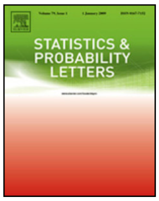
Statistics and Probability Letters 112 (2016) 72–78

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro



On randomization-based and regression-based inferences for 2^K factorial designs

Jiannan Lu

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA

ARTICLE INFO

Article history:
Received 11 December 2015
Received in revised form 5 January 2016
Accepted 15 January 2016
Available online 8 February 2016

Keywords:
Causal inference
Potential outcome
Unbalanced design
Huber–White estimator

ABSTRACT

We extend the randomization-based causal inference framework in Dasgupta et al. (2015) for general 2^K factorial designs, and demonstrate the equivalence between regression-based and randomization-based inferences. Consequently, we justify the use of regression-based methods in 2^K factorial designs from a finite-population perspective.

1. Introduction

Factorial designs, originally introduced for agricultural experiments (Fisher, 1935; Yates, 1937), have gained more popularity in recent times because of their abilities to investigate multiple treatment factors simultaneously. As pointed out by Ding (2014), although rooted in randomization theory (e.g., Kempthrone, 1952), factorial designs have been dominantly analyzed by regression methods in practice. Unfortunately, however, regression-based inference might not be suitable under certain circumstances. For example, several researchers (e.g., Miller, 2006; Lu et al., 2015) have pointed out that in many randomized experiments we cannot treat the experimental units as a random sample drawn from a hypothetical super-population, and should instead restrict the scope of inference to the finite-population of the experimental units themselves. Realizing the inherent deficiencies of regression-based inference, Dasgupta et al. (2015) advocated conducting randomization-based inference for factorial designs by utilizing the concept of potential outcomes (Neyman, 1990; Rubin, 1974). The proposed framework for balanced 2^K factorial designs is flexible, interpretable and applicable to both finite-population and super-population settings.

Given the advantages of randomization-based inference, it is necessary to generalize the framework in Dasgupta et al. (2015) for more general, i.e., unbalanced, 2^K factorial designs. Moreover, it is of great importance to reconcile randomization-based and regression-based inferences, i.e., the point estimators of the factorial effects and their corresponding confidence regions. However, although the equivalence between randomization-based and regression-based inferences for randomized treatment-control studies (i.e., 2^1 factorial designs) has been well established in the existing literature (Schochet, 2010; Samii and Aronow, 2012; Lin, 2013), similar discussions for 2^K factorial designs appear to be absent. In this paper, we fulfill the aforementioned two-fold task.

The paper proceeds as follows. Section 2 extends the randomization-based inference framework in Dasgupta et al. (2015) to general 2^K factorial designs. Section 3 demonstrates the equivalence between randomization-based and regression-based inferences for 2^K factorial designs. Section 4 considers extensions, and Section 5 concludes and discusses possible future directions.

E-mail address: jiannl@microsoft.com.

http://dx.doi.org/10.1016/j.spl.2016.01.010
0167-7152/© 2016 Published by Elsevier B.V.

2. Randomization-based inference for general 2^K factorial designs

2.1. 2^K factorial designs

Consider K distinct factors, each with two levels -1 and 1. We construct the model matrix (Wu and Hamada, 2009) $\boldsymbol{H} = (\boldsymbol{h}_0, \dots, \boldsymbol{h}_{2^K-1})$ as follows (Espinosa et al., 2015):

- let $\boldsymbol{h}_0 = \mathbf{1}_{2^K}$;
- for $k = 1, \dots, K$, construct $\boldsymbol{h}_k$ by letting its first 2^{K-k} entries be -1, the next 2^{K-k} entries be 1, and repeating 2^{k-1} times;
- for $k = K+1, \dots, K + \binom{K}{2}$, let $\boldsymbol{h}_k = \boldsymbol{h}_{k_1} \cdot \boldsymbol{h}_{k_2}$, where $k_1, k_2 \in \{1, \dots, K\}$;

  ...
- let $\boldsymbol{h}_{J-1} = \boldsymbol{h}_1 \cdot \dots \cdot \boldsymbol{h}_K$.

For $j = 1, \dots, 2^K$, let $\tilde{\boldsymbol{h}}_j$ denote the jth row of the model matrix $\boldsymbol{H}$. A well-known fact is that the model matrix $\boldsymbol{H}$ is orthogonal, i.e.,

$$\boldsymbol{H}\boldsymbol{H}' = (\tilde{\boldsymbol{h}}_j \tilde{\boldsymbol{h}}'_{j'})_{2^k \times 2^K} = 2^K \boldsymbol{I}_{2^K}, \qquad \boldsymbol{H}'\boldsymbol{H} = \sum_{j=1}^{2^K} \tilde{\boldsymbol{h}}'_j \tilde{\boldsymbol{h}}_j = 2^K \boldsymbol{I}_{2^K}. \tag{1}$$

The jth row of $\tilde{\boldsymbol{H}} = (\boldsymbol{h}_1, \dots, \boldsymbol{h}_K)$ is the jth treatment combination $\boldsymbol{z}_j$, and the columns of $\boldsymbol{H}$ define the factorial effects. To be specific, the first column $\boldsymbol{h}_0$ corresponds to the null effect, the next K columns $\boldsymbol{h}_1, \dots, \boldsymbol{h}_K$ correspond to the main effects of the K factors, the next $\binom{K}{2}$ columns $h_{K+1}, \dots, \boldsymbol{h}_{K+\binom{K}{2}}$ correspond to the two-way interactions, etc., and eventually the last column $\boldsymbol{h}_{J-1}$ corresponds to the K-factor interaction.

Example 1. For 2^2 factorial designs, the model matrix is:

$$\boldsymbol{H} = \begin{array}{c} \\ \tilde{\boldsymbol{h}}_0 \\ \tilde{\boldsymbol{h}}_1 \\ \tilde{\boldsymbol{h}}_2 \\ \tilde{\boldsymbol{h}}_3 \end{array} \begin{array}{c} \begin{array}{cccc} \boldsymbol{h}_0 & \boldsymbol{h}_1 & \boldsymbol{h}_2 & \boldsymbol{h}_3 \end{array} \\ \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \end{array}.$$

The four treatment combinations are $\boldsymbol{z}_1 = (-1, -1)$, $\boldsymbol{z}_2 = (-1, 1)$, $\boldsymbol{z}_3 = (1, -1)$ and $\boldsymbol{z}_4 = (1, 1)$. We represent the main effects of factors 1 and 2 by $\boldsymbol{h}_1 = (-1, -1, 1, 1)'$ and $\boldsymbol{h}_2 = (-1, 1, -1, 1)'$ respectively, and the two-way interaction by $\boldsymbol{h}_3 = (1, -1, 1, -1)'$.

2.2. Randomization-based inference

For consistency, we adopt the notations in Dasgupta et al. (2015). Let $N \geq 2^{K+1}$ be the number of experimental units. Under the Stable Unit Treatment Value Assumption (Rubin, 1980), for unit i, we denote its potential outcome under treatment combination $\boldsymbol{z}_j$ as $Y_i(\boldsymbol{z}_j)$, for all $j = 1, \dots, 2^K$. Let $\boldsymbol{Y}_i = \{Y_i(\boldsymbol{z}_1), \dots, Y_i(\boldsymbol{z}_{2^K})\}'$, and we define the factorial effect vector of unit i as

$$\boldsymbol{\tau}_i = \frac{1}{2^{(K-1)}} \boldsymbol{H}' \boldsymbol{Y}_i. \tag{2}$$

Having defined the potential outcomes and factorial effects on the individual-level, we shift focus to the population-level. For all j, we let

$$\bar{Y}(\boldsymbol{z}_j) = \frac{1}{N} \sum_{i=1}^{N} Y_i(\boldsymbol{z}_j)$$

be the average potential outcome under treatment combination $\boldsymbol{z}_j$, across all experimental units. Let $\bar{\boldsymbol{Y}} = \{\bar{Y}(\boldsymbol{z}_1), \dots, \bar{Y}(\boldsymbol{z}_{2^K})\}'$, and we define the population-level factorial effect vector as

$$\boldsymbol{\tau} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\tau}_i = \frac{1}{2^{(K-1)}} \boldsymbol{H}' \bar{\boldsymbol{Y}}. \tag{3}$$

We consider general 2^K factorial designs. For $j = 1, \dots, 2^K$, we randomly assign $n_j \geq 2$ units to treatment $\boldsymbol{z}_j$. Note that $\sum_{j=1}^{2^K} n_j = N$. For unit i, we let

$$W_i(\boldsymbol{z}_j) = \begin{cases} 1, & \text{if unit } i \text{ is assigned treatment } \boldsymbol{z}_j, \\ 0, & \text{otherwise.} \end{cases}$$