Smart Statistics for Smart Applications

Book of Short Papers SIS2019

SIS 2019
SMART STATISTICS FOR SMART APPLICATIONS

SIS Società Italiana di Statistica

Editors: Giuseppe Arbia, Stefano Peluso,
Alessia Pini and Giulia Rivellini

Published by Pearson

www.pearson.com

Giugno 2019 ISBN 9788891915108

Preface

Section 1. Plenary Sessions and Round Table

Section 2. Invited Papers

Section 3. Contributed Papers

Section 4. Posters

Design of Experiments, aberration and Market Basket Analysis

Pianificazione degli esperimenti, aberrazione e Market Basket Analysis

Roberto Fontana and Fabio Rapallo

Abstract In Design of Experiments, the Minimum Aberration criterion uses the aberrations of all the main effects and interactions to choose *suitable* fractions. In this work, we consider the aberrations in Market Basket Analysis, which is a completely different context from Design of Experiments. Using a real dataset, we show how aberrations could represent a meaningful measure of association.

Abstract *Nella pianificazione degli esperimenti, il criterio della Minima Aberrazione utilizza le aberrazioni di tutti gli effetti prinicipali e di interazione per selezionare* opportuni *piani fattoriali frazionari. In questo contributo consideriamo le aberrazioni nel contesto – totalmente diverso – della Market Basket Analysis. Basandoci su un esempio applicativo, mostriamo che le aberrazioni rappresentano una buona misura di associazione.*

Key words: Fractional factorial designs, Word-Length Pattern, association

1 Introduction

Fractional factorial designs are commonly used in Design of Experiments to determine the optimum mix of factors to predict a response variable. The need for efficient experimental designs has led to the definition of several criteria for the choice of the design points. In the case of binary designs, an important object associated to a design is its Word-Length Pattern (WLP). The WLP is used to discriminate among different designs through the Minimum Aberration (MA) criterion, which is based on the sequential minimization of the WLP. The MA criterion was introduced in

Roberto Fontana
DISMA, Dipartimento di Eccellenza 2018-2022, Politecnico di Torino, e-mail: roberto.fontana@polito.it

Rapallo Fabio
DISIT, Università del Piemonte Orientale e-mail: fabio.rapallo@uniupo.it

[6] for binary designs and then extended with the name of Generalized Minimum Aberration to non-regular multilevel designs in [7]. The WLP is computed using the aberrations of all the main effects and interactions.

In this work we use aberrations in the context of Market Basket Analysis (MBA). MBA finds association rules between sets of products bought in a unit of shopping by consumers (see e.g. [1]). In our application, given m products, a basket will be a binary vector $x = (x_1, \ldots, x_m)$, where $x_i = 1$ (resp. $x_i = -1$) means that i-th product has (resp. has not) been bought by the customer. The set of all the baskets is a multiset of $\{-1,1\}^m$ and can be considered as a fractional factorial design of $\{-1,1\}^m$. We use aberrations to describe the associations between products. To the best of our knowledge the use of aberrations as measures of association is new. An application to a set of Italian museums is shown.

2 Fractional factorial designs and aberrations

Let us consider an experiment with m binary factors. The full factorial design is $\mathscr{D} = \{-1,1\}^m$. We briefly recall here the basic definitions concerning fractional factorial designs (or simply fractions) and aberrations. For details refer to [4].

Definition 1. A fraction $\mathscr{F}$ is a multiset $(\mathscr{F}_*, f_*)$ whose underlying set of elements $\mathscr{F}_*$ is contained in $\mathscr{D}$ and f_* is the multiplicity function $f_* : \mathscr{F}_* \to \mathbb{N}$ that for each element in $\mathscr{F}_*$ gives the number of times it belongs to the multiset $\mathscr{F}$.

We recall that the underlying set of elements $\mathscr{F}_*$ is the subset of $\mathscr{D}$ that contains all the elements of $\mathscr{D}$ that appear in $\mathscr{F}$ at least once. We denote the number of elements of the fraction $\mathscr{F}$ by $\#\mathscr{F}$, with $\#\mathscr{F} = \sum_{x \in \mathscr{F}_*} f_*(x)$.

To describe the counting function of a fraction, we follow the theory in [4]. The simple terms of the form X_j, i.e., the j-th component function which maps a point $x = (x_1, \ldots, x_m)$ of $\mathscr{D}$ to its j-th component,

$$X_j \colon \mathscr{D} \ni (x_1, \ldots, x_m) \longmapsto x_j \in \{-1,1\}$$

and the interactions $X^\alpha = X_1^{\alpha_1} \cdot \ldots \cdot X_m^{\alpha_m}$, $\alpha \in L = \{0,1\}^m$ i.e., the monomial functions of the form

$$X^\alpha : \mathscr{D} \ni (x_1, \ldots, x_m) \mapsto x_1^{\alpha_1} \cdot \ldots \cdot x_m^{\alpha_m}$$

are a basis of all the real functions defined over $\mathscr{D}$. We use this basis to represent the counting function of a fraction according to the following definition.

Definition 2. The counting function R of a fraction $\mathscr{F}$ is a polynomial defined over $\mathscr{D}$ so that for each $x \in \mathscr{D}$, $R(x)$ equals the number of appearances of x in the fraction. We denote by c_α the coefficients of the representation of R on $\mathscr{D}$ using the monomial basis $\{X^\alpha,\ \alpha \in L\}$:

$$R(x) = \sum_{\alpha \in L} c_\alpha X^\alpha(x),\ x \in \{-1,1\}^m,\ c_\alpha \in \mathbb{R}\,.$$

Definition 3. The Word-Length Pattern (WLP) of a fraction $\mathscr{F}$ of the full factorial design $\mathscr{D}$ is the vector $A_{\mathscr{F}} = (A_0(\mathscr{F}), A_1(\mathscr{F}), \ldots, A_m(\mathscr{F}))$, where

$$A_j(\mathscr{F}) = \sum_{|\alpha|_0 = j} a_\alpha \quad j = 0, \ldots, m,$$

$$a_\alpha = (c_\alpha / c_0)^2,$$

$|\alpha|_0$ is the number of non-null elements of α, and $c_0 := c_{(0,\ldots,0)} = \#\mathscr{F}/\#\mathscr{D}$.

In the definition above, the number a_α is the aberration of the term X^α. In the case of binary design, using some results in [5], Prop. 1 below yields the aberration a_α as a function of the counts of the corresponding term X^α.

Proposition 1. *Given a fraction $\mathscr{F}$ of $\mathscr{D} = \{-1,1\}^m$ the aberration a_α of the term X^α is*

$$a_\alpha = (n_{1,\alpha} - n_{-1,\alpha})^2 / \#\mathscr{F}^2, \tag{1}$$

where $n_{1,\alpha}$ (resp. $n_{-1,\alpha}$) is the number of times X^α is equal 1 (resp. -1) in $\mathscr{F}$,

$$n_{1,\alpha} = \sum_{x \in \mathscr{D}} R(x) \delta_{x^\alpha}^1, \qquad n_{-1,\alpha} = \sum_{x \in \mathscr{D}} R(x) \delta_{x^\alpha}^{-1},$$

and δ_a^b denotes the Kronecker delta.

It is worth noting that the terms a_α can be easily interpreted for main effects and 2-factor interactions, i.e. when $|\alpha|_0 \in \{1,2\}$. If $|\alpha|_0 = 1$, without loss of generality, let us consider $\alpha = (1,0,\ldots,0)$. We get:

$$n_{1,\alpha} \equiv n_{1,(1,0,\ldots,0)} = \sum_{x=(x_1,\ldots,x_m)\in\mathscr{D}: x_1=1} R(x) \tag{2}$$

and a similar formula holds for $n_{-1,\alpha} \equiv n_{-1,(1,0,\ldots,0)}$. It follows that a_α is the squared difference between the relative frequency of the points for which $x_1 = 1$ and the relative frequency of the points for which $x_1 = -1$.

If $|\alpha|_0 = 2$ let us consider $\alpha = (1,1,0,\ldots,0)$. We get:

$$n_{1,\alpha} \equiv n_{1,(1,1,0\ldots,0)} = \sum_{\substack{x=(x_1,\ldots,x_m)\in\mathscr{D}: \\ x_1 x_2 = 1}} R(x) = \sum_{\substack{x=(x_1,\ldots,x_m)\in\mathscr{D}: \\ x_1 = x_2 = 1}} R(x) + \sum_{\substack{x=(x_1,\ldots,x_m)\in\mathscr{D}: \\ x_1 = x_2 = -1}} R(x) \tag{3}$$

and a similar formula holds for $n_{-1,\alpha} \equiv n_{-1,(1,1,0\ldots,0)}$. In this case a_α is the squared difference between the relative frequency of the points for which there is *agreement*, i.e. $x_1 = x_2 = 1$ or $x_1 = x_2 = -1$ and the relative frequency of the points for which there is *disagreement*, i.e. $x_1 = -1, x_2 = 1$ or $x_1 = 1, x_2 = -1$.

To correctly interpret the aberrations of order two, we compute the approximate distributions of such parameters through a Monte Carlo resampling (based on $1,000$ replicates) for different values of the sample size, in order to check also the robustness of the method for small data sets. Moreover, we compare the approximate distribution of each aberration with the expected aberration under independence.

For instance, the expected value under independence of the aberration $a_{(1,1,0,\ldots,0)}$ is:

$$\widehat{a}_{(1,1,0,\ldots,0)} = \left(\hat{n}_{1,(1,1,0\ldots,0)} - \hat{n}_{-1,(1,1,\ldots,0)}\right)^2 / \#\mathscr{F}^2 =$$

$$= \left\{ n_{1,(1,0,0,\ldots,0)} \left(n_{1,(0,1,0,\ldots,0)} - n_{-1,(0,1,0,\ldots,0)}\right) \right.$$
$$\left. - n_{-1,(1,0,0,\ldots,0)} \left(n_{1,(0,1,0,\ldots,0)} - n_{-1,(0,1,0,\ldots,0)}\right) \right\}^2 / \#\mathscr{F}^4 .$$

By comparison with Eqs. (1) and (2), we get:

$$\widehat{a}_{(1,1,0,\ldots,0)} = a_{(1,0,0,\ldots,0)} a_{(0,1,0,\ldots,0)} . \tag{4}$$

3 An application to Italian museums

In this paper we study a subset of museums participating in the Abbonamento Musei Torino Piemonte (AMTP, https://piemonte.abbonamentomusei.it) network. This network was created in 1995 and is available to people living in the Piemonte region (Italy). For a yearly subscription fee, AMTP card-holders have free entry to all the museums and all the temporary/permanent exhibitions participating in the program for the subscription year, from January to December. In recent years the number of subscribers has increased enormously: from 5,734 cards in 1999 to 87,237 cards in 2012, to 127,768 in 2016.

We analyze the 2012 AMTP transaction database. A descriptive analysis of the association structure of the 2012 AMTP network was conducted in [2]. Graphical models have been studied in [3]. The 2012 AMTP database collects information for each card-holder about the museums that he/she visited in 2012.

We consider which museums have been visited by each card-holder, regardless of the number of times he/she returned to the same museum. In this way we exclude repeated visits and we have a total $287,259$ visits to the main 23 museums. The number of card-holders who visited at least one of the 23 museums is $73,668$. Therefore, the final dataset has $N = 73,668$ observations and is organized as follows. Each row corresponds to a card-holder. The first 23 columns are binary variables and indicate whether the card-holder visited a museum one or more times, or did not visit a museum. As explained above we do not distinguish between single and multiple visits to the same museum. We note that the number of observations ($N = 73,668$) is less than the number of card-holders in 2012 because some of them did not use the card or did not visit any of the 23 museums which have been chosen for this analysis. Given the limited amount of space for this paper we select five museums, build up a $\{-1,1\}^5$ design with an appropriate counting function containing the absolute frequencies of visits, and show the use of aberrations of order two as one of the statistics to describe associations between any two of them.

In our study a *basket* is the set of museums which have been visited at least one time in the same year by an individual card-holder.

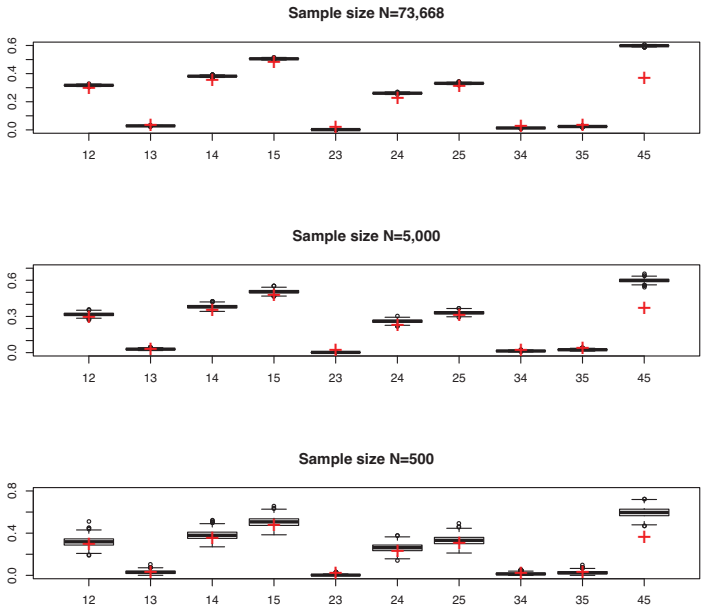

Fig. 1 Distribution of the aberrations of order two for different sample sizes. The symbol '+' denotes the expected aberration under independence.

In Fig. 1 the distributions (boxplots) of the aberrations of order two are compared with the expected aberration as in Eq. (4) for three sample sizes, namely the original sample size $N = 73,668$, $N = 5,000$ and $N = 500$. We note that the aberration 45 is significantly higher than expected, showing an agreement between the last two museums. We can also observe that the analysis remains robust when the sample size decreases.

We also compare the aberrations with the classical correlation coefficients. The results are displayed in Fig. 2. Here we note that there is a moderate correlation between the second and the third museum, not recognized by the aberration. This fact happens because the third museum has a large number of visitors compared with the other museums. Indeed, we can observe the same behavior for all the aberrations involving the third museum. In this example, the aberrations are less sensitive than correlations to non-homogeneous margins, and this may be an issue in favor of aberrations, especially when some museums has few visitors.

Future research will focus on the use of aberrations of order greater than two to measure the associations of more than two variables. The comparison of aberrations with other common measures of association (e.g. odds ratios and lifts) will also be studied.

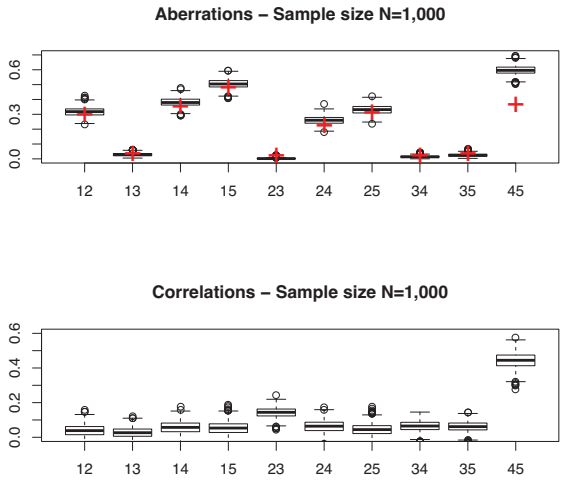

Fig. 2 Distribution of the aberrations of order two and the corresponding correlation coefficients.

Acknowledgements R. Fontana acknowledges that the present research has been partially supported by MIUR grant Dipartimenti di Eccellenza 2018-2022 (E11G18000350001). The authors thank Francesca Leon, past president of the Associazione Torino Città Capitale and currently member of the Turin city council, for providing the database of 2012 AMTP card-holders. R. Fontana also thanks Prof. Patrizia Semeraro (Politecnico di Torino) for helpful discussions.

References

1. Agresti, A.: Categorical data analysis, 3rd edn. John Wiley & Sons (2013)
2. Coscia, C., Fontana, R., Semeraro, P.: Market basket analysis for studying cultural consumer beahviour: AMTP card-holders. Statistica Applicata - Italian Journal of Applied Statistics **26**(2), 73–92 (2016)
3. Coscia, C., Fontana, R., Semeraro, P.: Graphical models for complex networks: an application to italian museums. J. Appl. Stat. **45**(11), 2020–2038 (2018)
4. Fontana, R., Pistone, G., Rogantin, M.P.: Classification of two-level factorial fractions. J. Statist. Plann. Inference **87**(1), 149–172 (2000)
5. Fontana, R., Rapallo, F., Rogantin, M.P.: Aberration in qualitative multilevel designs. J. Statist. Plann. Inference **174**, 1–10 (2016)
6. Fries, A., Hunter, W.G.: Minimum aberration 2^{k-p} designs. Technometrics **22**(4), 601–608 (1980)
7. Xu, H., Wu, C.F.J.: Generalized minimum aberration for asymmetrical fractional factorial designs. Ann. Statist. **29**(4), 1066–1077 (2001)