde{\mathbf{p}}_n - \mathbf{p}) \xrightarrow{d} N(\mathbf{0}, W(\mathbf{p})), \tag{8}$$

where  $\xrightarrow{d}$  denotes convergence in distribution, and  $N(\mathbf{0}, W(\mathbf{p}))$  represents a normal distribution with zero mean vector and covariance matrix  $W(\mathbf{p})$ .

The covariance matrix  $W(\mathbf{p})$  is defined as

$$W(\mathbf{p}) = D(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top, \tag{9}$$

where  $D(\mathbf{p})$  is a diagonal matrix with the components of  $\mathbf{p}$  along its diagonal. The probabilities in the initial distribution can be estimated from the observed frequencies through maximum likelihood estimation, where each  $\tilde{p}_{n,i} = \frac{n_i}{n}$  for  $i = 1, \dots, m$ . This result indicates that as the sample size grows, the distribution of the estimated probabilities becomes approximately normal, facilitating the application of classical statistical inference techniques. The covariance matrix  $W(\mathbf{p})$  characterizes the variability in the estimates, and its asymptotic normality allows for the construction of confidence intervals and hypothesis tests based on the estimated probabilities  $\tilde{\mathbf{p}}_n$ .

#### 3.1. One Sample

Let  $\alpha_1, \dots, \alpha_m$  constitute an orthonormal basis for the orthogonal complement,  $\Omega^\perp$ , of  $\Omega$ . Then, we can write

$$W(\tilde{\mathbf{p}}_n) = \sum_{j=1}^m v_j \alpha_j \alpha_j^\top, \tag{10}$$

where  $v_j$  is the eigenvalue associated with  $\alpha_j$ , see Horn [20]. We can further decompose the matrix  $W(\tilde{\mathbf{p}}_n)$  as

$$W(\tilde{\mathbf{p}}_n) = \sum_{j=1}^{m-1} v_j \alpha_j \alpha_j^\top + v_m \alpha_m \alpha_m^\top. \tag{11}$$

Let  $\alpha_m$  be the eigenvector associated with the largest eigenvalue  $v_m$ . Then, we can write

$$W(\tilde{\mathbf{p}}_n)^+ = \sum_{j=1}^{m-1} v_j^{-1} \alpha_j \alpha_j^\top + v_m^{-1} \alpha_m \alpha_m^\top, \tag{12}$$

where  $+$  denotes the Moore–Penrose inverse, expressed as

$$W(\tilde{\mathbf{p}}_n)^+ = \sum_{j=1}^m v_j^{-1} \boldsymbol{\alpha}_j \boldsymbol{\alpha}_j^\top. \tag{13}$$

According to Schott [16] and attending to the continuity of the Moore–Penrose inverse, if

$$W(\tilde{\mathbf{p}}_n) \xrightarrow{p} W(\mathbf{p}), \tag{14}$$

where  $\xrightarrow{p}$  denotes convergence in probability, then

$$W(\tilde{\mathbf{p}}_n)^+ \xrightarrow{p} W(\mathbf{p})^+. \tag{15}$$

Moreover, the continuity of the Moore–Penrose inverse implies that the covariance matrix of the vector  $\sqrt{n}(\tilde{\mathbf{p}}_n - \mathbf{p})$ ,  $\Sigma(\sqrt{n}(\tilde{\mathbf{p}}_n - \mathbf{p}))$  will converge in probability to  $W(\mathbf{p})$ ,

$$\Sigma(\sqrt{n}(\tilde{\mathbf{p}}_n - \mathbf{p})) \xrightarrow{p} W(\mathbf{p}). \tag{16}$$

Furthermore, the continuity of the Moore–Penrose inverse implies that, if  $\mathbf{c}$  is a vector, the inner product  $\mathbf{c}^\top W(\tilde{\mathbf{p}}_n)\mathbf{c}$  will converge in probability to the inner product  $\mathbf{c}^\top W(\mathbf{p})\mathbf{c}$ , i.e.,

$$\mathbf{c}^\top W(\tilde{\mathbf{p}}_n)\mathbf{c} \xrightarrow{p} \mathbf{c}^\top W(\mathbf{p})\mathbf{c} \tag{17}$$

and the variance of the vector  $\sqrt{n}\mathbf{c}^\top(\tilde{\mathbf{p}}_n - \mathbf{p})\mathbf{c}$  will also converge in probability to the inner product  $\mathbf{c}^\top W(\mathbf{p})\mathbf{c}$ , i.e.,

$$\text{Var}\left(\sqrt{n}\mathbf{c}^\top(\tilde{\mathbf{p}}_n - \mathbf{p})\mathbf{c}\right) \xrightarrow{p} \mathbf{c}^\top W(\mathbf{p})\mathbf{c}. \tag{18}$$

These properties of the Moore–Penrose inverse can be used to draw statistical inference from the data. For example, the covariance matrix of the vector  $\sqrt{n}(\tilde{\mathbf{p}}_n - \mathbf{p})$  can be used to estimate the variance of the vector  $\mathbf{p}$ , and the inner product  $\mathbf{c}^\top W(\tilde{\mathbf{p}}_n)\mathbf{c}$  can be used to estimate the inner product  $\mathbf{c}^\top W(\mathbf{p})\mathbf{c}$ . Besides this, we have

$$\begin{cases} \sqrt{n}(\tilde{\mathbf{p}}_n - \mathbf{p}) \rightsquigarrow N(\mathbf{0}, W(\mathbf{p})) \\ \sqrt{n}(\mathbf{c}^\top \tilde{\mathbf{p}}_n - \mathbf{c}^\top \mathbf{p}) \rightsquigarrow N(0, \mathbf{c}^\top W(\mathbf{p})\mathbf{c}) \end{cases}; \tag{19}$$

see Tsui [21]. We can then make use of the central limit theorem to show that

$$\sqrt{n} \frac{\mathbf{c}^\top(\tilde{\mathbf{p}}_n - \mathbf{p})\mathbf{c}}{\sqrt{\mathbf{c}^\top W(\tilde{\mathbf{p}}_n)\mathbf{c}}} \rightsquigarrow N(0, 1). \tag{20}$$

Using this result, we can construct a confidence interval for  $\mathbf{c}^\top \mathbf{p}$ , with confidence level  $(1 - q)$ , given by

$$\text{pr}\left(\mathbf{c}^\top \tilde{\mathbf{p}}_n - z_q \sqrt{\frac{\mathbf{c}^\top W(\tilde{\mathbf{p}}_n)\mathbf{c}}{n}} \leq \mathbf{c}^\top \mathbf{p} \leq \mathbf{c}^\top \tilde{\mathbf{p}}_n + z_q \sqrt{\frac{\mathbf{c}^\top W(\tilde{\mathbf{p}}_n)\mathbf{c}}{n}}\right) \rightarrow 1 - q. \tag{21}$$

We may then use duality to test the hypothesis

$$H_0(\mathbf{c}) : \mathbf{c}^\top \mathbf{p} = \mathbf{c}^\top \mathbf{p}_0 \tag{22}$$

at the limit  $q$ . This method is useful for testing hypotheses about parameters in a variety of different models, including linear mixed models.

### 3.2. Pair of Samples

We denote  $M(|m(l), \mathbf{p}(l))$  as the singular multivariate distribution for a pair of independent samples,  $l = 1, 2$ , with  $\mathbf{p}(l) = (p_1(l), \dots, p_m(l))$  and  $p_i(l) \geq 0$ . Thus, distributions and the corresponding statistics, such as estimators, will be independent.

We have

$$\sqrt{n(l)}(\tilde{\mathbf{p}}(l) - \mathbf{p}(l)) \rightsquigarrow N(\mathbf{0}, W(\mathbf{p}(l))). \tag{23}$$

We can also take the difference between each of the  $p_i(1)$  and  $p_i(2)$ ,  $i = 1, \dots, m$ , and end up with unbiased estimators for  $q_i$ , as follows:

$$\tilde{q}_i = \tilde{p}_i(1) - \tilde{p}_i(2), \quad i = 1, \dots, m. \tag{24}$$

If  $\frac{n(2)}{n(1)} \rightarrow r$ , we will have

$$\begin{cases} \frac{n(1)}{n} \rightarrow \frac{1}{1+r} \\ \frac{n(2)}{n} \rightarrow \frac{r}{1+r} \end{cases}, \tag{25}$$

with  $n = n(1) + n(2)$ . So,

$$\begin{cases} \text{Var}(\sqrt{n}\tilde{p}_i(1)) \rightarrow (r+1)p_i(1)(1-p_i(1)) \\ \text{Var}(\sqrt{n}\tilde{p}_i(2)) \rightarrow \frac{r+1}{r}p_i(2)(1-p_i(2)) \end{cases}, \quad i = 1, \dots, m \tag{26}$$

and, given the independence of the two samples,

$$\text{Var}(\sqrt{n}\tilde{q}_i) \rightarrow (r+1) \left[ p_i(1)(1-p_i(1)) + \frac{1}{r}p_i(2)(1-p_i(2)) \right], \quad i = 1, \dots, m. \tag{27}$$

When  $r \rightarrow 0$ , the variance of  $\sqrt{n}\tilde{q}_i$  tends to zero, so it becomes an unbiased estimator. The same is true for  $r \rightarrow \infty$ . To compare the two multivariate distributions,  $M(|m(1), \mathbf{p}(1))$  and  $M(|m(2), \mathbf{p}(2))$ , we can use the Hotelling's  $T^2$  statistic to measure the distance between the two distributions:

$$T^2 = n(\tilde{\mathbf{p}}(1) - \tilde{\mathbf{p}}(2))^T W(\mathbf{p}(1))^{-1} (\tilde{\mathbf{p}}(1) - \tilde{\mathbf{p}}(2)). \tag{28}$$

It is well known that if

$$\frac{n(2)}{n(1)} \rightarrow r, \tag{29}$$

then

$$T^2 \rightsquigarrow \chi_{m-1}^2(r). \tag{30}$$

Now, reasoning as in the previous subsection, it may be shown that

$$\frac{\sqrt{n}(\tilde{q}_i - q_i)}{\sqrt{(r+1)\tilde{p}_i(1)(1-\tilde{p}_i(1)) + \frac{1}{r}\tilde{p}_i(2)(1-\tilde{p}_i(2))}} \rightsquigarrow N(|0, 1), \quad i = 1, \dots, m, \tag{31}$$

and so

$$\begin{aligned} & pr \left[ \tilde{q}_i - z_q \sqrt{(r+1) \left( \tilde{p}_i(1)(1-\tilde{p}_i(1)) + \frac{1}{r}\tilde{p}_i(2)(1-\tilde{p}_i(2)) \right)} \right] \leq q_i \tag{32} \\ & \leq \tilde{q}_i + z_q \sqrt{(r+1) \left( \tilde{p}_i(1)(1-\tilde{p}_i(1)) + \frac{1}{r}\tilde{p}_i(2)(1-\tilde{p}_i(2)) \right)} \leq 1 - q, \quad i = 1, \dots, m. \end{aligned}$$

Thus, through duality, we get a  $1 - q$  limit level test for

$$H_0(q_{0,i}) : q_i = q_{0,i}, \quad i = 1, \dots, m. \tag{33}$$

In addition,

$$\begin{cases} \text{Var}(\sqrt{nc}^T \tilde{\mathbf{p}}(1)) \rightarrow (r+1)\mathbf{c}^T W(\mathbf{p}(1))\mathbf{c} \\ \text{Var}(\sqrt{nc}^T \tilde{\mathbf{p}}(2)) \rightarrow \frac{r+1}{r}\mathbf{c}^T W(\mathbf{p}(2))\mathbf{c} \end{cases}, \tag{34}$$

as  $n \rightarrow \infty$ , so that

$$\text{Var}\left(\sqrt{n}(\mathbf{c}^\top \tilde{\mathbf{p}}(1) - \mathbf{c}^\top \tilde{\mathbf{p}}(2))\right) \xrightarrow{n \rightarrow \infty} (r+1)\mathbf{c}^\top \left[ W(\mathbf{p}(1)) + \frac{1}{r}W(\mathbf{p}(2)) \right] \mathbf{c}, \tag{35}$$

and that

$$\frac{\sqrt{n}[(\mathbf{c}^\top \tilde{\mathbf{p}}(1) - \mathbf{c}^\top \tilde{\mathbf{p}}(2)) - (\mathbf{c}^\top \mathbf{p}(1) - \mathbf{c}^\top \mathbf{p}(2))]}{\sqrt{(r+1)(\mathbf{c}^\top W(\tilde{\mathbf{p}}(1)) + \frac{1}{r}W(\tilde{\mathbf{p}}(2))\mathbf{c})}} = \tag{36}$$

$$\frac{\sqrt{n}(\mathbf{c}^\top \tilde{\mathbf{q}} - \mathbf{c}^\top \mathbf{q})}{\sqrt{(r+1)(\mathbf{c}^\top W(\tilde{\mathbf{p}}(1)) + \frac{1}{r}W(\tilde{\mathbf{p}}(2))\mathbf{c})}} \rightsquigarrow N(|0, 1). \tag{37}$$

Thus,

$$pr\left(\mathbf{c}^\top \tilde{\mathbf{q}} - z_q \sqrt{(r+1)(\mathbf{c}^\top W(\tilde{\mathbf{p}}(1)) + \frac{1}{r}W(\tilde{\mathbf{p}}(2))\mathbf{c})} \leq \mathbf{c}^\top \mathbf{q} \leq \tag{38}$$

$$\mathbf{c}^\top \tilde{\mathbf{q}} + z_q \sqrt{(r+1)(\mathbf{c}^\top W(\tilde{\mathbf{p}}(1)) + \frac{1}{r}W(\tilde{\mathbf{p}}(2))\mathbf{c})}\right) \rightarrow 1 - q. \tag{39}$$

And, using duality once again, we can obtain limit level  $q$  tests for

$$H_0(\mathbf{c}^\top \mathbf{q}) : \mathbf{c}^\top \mathbf{q} = d_l, \quad l = 1, 2. \tag{40}$$

### 3.3. Structured Families of Samples

Let us consider  $d$  treatments of fixed-effects designs, with a total of

$$n = \sum_{l=1}^d n(l) \tag{41}$$

individuals, and  $d$  independent samples. This configuration is referred to as a structured family, and the distributions of the samples,

$$M(|n(l), \mathbf{p}(l)), \quad l = 1, \dots, d, \tag{42}$$

with  $\mathbf{p}(l) = (p_1(l), \dots, p_m(l))$ ,  $l = 1, \dots, d$ , and  $p_i(l) \geq 0$ ,  $i = 1, \dots, m$ ,  $l = 1, \dots, d$ , characterize it. Moreover, consider

$$\frac{n(l)}{n} \rightarrow r(l), \quad l = 1, \dots, d. \tag{43}$$

Now, we will have

$$\dot{\mathbf{X}}_n(l) = \frac{n}{r(l)} (\tilde{\mathbf{p}}_{n(l)}(l) - \mathbf{p}(l))^\top W^+(\mathbf{p}_l) (\tilde{\mathbf{p}}_{n(l)}(l) - \mathbf{p}(l)) \rightarrow G(|m - 1), \quad l = 1, \dots, d, \tag{44}$$

where  $G(|m - 1)$  is the gamma distribution with a shape parameter of  $m - 1$  and a scale parameter of 1. So, with  $\dot{\mathbf{X}}_n = (\dot{\mathbf{X}}_n(1), \dots, \dot{\mathbf{X}}_n(d))$ , given the independence of the  $\dot{\mathbf{X}}_n(l)$ ,  $l = 1, \dots, d$ , we have

$$\dot{\mathbf{X}}_n \rightsquigarrow G(|m - 1), \tag{45}$$

with

$$\dot{G}(\mathbf{x}|m - 1) = \prod_{l=1}^d G(x_l|m - 1), \tag{46}$$

whenever  $\mathbf{x} = (x_1, \dots, x_d)$ .

We now study the action of the factors in the base design, assuming that the hypotheses of absence of effects and interactions are linked to an orthogonal partition,

$$\mathcal{R}^d = \boxplus_{j=1}^m \nabla_j, \tag{47}$$

and that the row vectors of  $A_j$  constitute an orthonormal basis for  $\nabla_j, j = 1, \dots, m$ . This leads to the hypotheses

$$H_{0,j}(\mathbf{c}) : A_j \eta(\mathbf{c}) = \mathbf{0}, \quad j = 1, \dots, m, \tag{48}$$

where  $\eta(\mathbf{c}) = (\mathbf{c}^\top \mathbf{p}(1), \dots, \mathbf{c}^\top \mathbf{p}(d))$ .

More generally,

$$H_{0,j}(\mathbf{c}) : \Psi_j(\mathbf{c}) = \Psi_{0,j}(\mathbf{c}), \quad j = 1, \dots, m. \tag{49}$$

We have

$$\Psi_j(\mathbf{c}) = A_j \eta(\mathbf{c}). \tag{50}$$

So, we will have

$$\tilde{\eta}(\mathbf{c}) = (\mathbf{c}^\top \mathbf{p}(1), \dots, \mathbf{c}^\top \mathbf{p}(d)) \tilde{\Psi}_j(\mathbf{c}) = A_j \tilde{\eta}(\mathbf{c}) \curvearrowright N(|\Psi_j(\mathbf{c})|, A_j D(\mathbf{c}) A_j^\top), \tag{51}$$

where  $D(\mathbf{c})$  is the diagonal matrix, with

$$\mathbf{c}^\top W(\mathbf{p}(1))\mathbf{c}, \dots, \mathbf{c}^\top W(\mathbf{p}(d))\mathbf{c} \tag{52}$$

along its diagonal, and the estimator is given by

$$\tilde{\Psi}_j^\top(\mathbf{c}) A_j D(\mathbf{c}) (A_j^\top)^+ \tilde{\Psi}_j(\mathbf{c}) \curvearrowright G(|g_j|), \quad j = 1, \dots, m. \tag{53}$$

Moreover,  $x_{g_j, 1-q}$  will be the critical value for a limit level  $q$  test for

$$H_{0,j}(\mathbf{c}), \quad j = 1, \dots, m. \tag{54}$$

#### 4. Optimal Allocation Rules

##### 4.1. The Discrete Case

Consider a set of populations combined together, each placed into different groups. A randomly selected sample of  $n$  elements from this mixture belongs to a specified population and class, denoted as  $n_{i,j}$  for elements in class  $S_i$  and population  $P_j$ , with  $i = 1$  to  $w$  and  $j = 1$  to  $m$ .

Our objective is to determine an allocation rule that minimizes the overall cost of assigning elements to their respective populations. The cost of assigning an element from  $P_j$  to  $P_{j'}$  is represented by  $c_{j,j'}$ .

Let  $E_{i,j}$  be the event that a randomly chosen element belongs to  $S_i$  and  $P_j$ , with  $E_{i,\cdot}$  [ $E_{\cdot,j}$ ] representing the event that a randomly chosen element belongs to  $S_i$  [ $P_j$ ]. Define

$$\begin{cases} p_{i,j} = pr(E_{i,j}) \\ p_{i,\cdot} = pr(E_{i,\cdot}) \\ p_{\cdot,j} = pr(E_{\cdot,j}) \end{cases}, \tag{55}$$

for  $i = 1, \dots, w$  and  $j = 1, \dots, m$ . Additionally, introduce conditional probabilities

$$\begin{cases} p_{i|j} = pr(E_{i,\cdot} | E_{\cdot,j}) \\ p_{j|i} = pr(E_{\cdot,j} | E_{i,\cdot}) \end{cases}, \tag{56}$$

for  $i = 1, \dots, w$  and  $j = 1, \dots, m$ .

If the elements located in  $S_i$  are assigned to  $P_j, i = 1, \dots, w, j = 1, \dots, m$ , the average cost is given by

$$c_{i(j)} = \sum_{j'=1}^m p_{j',i} c_{j',j}, \quad i = 1, \dots, w, \quad j = 1, \dots, m. \tag{57}$$

This leads to consistent estimators

$$\begin{cases} \tilde{p}_{i,j} = \frac{n_{i,j}}{n} \\ \tilde{p}_{i,\cdot} = \sum_{j=1}^m \tilde{p}_{i,j} \\ \tilde{p}_{i|j} = \frac{\tilde{p}_{i,j}}{\tilde{p}_{i,\cdot}} \end{cases}, \tag{58}$$

as the sample size increases, and the estimated average cost

$$\tilde{c}_i(j) = \sum_{j'=1}^m \tilde{p}_{j',i} c_{j',j}, \quad i = 1, \dots, w, \quad j = 1, \dots, m. \tag{59}$$

An optimum allocation for  $x_i$  occurs if there exists a  $j(i)$  such that  $c_i(j(i)) < c_i(j)$  for  $j \neq j(i)$ , in which case

$$pr \left( \bigcap_{j \neq j(i)} (\tilde{c}_i(j(i)) < \tilde{c}_i(j)) \right) \xrightarrow{n \rightarrow \infty} 1, \quad i = 1, \dots, w. \tag{60}$$

Furthermore, if there is an optimum allocation for all  $x_i, i = 1, \dots, w$ ,

$$pr \left( \bigcap_{i=1}^w \bigcap_{j \neq j(i)} \tilde{c}_i(j(i)) < \tilde{c}_i(j) \right) \xrightarrow{n \rightarrow \infty} 1. \tag{61}$$

Discriminant analysis is consistent, indicating that the probability of the allocation rule being optimal increases as the initial sample size grows to infinity. This holds true for all  $\mathfrak{S}_i$ , with  $i$  ranging from 1 to  $w$ .

The global average cost of assigning elements of  $\mathfrak{S}_i, i = 1, \dots, w$ , to  $P_{j(i)}$  is given by

$$c. = \sum_{i=1}^w p_i c_i(j(i)), \tag{62}$$

with the consistent estimator

$$\tilde{c}. = \sum_{i=1}^w \tilde{p}_i c_i(j(i)), \tag{63}$$

and the limit level  $1 - p$  confidence interval

$$\tilde{c}. \pm z'_q \sqrt{\mathbf{d}^\top W(\tilde{\mathbf{p}}) \mathbf{d}}, \tag{64}$$

where  $z'_q$  is the critical value of a standard normal distribution at level  $1 - q$ , and  $\mathbf{d}$  is the vector with all components null except the one with index  $1 + (i - 1)j(i), i = 1, \dots, w$ , which is 1.

#### 4.2. The Continuous Case

For each event  $\mathcal{A}_i, i = 1, \dots, k$ , that occurs when an element of  $\mathcal{P}$  is randomly chosen to belong to  $\mathcal{P}_i, i = 1, \dots, k$ , the probability of the event is  $p_i = pr(\mathcal{A}_i)$ . To determine the allocation, a partition of the observations vector  $\mathbf{X} \in \mathbb{R}^n$  is created such that each element is assigned to  $\mathcal{P}_i$  when  $\mathbf{X} \in \Omega_i, i = 1, \dots, k$ . The probability of assigning to  $\mathcal{P}_j$  an element of  $\mathcal{P}_i$  is denoted as  $q_{i,j}, i = 1, \dots, k$  and  $j = 1, \dots, k$ , and is calculated from the densities of  $\mathbf{X}$  in the  $\mathcal{P}_i, i = 1, \dots, k$ . The expected cost of the decision is then determined by summing the product of the costs of each pair of events and their associated probabilities.

The average cost incurred by the decision will be equal to the sum of the probabilities multiplied by the expected cost for each action multiplied by the probability of that action. This can be expressed as the sum of integrals for each possible action over the region where that action is the optimal choice. That is,

$$E(C) = \sum_{j=1}^k \int \dots \int_{\Omega_j} g_j(\mathbf{x}) \prod_{l=1}^n dx_l, \tag{65}$$

with

$$g_j(\mathbf{x}) = \sum_{i=1}^k p_i c_{i,j} f_i(\mathbf{x}); \quad j = 1, \dots, k. \tag{66}$$

To minimize the expected cost, one must take the regions where each action is the optimal choice,

$$\Omega_j = \{\mathbf{x} : g_j(\mathbf{x}) \leq g_i(\mathbf{x}), \quad i = 1, \dots, k\}; \quad j = 1, \dots, k. \tag{67}$$

In the situation of identical decision costs, for which  $c_{i,j} = c, i \neq j$ , this is especially noteworthy. This means that

$$g_j(\mathbf{x}) = c \sum_{i \neq j} p_i f_i(\mathbf{x}) = c \sum_{i=1}^k p_i f_i(\mathbf{x}) - c p_j f_j(\mathbf{x}); \quad j = 1, \dots, k, \tag{68}$$

and the sets

$$\Omega_j = \{\mathbf{x} : p_j f_j(\mathbf{x}) \geq p_i f_i(\mathbf{x}), \quad i = 1, \dots, k\}; \quad j = 1, \dots, k. \tag{69}$$

are applicable. In addition, if  $p_i = \frac{1}{k}, i = 1, \dots, k$ , then

$$\Omega_j = \{\mathbf{x} : f_j(\mathbf{x}) \geq f_i(\mathbf{x}), \quad i = 1, \dots, k\}; \quad j = 1, \dots, k. \tag{70}$$

This means that, given  $\mathbf{x}$ , the population with the greatest probability will be chosen.

In summary, the expected cost associated with the decision can be expressed as a sum of integrals over the regions where each action is the optimal choice. This decision can be further simplified when the costs for each action are equal and the probabilities are the same. In this situation, the decision is simply to choose the population with the greatest probability.

### 5. Numerical Applications

#### 5.1. The Discrete Case—An Application to Real Data

The purpose of this subsection is to illustrate the theory, relative to the discrete case. To do this, we use an application with real data, obtained from Instituto Superior Dom Bosco [22], related to the impact of four economic policies on greenhouse gas emissions: Policy A, Policy B, Policy C, and Policy D. The data were provided by the Instituto Superior Dom Bosco, and will be used to highlight the effectiveness of the discrimination rule in guiding resource allocation decisions.

Let us assume that for each region, the four policies have four distinct categories. This would create  $4^4 = 256$  possible combinations. If our null hypothesis is true and all of the combinations are equally likely, then any given combination should have a probability of  $\frac{1}{256}$ .

We present a discrimination rule, which will classify regions into two groups, “High Emission” (Population 1) and “Low Emission” (Population 2), based on the values of the statistic:

$$v = |x_1 - x_2| + |x_1 - x_3| + |x_1 - x_4| + |x_2 - x_3| + |x_2 - x_4| + |x_3 - x_4| + 1, \tag{71}$$

where each policy category has a value of 1, 2, 3, or 4.

The choice of this statistic is based on its ability to capture the differences between the four economic policies in terms of their impact on greenhouse gas emissions. By calculating the absolute differences between the categories of each policy and summing them, we obtain a single value that represents the overall dissimilarity between the policies. In addition, the use of the absolute differences ensures that the magnitude of the differences is taken into account, regardless of their direction. This is important in the context of

evaluating policy effectiveness, as both positive and negative differences can contribute to the overall impact.

The collection of potential values for  $v$  is  $D = \{1, 4, 5, 7, 8, 9, 10, 12, 13\}$ . We use the values of  $r_h = \frac{p_{h,1}}{p_{h,2}}, h \in D$  of the conditional probabilities and allocate to population 2 or 1 those with  $v \in D(x)$  or  $v \notin D(x)$ , respectively.

Suppose  $n_{h,l}$  denotes the number of elements with  $v = h$  in population  $l$ . We can obtain the sample size for each population as follows:

$$\sum_{h \in D} n_{h,l}, \quad l = 1, 2 \tag{72}$$

and the probability of obtaining  $v = h$  in a randomly chosen element from that sample as

$$\frac{n_{h,l}}{\sum_{h \in D} n_{h,l}}, \quad h \in D, \quad l = 1, 2. \tag{73}$$

The overall sample size will be

$$\sum_{l=1}^2 \sum_{h \in D} n_{h,l}, \tag{74}$$

while the number of its elements, with  $v = h$ , will be

$$\sum_{l=1}^2 n_{h,l}, \quad h \in D. \tag{75}$$

The probability of a randomly chosen element of the global sample having  $v = h$  [belonging to population  $l$ ] will be

$$\frac{\sum_{l=1}^2 n_{h,l}}{\sum_{l=1}^2 \sum_{h \in D} n_{h,l}} \tag{76}$$

and

$$\frac{\sum_{h \in D} n_{h,l}}{\sum_{l=1}^2 \sum_{h \in D} n_{h,l}}. \tag{77}$$

We begin by finding the ratios  $r_h = \frac{p_{h,1}}{p_{h,2}}, h \in D$ , of the conditional probabilities, and set  $D(x) = \{h; r_h \leq x\}$ , allocating to population 2 [population 1] those with  $v \in D(x)$  [ $v \notin D(x)$ ], therefore to those with  $r_v \leq x$  [ $r_v > x$ ].

Table 1 displays the results, which indicate that  $x = 1$  should be chosen and  $D(1) = \{1\}$ , thus giving to population 1, the elements with  $v = 1$ , and to population 2, the elements with  $v > 1$ .

**Table 1.** Numbers, probabilities and ratios of cases for the values of  $v$  from the different population.

$h$	1	4	5	7	8	9	10	11	12	13	Total
$n_{h,1}$	17,456	738	755	939	1369	254	380	997	670	40	$n_{\cdot,1} = 23,598$
$n_{h,2}$	38,193	542	156	28	88	6	24	22	17	2	$n_{\cdot,2} = 39,078$
$n_{h\cdot}$	55,649	1280	911	967	1457	260	404	1019	687	42	$n_{\cdot\cdot} = 62,676$
$p_{h,1}$	0.740	0.031	0.032	0.040	0.058	0.011	0.016	0.042	0.028	0.002	1
$p_{h,2}$	0.977	0.014	0.004	0.001	0.002	0.000	0.001	0.001	0.000	0.000	1
$\Pi_{h\cdot}$	0.888	0.020	0.015	0.015	0.023	0.004	0.006	0.016	0.011	0.001	1
$r_h$	0.757	2.255	8.015	55.535	25.762	70.103	26.220	75.046	65.265	33.120	

By computing

$$\begin{cases} q_{1,l} = \sum_{v \in D(1)} p_{v,l} & ; l = 1, 2 \\ q_{2,l} = \sum_{v \notin D(1)} p_{v,l} & ; l = 1, 2 \end{cases} \quad (78)$$

the outcomes are summarized in Table 2.

**Table 2.** Contingency table of probabilities.

	Population 1	Population 2
$q_{1,l}$	0.740	0.977
$q_{2,l}$	0.260	0.023

The cost of wrong decisions for both populations was calculated to be  $\bar{c}_1 = 0.73972$  and  $\bar{c}_2 = 0.02265$ . The global average cost of wrong decisions can be estimated by the sum of the product of the probability of randomly chosen elements for the mixture,  $\Pi_{.,1}$  and  $\Pi_{.,2}$ , and the respective population average costs, which is

$$\bar{c} = 0.73972\Pi_{.,1} + 0.02265\Pi_{.,2}. \quad (79)$$

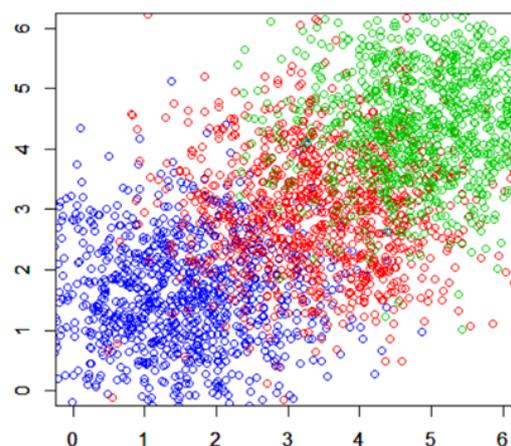
### 5.2. The Continuous Case—An Application to Simulated Data

In this subsection, we used R software, version 4.3.0, to simulate and compute a mixture of three populations with mean vectors

$$\begin{cases} \mu_1 = (1.5; 1.5) \\ \mu_2 = (3.5; 3) \\ \mu_3 = (5; 4.5) \end{cases} \quad (80)$$

and identical covariance matrices  $I_2$ . All measures were taken based on the corresponding normal homocedastic distributions. These values were assumed based on a reasonable assumption for the data generation process, taking into account the need for clear separation between the populations for effective simulation and analysis. In a real-world context, such as in the field of botany, these could, for example, represent averages of certain measurable traits, such as the lengths of petals from three different types of plants.

Simulated samples, with dimension 1000 were generated. These samples are presented in Figure 1.



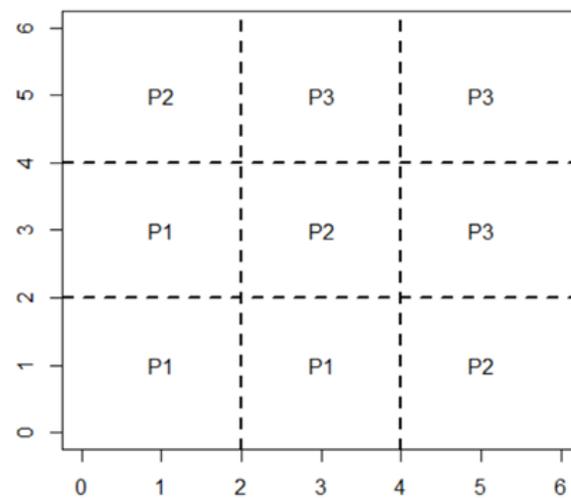
**Figure 1.** Simulated samples, with blue, red, and green points, corresponding to populations with mean  $\mu_1$ ,  $\mu_2$ , and  $\mu_3$ , respectively.

Table 3 shows the costs of misallocation. The rows represent the real population of the elements and the columns the populations they were allocated to.

**Table 3.** Costs of misallocation

	$P_1$	$P_2$	$P_3$
$P_1$	0	1	2
$P_2$	1	0	1
$P_3$	2	1	0

The shapes for the areas of distribution for the elements were established as illustrated in Figure 2. These are appropriate for normally distributed data with identical covariance matrices. The selected rectangular regions aligned with the axes allowed for efficient calculation of the probabilities of allocation and cost of misallocation.



**Figure 2.** Shape of the zones to be considered.

The average cost will depend on the amount of elements that are allocated to  $P_1$  from  $P_2$  and to  $P_2$  from  $P_3$ , as the probabilities of allocating to  $P_3$  one element of  $P_1$  and to  $P_1$  one element of  $P_3$  are very small.

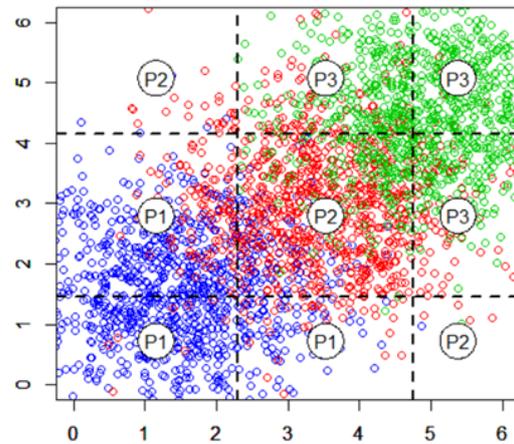
With  $C_1, C_2$  and  $C_3$  representing the regions in which the elements are allocated to populations  $P_1, P_2$  and  $P_3$ , respectively, we must minimize

$$q = \int_{C_1} dF_2 + \int_{C_2} dF_1 + \int_{C_3} dF_2 + \int_{C_2} dF_3.$$

As shown in Figure 2, our approach was to optimize the points where the horizontal and vertical lines intersect the x-axis at  $X = 2$  and  $X = 4$  and the y-axis at  $Y = 2$  and  $Y = 4$ , respectively. These intersection points act as delineations for the designated regions. The selection of those initial points was based on an informed estimation leveraging the mean vectors of the three populations. Noting the fact that the mean vectors for the three populations are  $\mu_1 = (1.5; 1.5)$ ,  $\mu_2 = (3.5; 3)$ , and  $\mu_3 = (5; 4.5)$ , the initial points were chosen to lie approximately midway between these centroids. Taking into account that the populations are distributed normally with identical covariance matrices, this ensures that the initial delineations for the partitioning areas ( $C_1, C_2, C_3$ ) in the clustering process are not skewed towards one or more populations. Once these initial parameters were set, the optimization process was run, using the ‘stats’ package of R software, version 4.3.0. in R, particularly the ‘optim’ function, continually adjusting the parameters in order to find the values that minimized the objective function, which, in this case, was the cost of misallocation.

After 1000 repetitions of this optimization process, the suitable partitions displayed in Figure 3 were obtained. These partitions represent the areas of allocation for the elements

within the three populations that produced the minimum misallocation cost, and, therefore, the most efficient allocation of elements within the identified clusters.



**Figure 3.** Considered cuts. The colors of the points, blue, red, and green, represent populations with mean values of  $\mu_1$ ,  $\mu_2$ , and  $\mu_3$ , respectively.

The cuts that are most cost-efficient can be represented by the straight lines

$$\begin{cases} x = 2.3; & x = 4.75 \\ y = 1.45; & y = 4.15 \end{cases} \quad (81)$$

Table 4 shows the number of elements in each area.

**Table 4.** Number of elements in each area.

$P_2$	$P_3$	$P_3$
$P_1 = 5$	$P_1 = 0$	$P_1 = 0$
$P_2 = 26$	$P_2 = 102$	$P_2 = 6$
$P_3 = 1$	$P_3 = 261$	$P_3 = 364$
$P_1$	$P_2$	$P_3$
$P_1 = 398$	$P_1 = 125$	$P_1 = 0$
$P_2 = 106$	$P_2 = 626$	$P_2 = 82$
$P_3 = 0$	$P_3 = 159$	$P_3 = 213$
$P_1$	$P_1$	$P_2$
$P_1 = 381$	$P_1 = 90$	$P_1 = 1$
$P_2 = 8$	$P_2 = 40$	$P_2 = 4$
$P_3 = 0$	$P_3 = 1$	$P_3 = 1$

The corresponding obtained total minimum cost was 638.

So, using simulated samples and different partitions of the sample, we were able to obtain the most cost-efficient cuts and the corresponding total minimum cost.

## 6. Conclusions

The paper discusses a comprehensive method to obtain optimal allocation rules for fixed-effect experiments. The study process is divided into four sections, namely, asymptotic distributions, confidence ellipsoids, inference, and optimal allocation rules.

Under the section of asymptotic distributions, the Central Limit Theorem (CLT) is accredited for providing valuable insights in terms of the behavior of sample averages.

In confidence ellipsoids, the ellipsoid properties are studied in developing a comprehensive understanding of the distribution of observations. The convex shape of the ellipsoid allows every point within the ellipsoid to be a convex combination of points on its boundary. Simultaneous confidence intervals can be constructed using these ellipsoids.

In the section on inference, the probability functions are detailed. These functions quantify the likelihood of obtaining specific results in experiments. As the number of trials increases, the asymptotic behavior of estimators is established. The distribution of estimated probabilities becomes approximately normal as the sample size grows.

The section on optimal allocation rules further elaborates on the methods for determining an allocation rule that minimizes the overall cost of assigning elements to their respective populations.

The paper concludes with two numerical applications. The first is a real data application showcasing the employment of computational techniques throughout the paper to analyze the real data. The second application undertakes the simulation of data to validate the findings of the study.

**Author Contributions:** Conceptualization, D.F.; Methodology, D.F. and S.S.F.; Investigation, S.S.F.; Writing—review & editing, D.F. and S.S.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partially supported by the Portuguese Foundation for Science and Technology through the projects UIDB/00212/2020, UIDB/04630/2020 and UIDB/00297/2020.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Conflicts of Interest:** The authors declare that they have no conflict of interest.

## References

- Hand, D.; Christen, P. A note on using the F-measure for evaluating record linkage algorithms. *Stat. Comput.* **2018**, *28*, 539–547. [CrossRef]
- Birch, M.W. Maximum likelihood in three-way contingency tables. *J. R. Stat. Soc. Ser. Methodol.* **1963**, *25*, 220–233. [CrossRef]
- Fienberg, S.E.; Rinaldo, A. Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *J. Stat. Plan. Inference* **2007**, *137*, 3430–3445. [CrossRef]
- Goodman, L.A. Interactions in multidimensional contingency tables. *Ann. Math. Stat.* **1964**, *35*, 632–646. [CrossRef]
- Haberman, S.J. Log-Linear Models for Frequency Tables with Ordered Classifications. *Biometrics* **1974**, *30*, 589–600. [CrossRef]
- Fisher, R.A. The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]
- Rao, R.C. The utilization of multiple measurements in problems of biological classification. *J. R. Stat. Soc. Ser.* **1948**, *10*, 159–203. [CrossRef]
- Friedman, J.H. Regularized Discriminant Analysis. *J. Am. Stat. Assoc.* **1989**, *84*, 165–175. [CrossRef]
- McFarland, H.R.; Richards, D.S.P. Exact Misclassification Probabilities for Plug-In Normal Quadratic Discriminant Functions: I. The Equal-Means Case. *J. Multivar. Anal.* **2001**, *77*, 21–53. [CrossRef]
- Perriere, G.; Thioulouse, J. Use of Correspondence Discriminant Analysis to predict the subcellular location of bacterial proteins. *Comput. Methods Programs Biomed.* **2003**, *70*, 99–105. [CrossRef] [PubMed]
- Flury, L.; Bourkai, B.; Flury, B.D. The discrimination subspace model. *J. Am. Stat. Assoc.* **1997**, *92*, 758–766. [CrossRef]
- Li, H.; Jia, M.; Mao, Z. Dynamic Feature Extraction-Based Quadratic Discriminant Analysis for Industrial Process Fault Classification and Diagnosis. *Entropy* **2023**, *25*, 1664. [CrossRef] [PubMed]
- Tong, Z.; Li, W.; Zhang, B.; Gao, H.; Zhu, X.; Zio, E. Bearing Fault Diagnosis Based on Discriminant Analysis Using Multi-View Learning. *Mathematics* **2022**, *10*, 3889. [CrossRef]
- McLachlan, G.J. *Discriminant Analysis and Statistical Pattern Recognition*; Wiley Interscience: New York, NY, USA, 2004; ISBN 978-0-471-69115-0.
- Ferreira, S.S.; Ferreira, D.; Nunes, C.; Mexia, J.T. Discriminant Analysis and Decision Theory. *Far East J. Math. Sci.* **2011**, *51*, 69–79.
- Schott, J.R. *Matrix Analysis for Statistics*, 3rd ed.; Wiley: New York, NY, USA, 2017.
- Kallenberg, O. *Foundations of Modern Probability*, 2nd ed.; Springer: New York, NY, USA, 2010.
- Scheffé, H. *The Analysis of Variance*; John Wiley & Sons: New York, NY, USA, 1959.
- Wilks, S.S. *Mathematical Statistics*; Wiley: New York, NY, USA, 1962.
- Horn, R.A.; Johnson, C.R. *Matrix Analysis*; Cambridge University Press: Cambridge, UK, 1985.
- Tsui, K.W. Asymptotic theory of statistical inference for random coefficients regression models. *J. Multivar. Anal.* **1995**, *54*, 281–299.
- Instituto Superior Dom Bosco—Maputo Mozambique. Available online: <https://isdb.co.mz/> (accessed on 7 August 2018).

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.