HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Statistics

have examined a dissertation entitled

A Bayesian Perspective on Factorial Experiments Using Potential Outcomes

presented by **Valeria Espinosa**

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Professor Donald B. Rubin

Signature Tirthankar Dasgupta

Typed name: Professor Tirthankar Dasgupta

Signature _____

Typed name: Professor Luke Miratrix

Date: December 5, 2013

A Bayesian Perspective on Factorial Experiments Using Potential Outcomes

A dissertation presented

by

Valeria Espinosa

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

December 2013

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Dissertation Advisors: Professors Tirthankar Dasgupta and Donald Rubin *Valeria Espinosa*

A Bayesian Perspective on Factorial Experiments Using Potential Outcomes

Abstract

Factorial designs have been widely used in many scientific and industrial settings, where it is important to distinguish "active" or real factorial effects from "inactive" or noise factorial effects used to estimate residual or "error" terms. We propose a new approach to screen for active factorial effects from such experiments that utilizes the potential outcomes framework and is based on sequential posterior predictive model checks. One advantage of the proposed method lies in its ability to broaden the standard definition of active effects and to link their definition to the population of interest. Another important aspect of this approach is its conceptual connection to Fisherian randomization tests. As in the literature in design of experiments, the unreplicated case receives special attention and extensive simulation studies demonstrate the superiority of the proposed Bayesian approach over existing methods. The unreplicated case is also thoroughly explored. Extensions to three level and fractional factorial designs are discussed and illustrated using a classical seat belt example for the former and part of a stem-cell research collaborative project for the latter.

Contents

Acknowledgments

I am especially thankful for my two advisors, Don Rubin and Tirthankar Dasgupta; for their constant support and advice. I thank Don Rubin for sharing his insightful and fascinating view of the world of statistics and, in particular, for the link between randomization tests and posterior predictive checks. I thank Tirthankar Dasgupta for immersing me in the world of design of experiments and sharing his wisdom and passion for the field. I thank the three members of my committee Don Rubin, Tirthankar Dasgupta and Luke Miratrix for very helpful comments, suggestions and discussions on the material presented in this thesis. I thank Dr. Quinn Peterson for the opportunity to work on an exciting and innovative collaborative project in stem cell research. I thank Viviana García, Jonathan Hennessy, Joseph Kelly, Nathan Stein and Samuel Wong for working with me on interesting projects and for the enthusiastic discussions about them.

I thank my parents Enrique and Rosa Luisa, for being the never-ending source of strength to keep going after stumbling and for always believing in me. I thank my brother Daniel, for encouraging me to continue exploring different paths and new challenges. I am thankful for these years in graduate school, for my professors, my students, my classmates and the wonderful friendships created both inside and outside Harvard; for Gabriela and Viviana. I thank Jon for making me a part of his world, within statistics and beyond.

To Enrique, Rosa Luisa and Daniel

Chapter 1

Introduction and importance of unreplicated experiments

Two-level full factorial designs have been used extensively in engineering and industrial applications. In many situations, where the unit-to-unit variation is reasonably assumed negligible, unreplicated factorial designs are used to reduce expenses. Screening factorial effects into active and inactive from such designs has been extensively studied by researchers. Hamada and Balakrishnan (1998) offer an extensive review and comparison of many of the methods, most of which rely on the assumption that the estimated factorial effects are independently and identically distributed (iid) normal random variables (e.g., the commonly used Lenth (1989) approach and Dong's (1993) method). Loughin and Noble (1997), proposed a method based on permutation tests, which is not included in the Hamada and Balakrishnan (1998) review. Focusing on the control of the false discovery rate (FDR) instead of the experimentwise error rate (EER), Tripolski et al. (2008) proposed modifications of the Lenth (1989) and Dong (1993) methods, and they performed a comparative study

with the *unmodified* methods. Bayesian approaches have also been proposed for screening factorial effects, for example Box and Meyer (1986) and Chipman et al. (1997). The former was included in the Hamada and Balakrishnan (1998) study but did not perform as well as Lenth's (1989) method. The approach proposed by Chipman et al. (1997) is, in principle, similar to the one proposed by Box and Meyer (1986), but has greater flexibility in terms of incorporating prior information through the effect heredity and effect hierarchy principles (Wu and Hamada (2009), Ch. 4).

Discriminating between active and inactive effects is the crucial goal of any screening experiment. In all the aforementioned frequentist methods, hypotheses regarding factorial effects are stated in terms of regression coefficients in the classical linear model. On the other hand, in the Bayesian approaches, active and inactive effects are distinguished in terms of their variance; whereas all effects are assumed to be normally distributed with mean zero, the standard deviation of active effects is assumed to be at least k times larger than the standard deviation of inactive effects, where k is a pre-set integer (both Box and Meyer (1986) and Chipman et al. (1997) suggest using $k \approx 10$).

Our motivation for proposing a new approach stems from the fact that the definition of an "active" effect is somewhat vague because it is not related to the experimental units or the population of units of interest to an experimenter but to model parameters. Consider, for example, experiments involving growth of nanostructures on substrates of silicon (Dasgupta et al., 2008), which is somewhat analogous to the yield of crops on plots of lands. Suppose that we are interested in assessing the effect of temperature on yield; whether a change of temperature increases or decreases the yield on one or more substrates. Considering the fact that the inference made from a small population of substrates in a particular laboratory can

hardly be extended and generalized to a larger population, a natural question is whether the temperature affects the yield of at least one of the substrates used for experimentation. Further, if we visualize a *potential yield* of each substrate for each level of temperature, then we may be interested in a summary of the distribution of these potential yields across units, e.g., the median or a percentile, rather than the average. None of the existing methods permit such an analysis.

The essential idea that develops from the foregoing discussion is that making an attempt to assess significance of factorial effects, without first defining both (a) the population of experimental units for which the inference is made and (b) the estimand, is not the right way to address *causal inference* questions, which is the sole objective of conducting a screening two-level factorial experiment. Here, we propose a Bayesian approach for screening active factorial effects from such designs, which addresses the limitations of current procedures. The proposed approach utilizes the concept of potential outcomes that lies at the center stage of causal inference Rubin (1974, 1980). Although such a framework for single-factor experiments with two levels is well-developed and popularly known as the Rubin Causal Model (Holland, 1986), RCM, it is not yet fully exploited for multiple-factor experiments. A theoretical framework for causal inference from two-level factorial designs has recently been proposed by Dasgupta et al. (2012).

In the next Section, we provide a brief historical review of the potential outcomes framework and the RCM, and describe how it can be applied to two level factorial designs. In Section 3, we describe the Fisher randomization test (Fisher, 1925, 1935) using the potential outcomes framework, extend it to the setting of two level factorial designs, and establish its connection to the permutation tests proposed by Loughin and Noble (1997). In Section

4, we show how the Fisherian approach to causal inference can be naturally extended to a Bayesian approach to screening factorial effects and propose a method based on sequential posterior predictive checks. In Section 5, we demonstrate the usefulness of our method in a super population setting by first calibrating the proposed algorithm to achieve the desired experimentwise error rate (EER), and then by comparing its performance with that of existing methods for screening factorial effects. Some concluding remarks are presented in Section 6.

1.1 A brief review of potential outcomes, evolution of RCM and extension to two-level factorial designs

As noted by Dasgupta et al. (2012), in the context of randomized experiments, Neyman (1923; 1990) introduced the first explicit notation for potential outcomes for randomization-based inference. Subsequently, Kempthorne (1955) and Cox (1958) continued its use for causal inference in randomized experiments. Later, Rubin (1974, 1975, 1977, 1978) formalized the concept and extended it to other forms of causal inference, including observational studies. An exposition of this transition appears in Rubin (2010).

The RCM was motivated by the need for a clear separation between the object of inference – *the science* – and what researchers do to learn about it (e.g., randomly assign treatments to units). In the context of causal inference, the science is usually represented by a matrix where the rows represent N units, which are physical objects at a particular point in time, and the columns represent every unit's potential outcomes under each possible exposure. Thus, for a study with one outcome variable Y and one experimental factor at two levels,

represented by 1 and -1, each row of the science can be written as $[Y_i(1), Y_i(-1)]$, where $Y_i(z)$ is the potential outcome of unit i if unit i receives treatment z, $z \in \{-1, 1\}$, indicated by $W_i(z) = 1$. This representation of the science is adequate under the stable unit treatment value assumption (SUTVA, Rubin (1980)).

The causal effect of Treatment 1 versus Treatment -1 for the ith unit is the comparison of the corresponding potential outcomes for that unit: $Y_i(1)$ versus $Y_i(-1)$ (e.g., their difference or their ratio). The "fundamental problem facing inference for causal effects" (Rubin, 1978) is that only one of the potential outcomes can ever be observed for each unit. Therefore, unit-level causal effects cannot be known and must be inferred. The RCM permits prediction of unit-level causal effects from either the Neymanian (1923/1990) perspective or from the Bayesian perspective (Rubin, 1978), although such estimations are generally imprecise relative to the estimation of population or subpopulation causal effects.

Although the average causal effects $\bar{Y}(1) - \bar{Y}(-1)$ are common estimands in many fields of application, other summaries of unit-level causal effects may also be of interest in specific scientific studies. As has been emphasized by Rubin (1974, 1975, 1977, 1978, 1980, 1984, 1990, 2008, 2010), there is no reason to focus solely on average causal effects, although this quantity is especially easy to estimate unbiasedly in randomized experiments under simple assumptions using standard statistical tools .

Rubin (2010) describes the RCM in terms of three legs – the first being to define causal effects using potential outcomes (define the science), the second being to describe the process by which some potential outcomes will be revealed (the assignment mechanism), and the third being the Bayesian posterior predictive distribution of the missing potential outcomes.

1.1.1 RCM for two-level factorial designs

For simplicity, an unreplicated 2^2 design is used to introduce the concepts, even though such a design is usually not of much use in practice. In this design, each of the two treatment factors can take one of two levels, typically denoted by -1 and 1, and thus, there are four treatment combinations denoted by $\boldsymbol{z} = (1,1), (1,-1), (-1,1)$ and $(-1,-1)$, and four experimental units. Let $Y_i(\boldsymbol{z}), i = 1, \ldots, 4$, denote the potential outcome of the ith unit if exposed to treatment combination $\boldsymbol{z}$. Note that when introducing this notation, we accept SUTVA, which means that the potential outcome of a particular unit depends only on the treatment combination it is assigned, and NOT on the assignments of the other units, and that there are no hidden versions of treatments not represented by the four combinations. Thus the ith unit has four potential outcomes $Y_i(1,1), Y_i(1,-1), Y_i(-1,1), Y_i(-1,-1)$, which comprise the 1×4 row vector $\boldsymbol{Y}_i$. Finally, we define *the Science* as the 4×4 matrix $\boldsymbol{Y}$ of potential outcomes in which the ith row is the 4-component row vector $\boldsymbol{Y}_i$, $i = 1, \ldots, 4$ as shown in Table 1.1. Only one potential outcome in each row of the Science is actually observed from an experiment, and the remaining three are missing, making the causal inference problem essentially a missing data problem.

Table 1.1: The *Science* for the full experiment.

Unit	Potential outcome for treatment combination				Unit-level factorial effects			
(i)	$(1,1)$	$(1,-1)$	$(-1,1)$	$(-1,-1)$	$\theta_{i,0}$	$\theta_{i,1}$	$\theta_{i,2}$	$\theta_{i,3}$
1	$Y_1(1,1)$	$Y_1(1,-1)$	$Y_1(-1,1)$	$Y_1(-1,-1)$	$\theta_{1,0}$	$\theta_{1,1}$	$\theta_{1,2}$	$\theta_{1,3}$
2	$Y_2(1,1)$	$Y_2(1,-1)$	$Y_2(-1,1)$	$Y_2(-1,-1)$	$\theta_{2,0}$	$\theta_{2,1}$	$\theta_{2,2}$	$\theta_{2,3}$
3	$Y_3(1,1)$	$Y_3(1,-1)$	$Y_3(-1,1)$	$Y_3(-1,-1)$	$\theta_{3,0}$	$\theta_{3,1}$	$\theta_{3,2}$	$\theta_{3,3}$
4	$Y_4(1,1)$	$Y_4(1,-1)$	$Y_4(-1,1)$	$Y_4(-1,-1)$	$\theta_{4,0}$	$\theta_{4,1}$	$\theta_{4,2}$	$\theta_{4,3}$
Average	$\overline{Y}(1,1)$	$\overline{Y}(1,-1)$	$\overline{Y}(-1,1)$	$\overline{Y}(-1,-1)$	θ_0	θ_1	θ_2	θ_3

For each unit, all levels of every factor (e.g., 1 or -1) appear on half of its potential outcomes. Therefore, at the unit-level we are generally interested in contrasting one half of the unit's potential outcomes with the other half. For example, the difference of the averages of the potential outcomes when factor 1 is at its high level (1) and at its low level (-1), is the so-called "main effect of factor 1". Of course, other definitions could be the difference in medians, or the difference in the logarithm of the averages, but the tradition in the study of factorial experiments is to deal with the difference of averages, and we adhere to this focus here. A factorial effect for each unit is the difference in the averages between one half of the potential outcomes and the other half. Consequently, we define three unit-level factorial effects representing the main effects of the two factors and their interactions as three contrasts (denoted by $\theta_{i,1}, \theta_{i,2}$ and $\theta_{i,3}$ respectively) of elements of the vector $\boldsymbol{Y}_i$. These contrasts are:

$$\begin{aligned}
\theta_{i,1} &= \frac{Y_i(1,1)+Y_i(1,-1)}{2} - \frac{Y_i(-1,1)+Y_i(-1,-1)}{2} = \frac{1}{2}\boldsymbol{Y}_i\boldsymbol{g}_1, \\
\theta_{i,2} &= \frac{Y_i(1,1)+Y_i(-1,1)}{2} - \frac{Y_i(1,-1)+Y_i(-1,-1)}{2} = \frac{1}{2}\boldsymbol{Y}_i\boldsymbol{g}_2, \\
\theta_{i,3} &= \frac{Y_i(1,1)+Y_i(-1,-1)}{2} - \frac{Y_i(1,-1)+Y_i(-1,1)}{2} = \frac{1}{2}\boldsymbol{Y}_i\boldsymbol{g}_3,
\end{aligned} \tag{1.1}$$

where $\boldsymbol{g}_1$, $\boldsymbol{g}_2$ and $\boldsymbol{g}_3$ are the three mutually orthogonal contrast column vectors $(1,1,-1,-1)'$, $(1,-1,1,-1)'$ and $(1,-1,-1,1)'$, where each element of $\boldsymbol{g}_3$ is obtained by multiplying the corresponding elements of $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$. For completeness, we label the vector generating the ith unit's mean potential outcome as $\boldsymbol{g}_0 = (1,\ldots,1)'$, which is orthogonal to $\boldsymbol{g}_1$, $\boldsymbol{g}_2$ and $\boldsymbol{g}_3$, so that this average potential outcome is

$$\theta_{i,0} = \frac{Y_i(1,1)+Y_i(1,-1)+Y_i(-1,1)+Y_i(-1,-1)}{4} = \frac{1}{4}\boldsymbol{Y}_i\boldsymbol{g}_0.$$

Defining

$$\boldsymbol{\theta}_i = (\theta_{i,0}, \theta_{i,1}/2, \theta_{i,2}/2, \theta_{i,3}/2),$$

as the 1×4 row vector of unit-level factorial effects, as shown in the last four columns of Table 1, and denoting by $\boldsymbol{G}$ the 4×4 matrix whose columns are the vectors $\boldsymbol{g}_0, \boldsymbol{g}_1, \boldsymbol{g}_2$ and $\boldsymbol{g}_3$, from (1.1), it immediately follows that

$$\boldsymbol{Y}_i = \boldsymbol{\theta}_i \boldsymbol{G}', \tag{1.2}$$

implying that $\boldsymbol{Y}_i$ and $\boldsymbol{\theta}_i$ are linear transformations of each other. The matrix $\boldsymbol{G}$ is often referred to as the model matrix in the literature (Wu and Hamada (2009), Ch. 4).

Consistent with the traditional definition of causal effects in the factorial design literature, the causal estimands at the *population level* could be the averages of the unit level factorial effects. These quantities, denoted by θ_1, θ_2 and θ_3, are the population level main effects of each treatment factor and their interaction respectively, and can be expressed in terms of potential outcomes as:

$$\theta_j = \frac{\sum_{i=1}^{4} \theta_{i,j}}{4} = \frac{1}{2} \overline{\boldsymbol{Y}} \boldsymbol{g}_j, \quad j = 1, 2, 3, \tag{1.3}$$

where

$$\bar{\boldsymbol{Y}} = \frac{1}{4} \sum_{i=1}^{4} \boldsymbol{Y}_i. \tag{1.4}$$

As in the case of unit-level effects, denoting by $\theta_0 = \sum_{i=1}^{N} \theta_{i,0}/4 = \frac{1}{4} \overline{\boldsymbol{Y}} \boldsymbol{g}_0$, and the row vector of population-level estimands by $\boldsymbol{\theta} = (\theta_0, \theta_1/2, \theta_2/2, \theta_3/2)$, it immediately follows that

$$\overline{\boldsymbol{Y}} = \boldsymbol{\theta} \boldsymbol{G}'. \tag{1.5}$$

As mentioned earlier, for each unit *only one* potential outcome is observed: the one that corresponds to the treatment the unit is assigned to receive; the other outcomes are *missing*. The treatment assignment mechanism selects the subset of potential outcomes that will be revealed and observed. The following assignment mechanism can be defined for a 2^2 factorial design

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

For an unreplicated completely randomized 2^2 factorial experiment, $Pr(W_i(\boldsymbol{z}) = 1) = 1/4$, where the probability $Pr(\cdot)$ is implicitly conditional on the science. Also, $\sum_z W_i(\boldsymbol{z}) = 1$ for $i = 1, \ldots, 4$, and $\sum_i W_i(\boldsymbol{z}) = 1$ for all $\boldsymbol{z}$. Let $w_i = \sum_{\boldsymbol{z}} \boldsymbol{z} W_i(\boldsymbol{z})$ be the treatment combination that the ith subject receives. Let $\boldsymbol{W}$ be the generic treatment assignment vector of random variables, and let $\boldsymbol{w}$ be a specific realization of $\boldsymbol{W}$, i.e., a vector that contains all the individual treatment assignments *after* randomization. Hence, each W_i is a random variable, and their joint probability distribution defines the treatment assignment mechanism of $\boldsymbol{W}$, implicitly conditional on the science. The vector of post randomization treatment assignments, $\boldsymbol{w}$, is a draw from this distribution.

Denote the observed outcome corresponding to the ith experimental unit by $Y_i^{\text{obs}}, i = 1, \ldots, 4$, so that

$$Y_i^{\text{obs}} = \sum_{\boldsymbol{z}} W_i(\boldsymbol{z}) Y_i(\boldsymbol{z}), \tag{1.6}$$

and by $\boldsymbol{Y}^{\text{obs}} = (Y_i^{\text{obs}}, \ldots, Y_4^{\text{obs}})'$, the 4×1 column vector of observed outcomes. For a *given* treatment assignment $\boldsymbol{w} = \big((1,1), (-1,-1), (-1,1), (1,-1)\big)'$, the table of *observed* potential outcomes looks analogous to the one displayed in Table 1.2. The missing potential outcomes are represented by question marks.

Table 1.2: *Observed Outcomes* for the full experiment with $\boldsymbol{w} = \big((1,1),(-1,-1),(-1,1),(1,-1)\big)'$.

Unit	Observed outcome for treatment combination				$\boldsymbol{w}$
(i)	$(1,1)$	$(1,-1)$	$(-1,1)$	$(-1,-1)$	
1	Y_1^{obs}	?	?	?	$(1,1)$
2	?	?	?	Y_2^{obs}	$(-1,-1)$
3	?	?	Y_3^{obs}	?	$(-1,1)$
4	?	Y_4^{obs}	?	?	$(1,-1)$

Let $Y^{\text{obs}}(\boldsymbol{z})$ denote the observed outcome for treatment combination $\boldsymbol{z}$, and let

$$\tilde{\boldsymbol{Y}}^{\text{obs}} = \big(Y^{\text{obs}}(1,1), Y^{\text{obs}}(1,-1), Y^{\text{obs}}(-1,1), Y^{\text{obs}}(-1,-1)\big)'$$

denote the vector of observed outcomes arranged according to the *natural order* of the treatment combinations. By *natural order* we refer to ordering the treatment combinations by specifying the levels of factors 1 through K, starting with the value 1 and then -1. That is, corresponding to factor j, the vector g_j is constructed by defining the first $N/2^j$ entries to be 1, the next $N/2^j$ entries to be -1, and repeating 2^{j-1} times (until the N entries of g_j are defined). The first treatment combination corresponds to the vector z defined by the first entry of each g_j, the second treatment combination corresponds to the second entries, and so on.

Therefore, $\tilde{\boldsymbol{Y}}^{\text{obs}}$ is simply a permutation of $\boldsymbol{Y}^{\text{obs}}$. Then, the same one to one relationship defined by the matrix $\boldsymbol{G}$ can be used, together with $\tilde{\boldsymbol{Y}}^{\text{obs}}$ to obtain unbiased estimators of population-level factorial effects θ_j given by (1.3) are:

$$\hat{\theta}_j = \frac{1}{2}\tilde{\boldsymbol{Y}}^{\text{obs}}\boldsymbol{g}_j, \quad j = 1,2,3. \tag{1.7}$$

The unbiasedness and other sampling properties of the estimators $\hat{\theta}_j$ under its randomization distribution for a general 2^K factorial design have been studied in details by Dasgupta et al. (2012), also a proof of unbiasedness appealing to symmetry arguments is given in the Appendix. Finally, denoting the vector of estimators by $\hat{\boldsymbol{\theta}} = (\bar{Y}^{\text{obs}}, \hat{\theta}_1/2, \hat{\theta}_2/2, \hat{\theta}_3/2)$, as in (1.5), it follows that

$$(\tilde{\boldsymbol{Y}}^{\text{obs}})' = \hat{\boldsymbol{\theta}}\boldsymbol{G}', \text{ or equivalently } \tilde{\boldsymbol{Y}}^{\text{obs}} = \boldsymbol{G}\hat{\boldsymbol{\theta}}',$$

from which, using the identity $\boldsymbol{G}'\boldsymbol{G} = 4\boldsymbol{I}$, we have

$$\hat{\boldsymbol{\theta}}' = (\boldsymbol{G}'\boldsymbol{G})^{-1}\boldsymbol{G}'\tilde{\boldsymbol{Y}}^{\text{obs}}. \tag{1.8}$$

This establishes the fact that the unbiased *point estimators* of the estimands defined in this Section are the same as those obtained by ordinary least squares in the classical linear model with an additive iid error. However, to perform inference, such as interval estimation or significance tests using the least squares formulation, we need to justify the assumption of additive errors and asymptotic normality of the estimators. In the set-up described so far in this section, the potential outcomes are considered fixed, and the estimands are functions of a finite population of size four. Therefore the only sampling distribution of the estimators that arises is the one induced by randomization.

It is now evident that inferential methods such as Lenth's test, that are based on asymptotic normality, cannot distinguish between the following two situations: (i) the four units considered here are the only ones that constitute the population of interest and the estimands θ_j's are those defined by (1.3), versus (ii) the four units are randomly sampled from

an infinitely large super-population and the estimands are counterparts of θ_j's for this super population. Methods that rely on asymptotic normality typically aim at addressing situation (ii).

There are two standard inferential methods that address situation (i) under the potential outcomes framework: (a) the Neymanian approach and (b) the Fisherian approach. We will not discuss the Neymanian approach here (see Dasgupta et al. (2012) for details), but we will discuss the Fisherian approach for two reasons. First, it has a natural connection to permutation tests, proposed by Loughin and Noble (1997) for the analysis of unreplicated two-level factorial experiments. Second, it has a natural Bayesian justification, which will be used to define the Bayesian approach that we eventually propose.

We conclude this chapter by noting that the framework illustrated using a 2^2 design can easily be extended to a 2^K design. For such a design, we have K factors labeled $1, \ldots, K$, $N = 2^K$ experimental units and $N - 1$ mutually orthogonal vectors $\boldsymbol{g}_j$, each representing a factorial effect θ_j. The first K vectors (and the corresponding θ_j's) represent the K main effects, the next $\binom{K}{2}$ vectors the two-factor interactions, and eventually the $(N-1)$-th vector represents the K-factor interaction. Thus, for example, for a 2^4 design, the factorial effect θ_3 represents the main effect of factor 3, θ_5 the interaction between factors 1 and 2, θ_{10} the two-factor interaction between factors 3 and 4, and θ_{15} the four factor interaction. The model matrix $\boldsymbol{G}$ is an $N \times N$ orthogonal matrix with an N-vector $\boldsymbol{g}_0 = (1, \ldots, 1)'$ as its first column, and such that $\boldsymbol{G}'\boldsymbol{G} = 1/2^K \boldsymbol{I}_{2^K} = 1/N \boldsymbol{I}_{2^K}$ where $\boldsymbol{I}_{2^K}$ is the 2^K identity matrix.

Chapter 2

The Unreplicated Case: Randomization tests for sharp null hypotheses

In the context of **unreplicated** factorial designs, we first present the permutation test in the potential outcomes framework. This discussion is key to understanding the proposed randomization-based methods. This is the most natural permutation test and it illustrates fundamental ideas of the potential outcomes framework in the Fisherian testing context. We then propose a sequential randomization-based method to screen active effects. This method initially presented an attempt to understand the method proposed in Loughin and Noble (1997) in the potential outcomes framework. However, our choice of test statistic was taken from the adaptive Lenth method (Ye et al., 2001): the estimated maximum absolute effect, scaled by the pseudo standard error, where both of these quantities are calculated using the

effects assumed null for that step. That is:

$$T_{L,i} = \frac{|\hat{\theta}_i|}{\text{PSE}_i},$$

where PSE_i is the PSE of $\hat{\theta}_{(2^k-1)}, \hat{\theta}_{(2^k-2)}, \ldots, \hat{\theta}_{(2^k-i)}$, as defined in Chapter 4. Nevertheless, other scalings might yield better results.

2.1 The Fisherian approach: randomization tests for sharp null hypothesis

Randomization tests (Fisher 1925,1935) are useful tools because they assess statistical significance of treatment effects from randomized experiments without making any assumption whatsoever about the distribution of the test statistic. Such tests can be used to test Fisher's *sharp null hypothesis* (see Fisher (1925) and Rubin (1980)) of no factorial effect at the unit levels, which is a much stronger hypothesis than the traditional one of no average factorial effects. Randomization tests have rarely been studied in the context of factorial experiments, except for the work by Loughin and Noble (1997), who studied such tests in the framework of a linear regression model for the observed response with additive error (in other words, invoking the assumption of strict additivity).

The plausibility of the sharp null hypothesis can be tested by using a randomization test. A suitable test statistic is chosen, and its observed value is compared with the statistic's sampling distribution induced by the randomization under the sharp null hypothesis. To obtain such a distribution, typically referred to as the "randomization distribution", we

enumerate all possible treatment assignments under the actual assignment mechanism (if the number of such assignments is very large, a sample can be considered). The assumption that the sharp null hypothesis is true permits us to complete the table of all missing potential outcomes using only the observed data. For example, if the sharp null hypothesis of no treatment effect on any unit is true, then all the missing potential outcomes in the ith row of Table 1.2 are equal to Y_i^{obs}. Consequently, under the sharp hypothesis, the table of potential outcomes would look like Table 2.1, where the missing potential outcomes for each unit, $\mathbf{Y}_i^{\text{mis}}$, are shown in a light gray, but under the sharp null hypothesis they take the same value as the observed one, Y_i^{obs}.

Table 2.1: Imputed Table of Potential Outcomes using the observed data and the sharp null hypothesis of absolutely no treatment effects

Unit	Observed outcome for treatment combination				$\boldsymbol{w}$
(i)	$(\ \ 1,\ \ 1)$	$(\ \ 1,-1)$	$(-1,\ \ 1)$	$(-1,-1)$	
1	Y_1^{obs}	Y_1^{obs}	Y_1^{obs}	Y_1^{obs}	$(\ \ 1,\ \ 1)$
2	Y_2^{obs}	Y_2^{obs}	Y_2^{obs}	Y_2^{obs}	$(-1,-1)$
3	Y_3^{obs}	Y_3^{obs}	Y_3^{obs}	Y_3^{obs}	$(-1,\ \ 1)$
4	Y_2^{obs}	Y_4^{obs}	Y_4^{obs}	Y_2^{obs}	$(\ \ 1,-1)$

For each possible assignment, the value of the test statistic that would have been observed under that assignment is calculated. The proportion of such computed values (with respect to the total number of possible randomizations) that are as large or larger than the actual observed test statistic is the p value (i.e., significance level) of the test statistic under the null hypothesis. The smaller the p-value, the greater is the degree of belief that the null hypothesis is not true, because the probability of the one observed result, even when that probability is combined with all more extreme results, would be a rare event.

There are many possible test statistics that could be used. A commonly used statistic

in the presence of replicates is the F-statistic associated with the decomposition of the total sum of squares of the observed outcomes. For the unreplicated case, the estimated factorial effect $\hat{\theta}_{(N-1)}$ that has the largest absolute value among all the $N-1$ estimated effects $\hat{\theta}_j, \quad j = 1, \ldots, N-1$ defined by (1.7) can be considered. A scaled version of $\hat{\theta}_{(N-1)}$ can also be considered.

Therefore, for a completely randomized treatment assignment mechanism, the Fisher randomization test with the test statistic $\hat{\theta}_{(N-1)}$ involves the following steps:

1. Compute $\hat{\theta}_{(N-1)}$ from the observed data and denote it by $\hat{\theta}^{\text{obs}}_{(N-1)}$.

2. Fill in the table of missing potential outcomes using the observed values Y_i^{obs} for $i = 1, \ldots, N$, and the sharp null hypothesis.

3. For each of the $N!$ possible treatment assignments of N units to N treatment combinations, generate the observed outcomes, implied by step 2.

4. Compute $\hat{\theta}_{(N-1)}$ for each of the $N!$ assignments. The set of values of $\hat{\theta}_{(N-1)}$ are its realizations across the randomization distribution under the sharp null hypothesis.

5. Compute the p-value as the proportion of values of $\hat{\theta}_{(N-1)}$ that are equal to or exceed $\hat{\theta}^{\text{obs}}_{(N-1)}$.

Because the execution of steps (2)-(3) described above essentially entails generating all possible permutations of the vector of observed outcomes, the test described above is the same as the first step of the permutation test proposed by Loughin and Noble (1997). We would like to re-emphasize the point that the procedure described above is more general and flexible because it permits generating the randomization distribution of *any* test statistic under *any* treatment assignment mechanism, under *any* sharp null hypothesis.

It is imperative that once the sharp null hypothesis described above is rejected and so $\theta_{(N-1)}$ is considered active, one needs to adopt a sequential approach for further screening of factorial effects. At every step, however, a new null hypothesis, that must take into consideration the effects already identified as active, needs to be defined. We illustrate such a strategy again with a 2^2 design. The first step involves testing the sharp null hypothesis,

$$H_{00} : Y_i(1,1) = Y_i(1,-1) = Y_i(-1,1) = Y_i(-1,-1), \quad i = 1, \ldots, N$$

which is equivalent to

$$H_{00} : \theta_{i,1} = \theta_{i,2} = \theta_{i,3} = 0, \quad i = 1, \ldots, N$$

To test this hypothesis, we can use the procedure described above in Steps 1–5. If the p-value is small enough to lead to the rejection of the null hypothesis, we move on to the next step. Without loss of generality (WLOG), assume that the factorial effect identified as active is θ_1, the main effect of factor 1. Assuming that the effect of factor 1 is additive (the same for each unit), we *temporarily* define the second *sharp* null hypothesis as:

$$H_{01} : \theta_{i,1} = \theta_1^*, \theta_{i,2} = \theta_{i,3} = 0, \quad i = 1, \ldots, N$$

where θ_1^* is the estimated value of θ_1. This hypothesis can be tested using $\hat{\theta}_{(2)}$, the estimated effect having the second largest absolute value, as the test statistic and using steps 1–5 as before. Again, assuming WLOG that the second largest active effect is identified as θ_2, and the p-value for H_{02} is small enough to lead to its rejection, we define the third and final *sharp*

null hypothesis as:

$$H_{02} : \theta_{i,1} = \theta_1^*, \theta_{i,2} = \theta_2^*, \theta_{i,3} = 0, \quad i = 1, \ldots, N$$

where θ_2^* is the estimated value of θ_2.

One important aspect about the aforementioned sequential procedure is to fill in the missing potential outcomes at each step, i.e., under each sharp null hypothesis. Defining $\dot{\boldsymbol{\theta}}_i$ as the $(N-1)$-vector $\boldsymbol{\theta}_i$ without its first element θ_{i0} (i.e., all the unit level factorial effects), we can express the sequence of sharp null hypotheses as

$$H_{0s} : \dot{\boldsymbol{\theta}}_i = \dot{\boldsymbol{\theta}}_i^{(s)} \ \ \forall \ \ i = 1, \ldots, N,$$

where $\dot{\boldsymbol{\theta}}_i^{(s)}$ is an $(N-1)$-component row vector with s non-zero entries that correspond to the s factorial effects largest in magnitude for $s = 0, \ldots, N-2$. Note that we do not consider the case when all factorial effects take non zero values.

Recall that for the ith experimental unit, the experimenter observes only one potential outcome Y_i^{obs}. As exemplified in Table 2.1, let $\boldsymbol{Y}_i^{\text{mis}}$ denote the $(N-1)$-component row vector of the missing potential outcomes. Note that the rows of the $N \times K$ submatrix formed by columns 2 to $K+1$ of $\boldsymbol{G}$ represent the N treatment combinations, which is usually referred to as the design matrix (i.e., the column $\boldsymbol{w}$ in Table 2.1). Denote by $\boldsymbol{g}_i^{\prime\text{obs}}$ the column of $\boldsymbol{G}'$ that contains the treatment combination $\boldsymbol{z}$ assigned to unit i, and let the submatrix formed by the remaining $N-1$ columns be $\boldsymbol{G}_i^{\prime\text{mis}}$. Then from (1.2) we can write

$$(Y_i^{\text{obs}}, \boldsymbol{Y}_i^{\text{mis}}) = \left(\theta_{i,0}, \frac{\dot{\boldsymbol{\theta}}_i^{(s)}}{2}\right) (\boldsymbol{g}_i^{\prime\text{obs}} : \boldsymbol{G}_i^{\prime\text{mis}}) \tag{2.1}$$

Imputation of the vector of missing potential outcomes $\boldsymbol{Y}_i^{\text{mis}}$ under H_{0s} requires two simple steps:

1. Estimate θ_{i0} as $\hat{\theta}_{i0} = Y_i^{\text{obs}} - (1/2)\dot{\boldsymbol{\theta}}_i^{(s)}\tilde{\boldsymbol{g}}_i^{\prime\text{obs}}$, where $\tilde{\boldsymbol{g}}_i^{\prime\text{obs}}$ is the column vector $\boldsymbol{g}_i^{\prime\text{obs}}$ without its first element (which is unity).

2. Impute the missing potential outcomes for the ith unit using

$$Y_i^{\text{mis}} = \left(\hat{\theta}_{i0}, \frac{\dot{\boldsymbol{\theta}}_i^{(s)}}{2}\right)\boldsymbol{G}_i^{\prime\text{mis}}.$$

Two important drawbacks of this imputation-based approach involving testing a sequence of sharp-null hypotheses are, (i) it assumes a constant additive effect and (ii) it ignores uncertainty of the estimate. Rubin (1984) provided the following Bayesian justification of the Fisherian approach to inference: *it gives the posterior predictive distribution of the estimand of interest under a model of constant treatment effects and fixed units with fixed responses.* Thus, a natural extension of the Fisherian approach is the Bayesian inferential procedure described in the following section.

2.2 Comparison of Loughin & Noble and Single Imputation

The L&N approach does not use the usual randomization based p value to calculate the probability of observing something as extreme or more extreme than what was observed. They correct the p-value because the obtained randomization distribution using all factorial effects is an approximation of the one they actually want: the distribution of all the effects

that are null, completely excluding those that are assumed active. Therefore their target distribution does not have the effects that were previously tested. However, the actual distribution they get does include them as noise because the residuals only ensure that the average effect is zero. We do not believe this is appropriate for the randomization tests in the potential outcomes framework, hence we use the usual randomization based p value (i.e., the direct p value obtained from comparing the observed value to the re-randomization distribution).

Given all of the differences mentioned above, even though this method came about by trying to understand Loughin and Noble (1997) in the potential outcomes framework, they lead to different results. The following theorem together with the observation that for the Single Imputation method the bound can be exceeded show that these two methods can in fact lead to different results.

Theorem 2.2.1. *For a 2^k unreplicated factorial experiment there are at most*

$$\frac{\binom{2^k}{2^{k-1}}}{2} - 2^k + 2$$

distinct values that the maximum absolute effect, $|\hat{\theta}|_{(2^k-1)}$, can take across randomizations.

Proof:

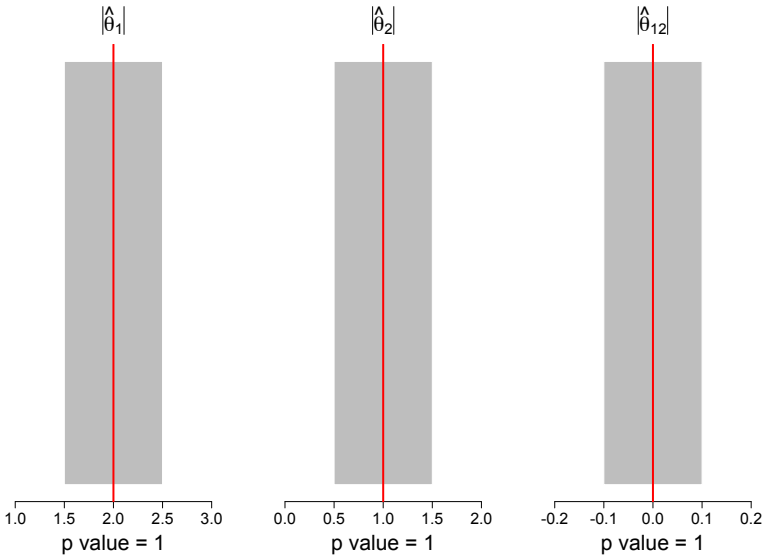
$\hat{\theta}_1$ is the difference in means between two distinct halves of the finite population in the experiment. Assuming each partition leads to a different value, then there are $\binom{2^k}{2^{k-1}}$ distinct values $\hat{\theta}_1$ can take. Now, because of the symmetry of the $\hat{\theta}_1$ we know that $|\hat{\theta}_1|$ can only take $N = \binom{2^k}{2^{k-1}}/2$ values. Let $\mathscr{S}$ denote the set of these values.

Due to the symmetry between factorial effect estimators, for a given randomization the observed absolute factorial effects are a sample of $2^k - 1$ values of the $\mathscr{S}$ set. Let $\mathscr{S} = \{|\hat{\theta}|_i : |\hat{\theta}|_{(1)} < |\hat{\theta}|_{(2)} < \cdots < |\hat{\theta}|_{(N)}\}$.

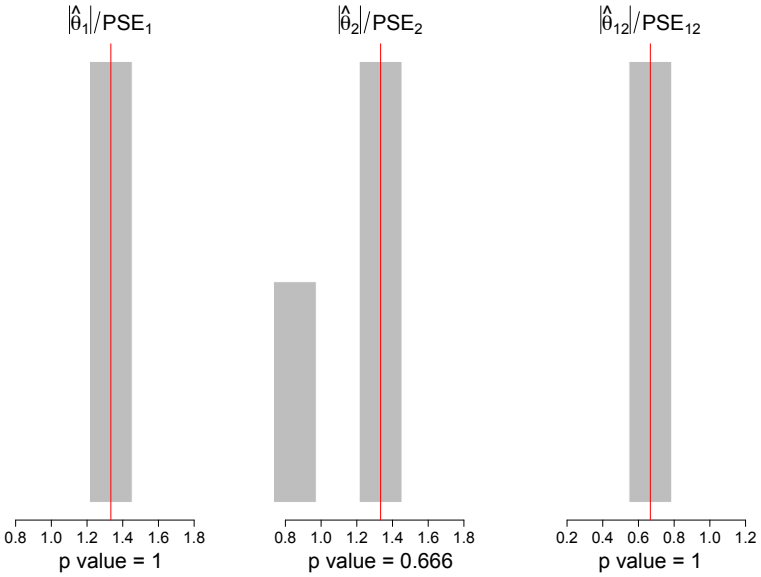
Hence, the maximum in this sample will be equal to $|\hat{\theta}|_{(N+1-j)}$ for $\binom{N-j}{2^k-2}$ randomizations, for $j = 1, 2, \ldots, N - 2^k + 2$. □

This theorem is relevant for every step in the Loughin and Noble procedure, because every step consists of a permutation test all 2^K values used to calculate the value of the test statistic (i.e., the maximum absolute value of *all* factorial effects). The difference between steps is the 2^K residuals in $\tilde{\mathbf{y}}$ that are permuted. However, this theorem only applies to the first step in the Single Imputation approach.

For $\mathbf{Y}^{obs} = (1, 2, 3, 4)'$, the comparison of results obtained using the Loughin & Noble and Sequential Single Imputation methods is displayed in Figure 2.1. Note that the the first step is equivalent in both sequential procedures for the 2^2 case because PSE_1 is the same for every randomization, therefore the result in the Theorem above follows for both procedures. Now, as in the L&N approach, for any k, the smallest absolute effect cannot be tested. In the Single Imputation approach, this happens because the p value obtained will be 1 due to the fact that all randomizations will lead to the same value of the test statistic, $T_{L,i} = 2/3$ since $PSE_{2^k-1} = 1.5 * |\hat{\theta}|_{(2^k-1)}$. Therefore in this simple case, the only step where we can see the differences is the second one.



(a) Loughin and Noble



(b) Single Imputation

Figure 2.1: Results of Two Sequential Testing Procedures for $\mathbf{Y}^{obs} = (1, 2, 3, 4)'$

Chapter 3

The Unreplicated Case: Sequential Posterior Predictive Checks

In this chapter we propose a Bayesian approach that is different from both Box and Meyer (1986) and Chipman et al. (1997), which have been proposed to screen for active effects.

3.1 A Bayesian approach to screening factorial effects using Sequential Posterior Predictive Checks

We now propose a Bayesian approach for screening active factorial effects from an unreplicated 2^K design that overcomes the drawbacks of the Fisherian and Loughin and Noble (1997) approaches and makes the inferential procedure more flexible. Our method extends the Bayesian framework proposed by Dasgupta et al. (2012) to a sequential screening procedure using posterior predictive checks (PPC) proposed by Rubin (1984) and investigated

by Meng (1994), Gelman et al. (1996) and, from a randomization perspective, Rubin (1998). The use of PPC is motivated by an additional intuitive appeal: in hypothesis testing, it is common practice to stop once we have a p-value that is "small enough". Ironically, we stop when we find a model that does not fit (the null model). We believe that a better, cleaner and more principled strategy is one that stops when we find a model that *does* fit the data. Rubin (1984), wrote "*Although the frequentist can stop with a rejection of the null hypothesis, I believe that the Bayesian is obliged to seek and build a model that is acceptable to condition on*". The proposed Bayesian procedure does make use of distributional assumptions in contrast to the Fisherian approach, making it more general than the randomization-based approach, which can be derived as a special case of the former by putting point-mass prior distributions on the potential outcomes and the unit-level factorial effects.

The key steps in the proposed approach are: (i) postulating a suitable "null model" (a probabilistic model specifying the active effects) for the potential outcomes; (ii) obtaining an imputation model for the missing potential outcomes $\boldsymbol{Y}^{\text{mis}}$, conditional on the observed outcomes $\boldsymbol{Y}^{\text{obs}}$ and the observed assignment vector $\boldsymbol{W}$; and (iii) using the imputation model to obtain the posterior predictive distribution of a suitable test statistic (or discrepancy measure) T, and consequently to compute the posterior predictive p value. For example, for a one-sided assessment of certain test statistics, the posterior predictive p value, $Pr(T > T^{obs})$, is the posterior probability of observing a value of T at least as large as the value observed, T^{obs}, given the observed data, the model being assumed, and the assignment mechanism.

To implement the sequential posterior predictive check (S-PPC), the aforementioned three steps are iterated either through a "step-down" or a "step-up" approach. The former approach starts with the sharp null model of no active effect, i.e., Fisher's sharp null hy-

pothesis of no treatment effects and then creates a sequence of non-sharp null models by including effects to the set of postulated active effects one by one, starting with the largest estimated effect. We stop as soon as we find the most parsimonious model that is consistent with the data. In contrast, the latter procedure starts with the saturated model (all effects active) as a default, and then tests whether more parsimonious models are consistent with the data by eliminating the factorial effects one by one, starting with the smallest estimated one. We then stop when we find a model that is inconsistent with the data, keeping the last one that seemed adequate. In the following three subsections, we describe this procedure in detail, starting with the definition of "active" effects.

3.1.1 Definition of inactive effects and their related notation

As before, we assume that the N experimental units are *fixed*. Under the potential outcomes perspective there are different definitions of *active* effects that we could use (see page 17). We now give a sharp definition of *activeness* of an effect as follows. However, at the end of this section we mention different ways to relax the definition, although we do not explore these any further in this paper. For $j = 1, \ldots, N-1$, we call the jth factorial effect (indexed by the vector $\boldsymbol{g}_j$) *inactive* if $\theta_{ij} = 0$ for all $i = 1, \ldots, N$ and *active* otherwise. Thus, a factorial effect is active if it is non-zero for at least one unit of the finite population. Let $\mathcal{A}$ denote the set of active effects with cardinality a where $0 \leq a \leq N-1$. Note that for the individual level definition, $a = 0$ is equivalent to the sharp null hypothesis of no treatment effects. Also, let $\mathcal{I}$ denote the set of effects that are "inactive", having cardinality $N-1-a$. It is therefore possible to partition the unit-level vector of factorial effects $\boldsymbol{\theta}_i$ as $(\boldsymbol{\theta}_i^{\mathcal{A}} : \mathbf{0})$, where $\boldsymbol{\theta}_i^{\mathcal{A}}$ is an $(a+1)$-component row vector that includes the mean term and $\mathbf{0}$ is a null

vector with $N-1-a$ components. Therefore $(\boldsymbol{\theta}_i^{\mathcal{A}} : \mathbf{0})$ is a permutation of $\boldsymbol{\theta}_i$. Each row of $\boldsymbol{G}'$ represents a factorial effect, hence $\boldsymbol{G}'$ can also be rearranged to form the matrix

$$\begin{pmatrix} \boldsymbol{G}'^{\mathcal{A}} \\ \boldsymbol{G}'^{\mathcal{I}} \end{pmatrix},$$

so that $\boldsymbol{G}'^{\mathcal{A}}$ and $\boldsymbol{G}'^{\mathcal{I}}$ are matrices of order $(a+1) \times N$ and $(N-a-1) \times N$ corresponding to the active and inactive effects respectively.

Consequently, it follows from (1.2) that

$$\boldsymbol{Y}_i = (\boldsymbol{\theta}_i^{\mathcal{A}} : \mathbf{0}) \begin{pmatrix} \boldsymbol{G}'^{\mathcal{A}} \\ \boldsymbol{G}'^{\mathcal{I}} \end{pmatrix} = \boldsymbol{\theta}_i^{\mathcal{A}} \boldsymbol{G}'^{\mathcal{A}}. \tag{3.1}$$

To express the observed and missing outcomes in terms of the active effects, we partition the vector $\boldsymbol{g}_i'^{\text{obs}}$ and the matrix $\boldsymbol{G}_i'^{\text{mis}}$, defined in Section 2, as

$$\begin{pmatrix} \boldsymbol{g}_i'^{\text{obs},\mathcal{A}} \\ \boldsymbol{g}_i'^{\text{obs},\mathcal{I}} \end{pmatrix} \text{ and } \begin{pmatrix} \boldsymbol{G}_i'^{\text{mis},\mathcal{A}} \\ \boldsymbol{G}_i'^{\text{mis},\mathcal{I}} \end{pmatrix},$$

respectively, just as we partitioned $\boldsymbol{G}'$. Therefore, from (2.1) we can write

$$Y_i^{\text{obs}} = \boldsymbol{\theta}_i \boldsymbol{g}_i'^{\text{obs}} = (\boldsymbol{\theta}_i^{\mathcal{A}} : \mathbf{0}) \begin{pmatrix} \boldsymbol{g}_i'^{\text{obs},\mathcal{A}} \\ \boldsymbol{g}_i'^{\text{obs},\mathcal{I}} \end{pmatrix} = \boldsymbol{\theta}_i^{\mathcal{A}} \boldsymbol{g}_i'^{\text{obs},\mathcal{A}}, \tag{3.2}$$

$$\boldsymbol{Y}_i^{\text{mis}} = \boldsymbol{\theta}_i \boldsymbol{G}_i'^{\text{mis}} = (\boldsymbol{\theta}_i^{\mathcal{A}} : \mathbf{0}) \begin{pmatrix} \boldsymbol{G}_i'^{\text{mis},\mathcal{A}} \\ \boldsymbol{G}_i'^{\text{mis},\mathcal{I}} \end{pmatrix} = \boldsymbol{\theta}_i^{\mathcal{A}} \boldsymbol{G}_i'^{\text{mis},\mathcal{A}}. \tag{3.3}$$

Note that the potential outcomes framework also permits us to define active/inactive effects in terms of the finite population factorial effects and super-population factorial effects. For example, although we do not do this here, the jth factorial effect could be called inactive at the finite population level if $\bar{\theta}_{\cdot j} = 0$, and at the super population level if $\mu_j = 0$, where θ_{ij} is assumed to be, say, $N(\mu_j, \sigma^2)$ for $i = 1, \ldots, N$. In Section 5, we show that the current approach (see page 16) is robust enough to perform well in super-population settings, where inferences are more uncertain.

3.1.2 The imputation model and computation of the posterior predictive distribution of T

Throughout this section, we assume that the units are fixed, but the potential outcomes can be random. Also, we assume that the potential outcomes are conditionally independent across units, given the values of the hyperparameters. Whereas this assumption is not necessary for the developments that follow, it is a reasonable assumption in many realistic situations.

Let $p(\boldsymbol{\theta}_i|\eta^{\mathcal{A}})$ denote a null probabilistic model for $\boldsymbol{\theta}_i$, where $\boldsymbol{\eta}^{\mathcal{A}}$ is a vector of parameters having a suitable prior distribution for the set of active effects $\mathcal{A}$. Because of the identity $\boldsymbol{Y}_i = \boldsymbol{\theta}_i \boldsymbol{G}$, the model can also be specified through $\boldsymbol{Y}_i$. Then, following Rubin (1978), for a completely randomized factorial experiment, we have the following lemma

Lemma 1. *The conditional distribution of* $\mathbf{Y}_i^{\text{mis}}$ *given* $\boldsymbol{Y}^{\text{obs}}$ *and* $\boldsymbol{W}$ *is given by*

$$p(\boldsymbol{Y}_i^{\text{mis}}|\boldsymbol{Y}^{\text{obs}}, \boldsymbol{W}) \propto \int\int p(Y_i^{\text{mis}}|\boldsymbol{\eta}^{\mathcal{A}}, \boldsymbol{\theta}_i, \boldsymbol{Y}^{\text{obs}}, \boldsymbol{W}) p(\boldsymbol{\theta}_i|\boldsymbol{\eta}^{\mathcal{A}}, \boldsymbol{Y}^{\text{obs}}) p(\boldsymbol{\eta}^{\mathcal{A}}|\boldsymbol{Y}^{\text{obs}}) d\boldsymbol{\eta}^{\mathcal{A}} d\boldsymbol{\theta}_i \tag{3.4}$$

Then, obtaining the posterior predictive p-value for the null model involves the following

steps:

1. Obtain the imputation model $p(\boldsymbol{Y}_i^{\text{mis}}|\boldsymbol{Y}^{\text{obs}}, \boldsymbol{W})$ using Lemma 1.

2. Repeat the following steps M times:

   (a) Impute a draw of the missing potential outcomes using the model obtained in step 1.

   (b) Re-randomize the N units to the N treatment combinations (i.e., generate a draw from $\boldsymbol{W}$) and generate a new set of observed data.

   (c) Compute the discrepancy measure T^{rep}, given this specific imputation and re-randomization.

3. Compute the posterior predictive p-value for the null model as the proportion of cases (out of M) in which T^{rep} equals or exceeds T^{obs}.

In principle, the steps described above can be carried out for any model. However, here we restrict ourselves to a simple normal hierarchical model described below, where δ denotes an indicator function such that $\delta(A) = 1$ if A is true and zero otherwise.

$$\begin{aligned}
p(\boldsymbol{Y}_i|\boldsymbol{\theta}_i) &= \delta(\boldsymbol{Y}_i = \boldsymbol{\theta}_i^{\mathcal{A}}\boldsymbol{G}'^{\mathcal{A}}), \quad i = 1, \ldots, N, \\
p(\boldsymbol{\theta}_i^{\mathcal{A}}|\mu^{\mathcal{A}}, \sigma^2) &= N(\mu^{\mathcal{A}}, \sigma^2\boldsymbol{I}_a), \quad i = 1, \ldots, N, \\
p(\boldsymbol{\theta}_i^{\mathcal{I}}) &= \delta(\boldsymbol{\theta}_i^{\mathcal{I}} = \boldsymbol{0}), \quad i = 1, \ldots, N, \\
p(\boldsymbol{\mu}^{\mathcal{A}}, \sigma^2) &\propto \frac{1}{\sigma^2}.
\end{aligned} \tag{3.5}$$

Also, we assume each unit's $\boldsymbol{\theta}_i$ is independent of the other units conditional on the hyperparameters, and hence that of $\boldsymbol{Y}_i$'s for $i = 1, \ldots, N$. The hierarchical normal model specified

by (3.5) permits a fair comparison of the proposed approach with the standard approaches. Such a comparative study is conducted in Section 5.

We now discuss how Steps 1 and 2, described earlier, can be implemented under model (3.5). Substituting $(\boldsymbol{\mu}^{\mathcal{A}}, \sigma^2)$ for $\boldsymbol{\eta}^{\mathcal{A}}$ in (3.4), and after some minor manipulations, we obtain the imputation model for missing outcomes as:

$$\int\int\int p(Y_i^{\text{mis}}|\boldsymbol{\mu}^{\mathcal{A}}, \sigma^2, \boldsymbol{\theta}_i, \boldsymbol{Y}^{\text{obs}}, \boldsymbol{W}) p(\boldsymbol{\theta}_i|\boldsymbol{\mu}^{\mathcal{A}}, \sigma^2, \boldsymbol{Y}^{\text{obs}}) p(\boldsymbol{\mu}^{\mathcal{A}}|\sigma^2, \boldsymbol{Y}^{\text{obs}}) p(\sigma^2|\boldsymbol{Y}^{\text{obs}}) d\boldsymbol{\mu}^{\mathcal{A}} d\sigma^2 d\boldsymbol{\theta}_i, \tag{3.6}$$

where

$$\begin{aligned} p(\sigma^2|\boldsymbol{Y}^{obs}, \mathcal{A}) &= \text{Inv}\boldsymbol{\chi}^2(N-a-1, MS_{res}^{\mathcal{A}}), \text{ where } MS_{res}^{\mathcal{A}} \text{ is the} \\ &\quad \text{residual mean square of the model that} \\ &\quad \text{corresponds to the regression of } \boldsymbol{Y}^{obs} \text{ on } \boldsymbol{G}'^{\mathcal{A}}, \\ p(\boldsymbol{\mu}^{\mathcal{A}}|\sigma^2, \boldsymbol{Y}^{obs}) &= N(\hat{\boldsymbol{\theta}}^{\mathcal{A}}, \sigma^2 \mathbf{I}_a/N), \text{ where } \hat{\boldsymbol{\theta}}^{\mathcal{A}} \text{ is the posterior} \\ &\quad \text{mean of } \boldsymbol{\mu}^{\mathcal{A}} \text{ (and } \boldsymbol{\theta}_i^{\mathcal{A}} \text{), which is also} \\ &\quad \text{its unbiased OLS estimator.} \end{aligned} \tag{3.7}$$

For the unreplicated case, $MS_{res}^{\mathcal{A}}$ is well defined except for the saturated model. The imputation of missing potential outcomes, i.e., step 2(a), can be executed using the following simulation procedure based on (3.6) and (3.7):

Step 1. Draw σ_*^2 from $p(\sigma^2|\boldsymbol{Y}^{obs})$.

Step 2. With σ^2 set to σ_*^2, draw $\boldsymbol{\mu}_*^{\mathcal{A}}$ from $p(\boldsymbol{\mu}^{\mathcal{A}}|\sigma_*^2, \boldsymbol{Y}^{obs})$.

Step 3. For each unit,

(a) With $\sigma^2 = \sigma_*^2$ and $\boldsymbol{\mu}^{\mathcal{A}} = \boldsymbol{\mu}_*^{\mathcal{A}}$, draw $\dot{\boldsymbol{\theta}}_{i,*}^{\mathcal{A}}$ from $N(\boldsymbol{\mu}_*^{\mathcal{A}}, \sigma_*^2 \mathbb{I})$.

(b) Complete $\boldsymbol{\theta}_{i,*}^{\mathcal{A}}$ by calculating $\theta_{i,0} = Y_i^{obs} - \dot{\boldsymbol{\theta}}_{i,*} \dot{\boldsymbol{g}}_i^{'\text{obs}}/2$, where $\dot{\boldsymbol{g}}_i^{'\text{obs}}$ excludes the first entry of $\boldsymbol{g}_i^{'\text{obs}}$.

(c) Fill in missing potential outcomes for unit i as $\boldsymbol{Y}_{i,*}^{\text{mis}} = \boldsymbol{\theta}_{i,*}^{\mathcal{A}} \boldsymbol{G}_i^{'\text{mis},\mathcal{A}}$. draw $\boldsymbol{\theta}_{i,*}$ from $p(\boldsymbol{\theta}_i | \boldsymbol{\mu}_*^{\mathcal{A}}, \sigma_*^2)$.

More precisely,

- For every $\theta_{i,j} \in \mathcal{I}$, the posterior distribution equals the prior distribution (i.e., point mass at zero):

$$p(\theta_{i,j} | \boldsymbol{\mu}, \sigma) = \delta(\theta_{i,j} = 0).$$

- For the active set of θ_{ij}, set $\sigma^2 = \sigma_*^2$ and $\boldsymbol{\mu}^{\mathcal{A}} = \boldsymbol{\mu}_*^{\mathcal{A}}$ to draw $\boldsymbol{\theta}_{i,*}^{\mathcal{A}}$ from the normal distribution $N(\boldsymbol{\mu}_*^{\mathcal{A}}, \sigma_*^2)$.
- The missing potential outcomes for this unit are filled in as $\boldsymbol{Y}_{i,*}^{\text{mis}} = \boldsymbol{\theta}_{i,*}^{\mathcal{A}} \boldsymbol{G}_i^{'\text{mis},\mathcal{A}}$.

3.1.3 Discrepancy measures and definitions of extremeness

As in most statistical inference problems, the choice of discrepancy measure T and definition of extremeness (e.g., $P(T \geq T^{\text{obs}})$) are fundamental, and should be determined before the experiment is performed. In unreplicated experiments, zero degrees of freedom are available to estimate the standard deviation of the estimated factorial effects $\hat{\theta}_j$ in the usual way, i.e., using the residual mean squared error. In this context, Lenth (1989) proposed the *pseudo standard error* as an estimate of the standard deviation of the $\hat{\theta}_j$'s. This estimate was presented in Equation 4.1 in Chapter 4.

Here we consider three discrepancy measures. In the proposed sequential procedure, at

each step we compare the observed value of T to its posterior predictive distribution, where a specific set of effects are assumed active,$\mathcal{A}$, and the rest are assumed inactive, $\mathcal{I}$. Let $PSE_{\mathcal{I}}$ denote the pseudo standard error of the factorial effects in $\mathcal{I}$. The three discrepancy measures we study are the maximum absolute value of the effects in $\mathcal{I}$ in the current prior distribution with and without standardizing by the pseudo standard error, i.e., $\max_{j\in\mathcal{I}}\left|\frac{\hat{\theta}_j}{PSE_{\mathcal{I}}}\right|$ and $\max_{j\in\mathcal{I}}\left|\hat{\theta}_j\right|$, and the $PSE_{\mathcal{I}}$.

In particular, the use of the order statistics in the Loughin and Noble method motivated the inclusion of discrepancy measures involving $\max_{j\in\mathcal{I}}\left|\hat{\theta}_j\right|$ in our study. Extremeness for these statistics is defined by the right tail of the distribution of T because, given the assumed prior distribution is correct, it is unlikely that a null effect would be as large or larger than the observe value, T^{obs}.

On the other hand, the interpretation of $PSE_{\mathcal{I}}$ as a measure of what is not explained by the model motivated its use as a discrepancy measure here. Thus, a small value of the $PSE_{\mathcal{I}}$ indicates that the assumed model is not consistent with the data because there is a smaller variability between the factorial effects assumed null than what would be expected under the assumed prior model. Hence, extremeness for this discrepancy measure is based on a one sided assessment focusing on the lower tail of the posterior predictive distribution of T.

Chapter 4

The Unreplicated Case: Brief overview of related existing methods

In this chapter we review a subset of the existing methods that are relevant to our work. In 1959, Daniel proposed a graphical method to visually screen for active effects in two-level designs. He wrote "Plotting the empirical cumulative distribution of the usual set of orthogonal contrasts computed from a 2^k experiment on a special grid may aid in its criticism and interpretation.[...] The half-normal plot can be used to estimate the error standard deviation and to make judgements about the reality of the observed effects." Hamada and Balakrishnan (1998) mention that Daniel (1959) did provide an objective method that for the most part has been ignored. It has been the subjective assessment of the plot that has withstood the test of time, continuing to make his half normal probability plot a standard approach in the screening of non replicate factorial experiments. A figure from the original paper is replicated in Figure 4.1. The idea behind it is that many effects are inactive and plotting them this way will suggest an inactive distribution (variance) and will highlight

those effects that don't follow this trend to identify them as active.

Lenth (1989) gave a formalization of Daniel's halfnormal probability plots, making it a popular analytical tool for screening in unreplicated experiments. Ye et al. (2001) and Tripolski et al. (2008) gave extensions of Lenth's method; the former from a sequential perspective and the latter motivated by the need to control the false discovery rate (FDR). Finally, we review two existing Bayesian methods because our proposal follows a Bayesian perspective. However, unlike the rest of the presented methods, the Bayesian ones were not included in our simulation study because of the reported sensitivity to the specification of the hyperparameters in the prior distributions.

4.1 Lenth Method (1989)

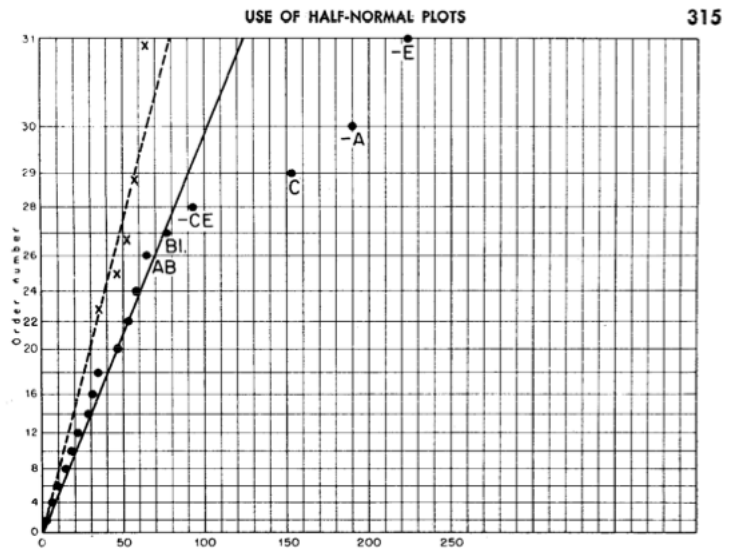
Lenth (1989) defined an estimator for the the standard error of the factorial effects, which he called the *pseudo standard error* (PSE),

$$PSE = 1.5 \cdot \text{median}_{|\hat{\theta}_j| \leq 2.5 s_0} |\hat{\theta}_j| \tag{4.1}$$

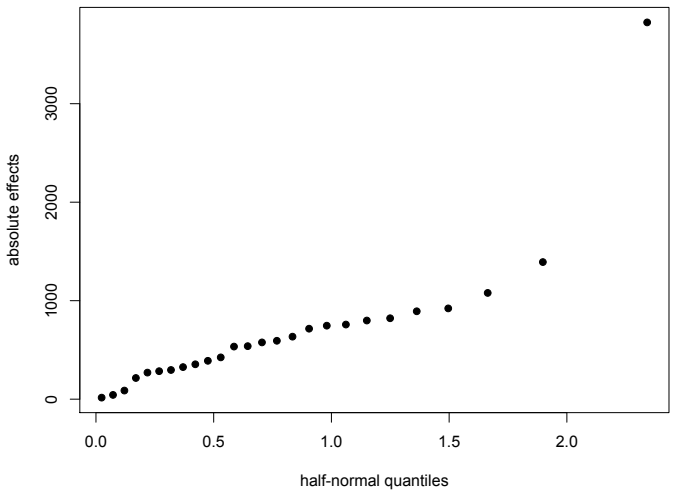
where $s_0 = 1.5 \cdot \text{median}|\hat{\theta}_j|$, and $\hat{\theta}_j$ is the estimate of the j-th factorial effect. He then proposed the calculation of the statistic

$$T_{L,j} = \frac{\hat{\theta}_j}{\text{PSE}}$$

to test the effects. He proposed to use a t distribution with $I/3$ degrees of freedom for controlling the individual error rate (IER), where I corresponds to the number of mutually



(a) Screenshot of Daniel (1959) original paper proposing the use of half-normal probability plots to screen for active effects in two level factorial designs.



(b) A modern halfnormal probability plot. The y-axis corresponds to the absolute value of the factorial effects and the x-axis corresponds to the halfnormal quantiles. Here, the subjectiveness of the results is clear. This plot was produced by pretending the seat belt example (see section 7.1.3) in Chapter 7 was unreplicated, and only using the first value for each treatment combination to screen for active effects.

Figure 4.1: Halfnormal probabiltiy plots

orthogonal estimated factorial effects (e.g. 2^K-1 in a 2^K full factorial design). He proposed to set the critical value to $(1-(1-\alpha)^{1/I})$ to control for the experiment-wise error rate (EER). Loughin (1998) and Ye and Hamada (2000) derived calibrated critical values for Lenth's method in terms of $|T_{L,j}|$. In this paper we use the critical value that Ye, Hamada and Wu (2001) reported, which was based on Ye and Hamada (2000). In the Hamada and Balakrishnan study, Lenth's method was shown to be, not only simple, but one of the most powerful approaches. Multiple testing, or equivalently the experiment-wise error rate (EER) or false discovery rate (FDR), is an issue in this context. There are two ways in which this is handled. The most common one is to use a modified critical value. The Bonferroni correction is an example of this type of solution, in which an EER of α is obtained by using the critical value found by modifying the individual significance levels to α/n, where n is the number of tests performed. Therefore, in a 2-level factorial experiment with k factors $n = 2^K - 1$. Daniel's and Lenth's methods are of the this type. Zahn (1975) was the first to propose sequential procedures to control the EER. In the following sections common sequential procedures for controlling the EER are reviewed. The total number of treatment combinations 2^K will henceforth be replaced by m for notational convenience.

4.2 Step Down Lenth Method (2001)

The method proposed by Ye et al. (2001) is a sequential version of Lenth's approach that uses the order statistics of the factorial effects, $|\hat{\theta}|_{(j)}$. The method proceeds as follows, let $m = 2^K$. Then, at step s calculate the test statistics $T_{L,s} = \frac{|\hat{\theta}|_{(m-s)}}{\text{PSE}_s}$, where PSE_s is the PSE of $\hat{\theta}_{(1)}, \hat{\theta}_{(2)}, \ldots, \hat{\theta}_{(m-s)}$. Let C_α^j denote the EER critical value at the α significance level of the original Lenth method with j contrasts. If $T_{L,j} > C_\alpha^j$ for all $j > m - s$ then the largest s

factorial effects are declared active. The simulation study in Ye et al. (2001) shows that it is slightly more powerful than Lenth's method and has a closer value to the nominal level of the EER.

4.3 FDR corrected Lenth Method (2008)

Tripolski et al. (2008) propose a method that targets controlling the False Discovery Rate (FDR), instead of the Experiment-wise Error Rate. As the name suggests, the FDR refers to the proportion of true null effects that are declared active. The authors, using the Benjamini and Hochberg (1995) controlling procedure, propose adaptive and non-adaptive versions of Lenth (1989) and Dong (1993) methodologies to screen for active effects. The authors proposed regular and adaptive versions of their methods, but did not find a significant improvement in the use of the adaptive ones. In this paper, the *the FDR procedure* refers to the non-adaptive proposal that is based on Lenth's method. All estimated absolute effects are standardized by the pseudo standard error, and then converted to "raw" p values using a t distribution with $I/3$ degrees of freedom, were I is the number of effects being tested. Then, these p values are arranged in ascending order, and effects for which the p value is smaller than the largest p value that upholds $p_{(i)} \leq iq/I$ are identified as active. In Tripolski et al. (2008) the original t-distribution proposal was used, instead of the Ye and Hamada (2000) calibrated version that was mentioned above and that was implemented for Lenth's method in our simulation studies. Therefore, for the FDR procedure we use the t distribution that Tripolski et al. (2008) used in their implementation of their method. Hence, the two Lenth methods are not identical, but both are calibrated. Relevant to this, Tripolski et al. (2008) note that their calibration leads to similar results to those reported by Ye and Hamada

(2000).

4.4 Loughin and Noble (1997) Permutation Approach

To our knowledge, this method is the main randomization-based approach for sequential testing. It was proposed in the context of unreplicated factorial experiments. The key idea behind it is to sequentially redefine the response variable used for the permutation test by eliminating the orthogonal projection of the $\boldsymbol{Y^{obs}}$ vector on to the space generated by the largest factorial effects. An outline of the method is:

1. Compute $\hat{\theta}$ from $\mathbf{y} = \mathbf{Y}^{obs}$ and order the effects and columns of $\boldsymbol{G}$ to correspond to $|\hat{\theta}|_{(m)} \geq |\hat{\theta}|_{(m-1)} \geq \cdots \geq |\hat{\theta}|_1$.

2. At step s, set $\hat{W}_s = |\hat{\theta}|_{(s)}$ and obtain

$$\tilde{\mathbf{y}}_s = \mathbf{y} - \hat{\theta}_{(m)}\boldsymbol{g}_{(m)} - \cdots - \hat{\theta}_{(m-s+1)}\boldsymbol{g}_{(m-s+1)}; \quad (\tilde{\mathbf{y}}_1 = \mathbf{y}).$$

3. Do a permutation test on $\tilde{\mathbf{y}}_s$

   - Repeat **nsim** (large) times:

     (a) Permute $\tilde{\mathbf{y}}_s$

     (b) Calculate the discrepancy measure $W_s^* = \left(\frac{m-1}{m-s}\right)^{1/2} |\hat{\tilde{\theta}}^*|_{(1)}$.

   - Compute the p value as $P_s = 1 - \left[\frac{\#W_s^* < \hat{W}_s}{\mathbf{nsim}}\right]^{\frac{m-s}{m-1}}$.

4. Repeat steps 2 and 3 for as many effects as desired.

They propose a *step up* procedure, starting with the smallest absolute effect. This approach is does not impose either the sparsity or the normality assumptions. However, this procedure is unable to identify as active any effects with magnitude equal to that of the smallest effect.

4.5 Two Bayesian Options

These methods distinguish between active and inactive effects through different variances. There are two main ones, which we now describe.

4.5.1 Box and Meyer (1986)

To our knowledge, this is the first Bayesian method proposed to screen for active effects. The active effects have a $N(0, \sigma^2_{active})$ distribution and inactive effects have a $N(0, \sigma^2_{inactive})$ distribution, where $\sigma_{inactive} << \sigma_{active} = K\sigma_{inactive}$. Hence, the prior distribution of factorial effects is a mixture of the active and inactive normal distributions (centered at zero), where the proportion of active effects (α_{active}) is small, assuming effect sparsity. There are many hyperparameters that need to be set (also referred to as tuning parameters) for which the authors recommend $\alpha_{active} = 0.2$, $K = 10$ and marginal posterior probability threshold to identify as active of 0.5. It was the only Bayesian method compared in Hamada and Balakrishnan (1998), where it performed best for small number of active effects but not that much better than Lenth's method. It performed poorly for large number of active effects (with a considerable underperformance relative to Lenth's method). Therefore, on average, it was found to be less powerful than Lenth's method.

4.5.2 Chipman et al. (1997)

This method was proposed to deal with complex aliasing structures. Therefore, we review it in more detail in Chapter 7 where we believe it is more relevant because we discuss fractional factorial designs. However, a short description is relevant at this point because of the Bayesian motivation of the method proposed. A key idea of this procedure (originally proposed in Chipman (1996) as a Bayesian variable selection procedure) was to include a vector of latent variables, $\boldsymbol{\delta}$, indicating the activeness for each effect, such that conditional on it, the effects have normal distributions. That is, the conditional prior distribution of the factorial effects is

$$p(\boldsymbol{\theta}_j) = \begin{cases} N(0, \sigma^2\tau_j^2) & \text{if } \delta_j = 0, \\ N(0, \sigma^2(c_j\tau_j)^2) & \text{if } \delta_j = 1. \end{cases}$$

An inverse gamma prior is given for σ^2. Two main options are given for the priors of $\boldsymbol{\delta}$. The first option assumes independence between these latent variables, such that $p(\boldsymbol{\delta}) = \prod_{j=1}^{p+1} p_j^{\delta_j}(1-p_j)^{1-\delta_j}$, independence prior on $\boldsymbol{\delta}$, where p_j is the probability that $\delta_j = 1$, looks like this where p_j is the prior probability that $\delta_j = 1$. The second option is to use hierarchical priors that incorporate design of experiments principles. For example, factor the prior distribution of $\boldsymbol{\delta}$, $p(\boldsymbol{\delta})$, using the effect heredity principle into

$$p(\boldsymbol{\delta}) = p(\delta_1)p(\delta_2)p(\delta_3|\delta_1, \delta_2),$$

and model the probability that $\delta_3 = 1$ as being higher the more parent factorial effects are active. This is further explained in Chapter 7.

The hyperparameters of the inverse gamma, in addition to c_j and τ_j for every j, are tuned

to give reasonable posterior probabilities to the models. The recommendation is, following Box and Meyer (1986) $c_j = 10$, high sensitivity to τ_j

4.6 Why are these relevant?

All of the non Bayesian methods presented in this review were implemented in our simulation study for the unreplicated case for comparison with our proposals. We also implemented the permutation test based on Fisher's sharp null hypothesis of absolutely no treatment effect, which we explain in detail in Chapter 2. To our knowledge, Loughin and Noble (1997) is the standard permutation based method. Hence, including it is not only natural but necessary, and even more so because it was not included in the extensive comparative study performed in Hamada and Balakrishnan (1998). All the other non Bayesian methods are standard in the literature and serve as reference points.

The relevance of the Bayesian procedures stems from the fact that our main proposal has a Bayesian motivation and uses that framework. However, these Bayesian options were not included in the simulation study. The reasons behind excluding these are that 1) Box and Meyer (1986) did not outperform Lenth (1989) in comparative study carried out in Hamada and Balakrishnan (1998), and 2) both Bayesian procedures have hyperparameters that need to be specified. Chipman et al. (1997) point out that their results are sensitive to the specification of τ_j, so implementing this method needs some thought about each of the situations being tested (i.e., subject matter expertise). It is unclear how to implement this approach, taking into account this sensitivity to prior specification, in simulation studies like those we pursue. These "tuning" parameters are not something our current Bayesian proposal requires.

Chapter 5

The Unreplicated Case: Simulation study

In this chapter we present results from a simulation study comparing the Single Imputation method presented in Chapter 2 and the Sequential Posterior Predictive Checks of 3.1 to the methods reviewed in 4.

The general framework presented in Section 3.1 offers flexibility for estimands, discrepancy measures and sequences of models to be assessed other than those explored here, which may be more relevant to the study at hand. However, we believe it is fundamental to compare its performance relative to standard approaches in the literature to evaluate what benefits our approach can have, even in the usual setting. As already mentioned in 3.1.2, the specific hierarchical normal model given by (3.5) was proposed with this goal in mind. To agree with the traditional setting, the simulation is done using the superpopulation definition of an active effect, which in our procedure can be understood as $\mu^{\mathcal{I}} = \mathbf{0}$ and $\mu^{\mathcal{A}}$ consists of non-zero elements. Nevertheless, as explained in Section 3.1.2, the S-PPC does fill in the

missing data, assuming $\theta_{ij} = 0$ for effects in $\mathcal{I}$ and all units. We believe this simulation allows for a fair comparison of the frequency properties of the different methods.

5.1 Calibration Study - Under the Null Hypothesis

Tukey (1953) coined the term *experimentwise error rate* (EER) to refer to the probability of making one or more false discoveries when testing multiple factors. It is standard practice in the literature to calibrate proposed methods to thresholds that are most frequently used in screening. We calibrate our proposals to satisfy a 0.05 experimentwise error rate to be able to compare them to other methods in the literature, using the cutoff values reported in their papers for an EER of 0.05. Two other error rates that are commonly used in the literature are compared here but not calibrated. One is the *individual error rate* (IER), which is the probability of incorrectly identifying a null effect as active, but does not account for the other effects being assumed null. The other one is the *false discovery rate* (FDR), which corresponds to the expected proportion of false positives among all the effects declared active. For the null model ($\mathcal{A} = \emptyset$), the FDR is equivalent to the EER.

The calibration study was performed under the null hypothesis of all individual level factorial effects being inactive in a 2^4 full factorial design (i.e. $\mu_j = 0$ for all $i = 1, \ldots, 16$ and $j > 1, \ldots, 15$, where $\mu_j = E(\theta_{ij})$. We calibrated the sequential posterior predictive checks, using the three discrepancy measures described in Section 3.1.3 and the different screening rules (i.e., "step-down" or "step-up"). The procedure we followed in the simulations is described below, and satisfies the classic assumptions of the existing methods.

We simulated 1000 different science tables assuming the null hypothesis. Each science

table was obtained by simulating θ_{ij}'s from a $N(0,1)$ distribution, and transforming them to the potential outcomes $Y_i(\boldsymbol{z})$. For each method, the following two summary measures were recorded.

(i) The proportion of null effects incorrectly declared significant.

(ii) The indicator of whether there was at least one null effect incorrectly declared significant for this data set.

The IER and EER were estimated by averaging summary measures (i) and (ii) respectively over the 1000 different simulations.

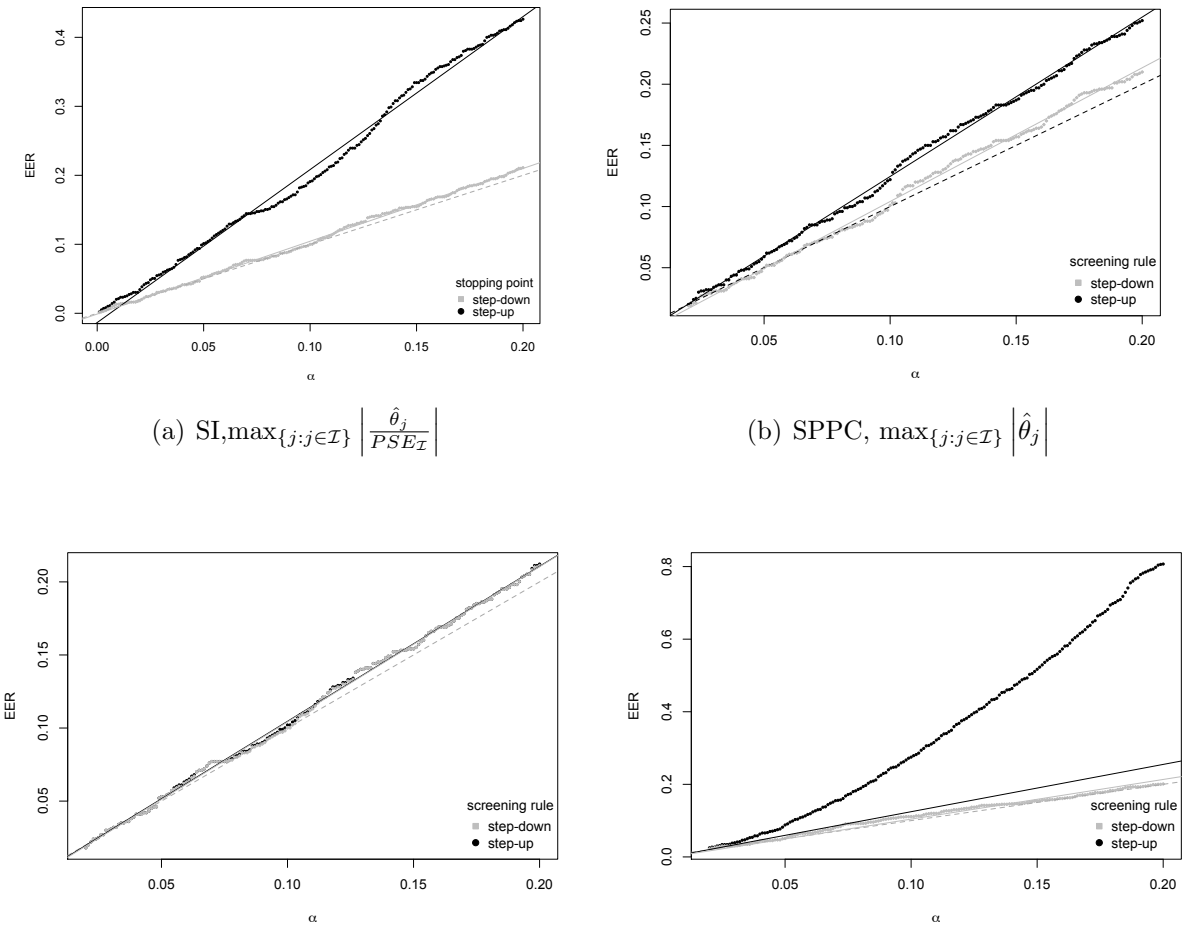
The calibration results for the Single Imputation and **S-PPC** are shown in Figure 5.1. The points correspond to the EER observed using different cut off values (α). The greyscale denotes the two screening rules. The dashed line represents the identity line, and each continuous line denote the OLS fit of the observed EER on the α level used for the corresponding screening rule. Each subfigure corresponds to a different discrepancy measure. For every selection of a cut off we looked at the α value that lead to an observed EER closest to and below 0.05. For each method, discrepancy measure and screening rule combination, we corroborated the selection of α by fitting a linear model of the observed EER values to the different α cutoffs.

The calibration study results of the Single Imputation approach are displayed below, in Figure 5.1(a). The test statistic used for this approach is $\max_{\{j:j\in\mathcal{I}\}}\left|\frac{\hat{\theta}_j}{PSE_s}\right|$. For the step-down approach the OLS fit is $EER = -0.013 + 2.21\alpha$. However the the observed values suggest a cut off of 0.026, this is the value we used. For the step-up approach both the observed values and the OLS (i.e., $EER = -0.001 + 1.055\alpha$) agree on a cut off of 0.048.

The rest of the subfigures in Figure 5.1 correspond to the calibration results of the **S-PPC**. As shown in Figure 5.1(b), to achieve an EER of 0.05 with the $\max_{\{j:j\in\mathcal{I}\}}\left|\hat{\theta}_j\right|$ discrepancy measure, a 0.043 cut off (α) should be used for each test to identify the active effects when using the step-up procedure. This 0.043 value corresponds to the observed EER of 0.049 and a predicted 0.05 for the OLS fit. The linear regression of the observed EER on α resulted in this expression $EER = -0.006 + 1.303\alpha$. Furthermore, when the step-down procedure is used with this discrepancy measure, then a 0.05 cut off should be used. Again, this value agrees with both the observed simulation results and the OLS fit, which is $EER = -0.005 + 1.093\alpha$.

Figure 5.1(c) shows that the use of the $\max_{\{j:j\in\mathcal{I}\}}\left|\frac{\hat{\theta}_j}{PSE_{\mathcal{I}}}\right|$ discrepancy measure appears to eliminate the difference in the results between the explored screening rules, which might be an attractive feature. In this case, the OLS fits for both screening rules are practically identical. The OLS results are $EER = -0.002 + 1.06\alpha$ and $EER = -0.001 + 1.06\alpha$ for the step-down and step-up procedures, respectively. Observed values from the simulation, with both screening rules and the step-up OLS fit, suggest the 0.048 value as cut off (the first one below 0.05). The OLS fit for the step-down procedure suggests a 0.049 cut off. We chose to use 0.048. Although it is not shown here, the screening rules do lead to different IER values for this discrepancy measure

Figure 5.1(d) displays the results for the $PSE_{\mathcal{I}}$ as a discrepancy measure. Clearly, the step-up procedure is not viable for this statistic. Henceforth we will only use the step-down procedure for this measure, for which the OLS fit is $EER = 0.008 + 0.988\alpha$. The cut off that the OLS fit suggests (0.043) is not the same as that suggested by the observed rates in the simulation, which is 0.049. In the comparative study we use the latter.



(a) SI,$\max_{\{j:j\in\mathcal{I}\}}\left|\frac{\hat{\theta}_j}{PSE_{\mathcal{I}}}\right|$

(b) SPPC, $\max_{\{j:j\in\mathcal{I}\}}\left|\hat{\theta}_j\right|$

(c) SPPC, $\max_{\{j:j\in\mathcal{I}\}}\left|\frac{\hat{\theta}_j}{PSE_{\mathcal{I}}}\right|$

(d) SPPC, $PSE_{\mathcal{I}}$

Figure 5.1: Scatter plot of the Calibration Study for the Single Imputation and for the Sequential Posterior Predictive Checks. The greyscale denotes the two screening rules. The dashed line represents the identity line, and each continuous line denotes the OLS fit of the observed EER on the α level used for the corresponding screening rule. Each subfigure corresponds to a different discrepancy measure. In the label of each subfigure the first element corresponds to the method and the second one to the discrepancy measure used.

Table 5.1: Cutoffs obtained in the calibration of the Single Imputation and Sequential Posterior Predictive Checks methods using 2^4 factorial designs for the different methods, discrepancy measures and screening rules proposed.

Method	**Discrepancy Measure**	**Screening Rule**	α
SI	$\max_{\{j:j\in\mathcal{I}\}} \left\lvert \frac{\hat{\theta}_j}{SE_{\mathcal{I}}} \right\rvert$	step-down	0.048
		step-up	0.026
S-PPC	$\max_{\{j:j\in\mathcal{I}\}} \left\lvert \hat{\theta}_j \right\rvert$	step-down	0.050
		step-up	0.043
	$\max_{\{j:j\in\mathcal{I}\}} \left\lvert \frac{\hat{\theta}_j}{PSE_{\mathcal{I}}} \right\rvert$	both	0.048
	PSE_s	step-down	0.049

5.2 Simulation across Alternative Hypotheses

A similar simulation was used to compare the performance of the proposed S-PPC to the Permutation Test (also called the Fisher's sharp Null approach) as described in Chapter 2 with and without the Bonferroni correction; the Loughin and Noble (1997) approach (L&N); the Lenth (1989) method; the Step Down Lenth method, as described in Ye et al. (2001); and the FDR corrected Lenth method described in Tripolski et al. (2008). The methods of Box and Meyer (1986) and Chipman et al. (1997) are not included in this study; we excluded the former because it performs poorly for large numbers of active effects, and, although it performed well for low numbers of active effects, on average, it is less powerful than Lenth's method (Hamada and Balakrishnan, 1998); we excluded the latter because of the sensitivity of the results to the choice of some hyperparameter values (Wu and Hamada, 2009).

We simulated potential outcomes repeatedly (1000 times) from alternatives defined by setting $\max_{\{j:j\in\mathcal{A}\}} |\mu_j|$ to 4, and different levels of noise ($\sigma = 0.5, 1, 2$), number of active effects ($a = 1, 2, 4, 6$), and the range, ρ between active effects (defined as $\rho = \max_{\{j:j\in\mathcal{A}\}} |\mu_j| - \min_{\{j:j\in\mathcal{A}\}} |\mu_j| = 1, 2, 3$). Each combination of these simulation factors determines a true

value of $\boldsymbol{\mu}$ by selecting a factorial effects at random to be active, and letting the corresponding μ_j's assume the value $4-\rho$ if $a=1$, and the values $4-\rho\left(1-\frac{t}{a-1}\right)$ for $t=0,\ldots,a-1$ if $a>1$. The remaining μ_j's are set at zero. Finally, the potential outcomes for each unit are drawn from $\boldsymbol{Y}_i \sim N(\boldsymbol{\mu}\boldsymbol{G}', \sigma^2\mathbb{I})$.

We compare these methods using five summary measurement based on averages across 1000 simulated data sets: **IER** (average proportion of false positives), **EER** (proportion of data sets with at least one false positive), **FDR** (average proportion of false positives among all the effects declared active), **RR** (average proportion of true positives), and **ANP** (average number of positives effects declared active). In this context, positives are the factorial effects declared active, false positives are inactive effects incorrectly identified as active, and true positives are active effects that are correctly identified.

The results of the simulation study are shown in Table 5.2. Although the expressions for the test statistic for the step-down Lenth and FDR-corrected Lenth methods in Table 5.2 appear to be different from those in Ye et al. (2001) and from those in Tripolski et al. (2008), respectively, they are essentially the same. The values obtained in the calibration study in Section 5.1 are used as threshold p-values in the proposed S-PPC methods. For the existing methods, we used the values reported in the literature, which were also determined via calibration studies. We include the results for the permutation test, which does not account for multiple comparisons, as a reference point where it achieves an IER of 0.05 but a very high EER of 0.661. The Step Down Lenth method has a very high EER under the null hypothesis (0.109). The levels for the remainder of the methods are reasonably close to the intended 0.05.

Table 5.2 shows that the sequential posterior predictive checks *step-up* procedure using

the maximum of the absolute effects assumed null as the discrepancy measure has the best performance. It has an EER below 0.049 under the null hypothesis, as well as low IER and ANP values relative to the other methods. Averaging across the alternative hypotheses, it has the highest RR, achieving 0.637, with the closest competitors achieving 0.560 (L&N, Step Down Lenth and Lenth-with 0.551), the highest ANP (2.146) and all the error rates are below 0.05. This last statement is true for the S-PPC variations, unlike the Step Down Lenth (EER = 0.095, FDR = 0.052) and the FDR corrected Lenth (EER = 0.082). The Single Imputation approach using the step-down procedure performs better than using the step-up procedure. This is probably due to the large difference between the cutoffs of the two screening rules obtained from the calibration.

Both step-up and step-down procedures for the Single Imputation approach are dominated by the Loughin and Noble (1997) method.

In Figure 5.2 we display the RR, FDR, and EER, across the different simulation alternative hypotheses settings for the previously established methods and compare them to the S-PPC with the best performance. This figure clearly shows that the step-up S-PPC with discrepancy measure $\max_{\{j:j\in\mathcal{I}\}}\left|\hat{\theta}_j\right|$, the best S-PPC (BS-PPC), is much better than the rest with clear and distinct modes around 1 for the RR, and 0 for the FDR and EER. For our proposal, there are no combinations with an FDR above 0.05. However, there are eight combinations with EER above 0.05. In contrast, the L&N and Lenth methods have none above this threshold, at the cost of lower rejection rates. In contrast, the Lenth method with the FDR correction leads to EER values above 0.05 for 30 of the 36 settings for the alternative hypotheses, which is not surprising because the goal is to control the FDR. Nevertheless, their method has 5 combinations with FDR above 0.05.

Table 5.2: Summary of results of the simulation study. Average rates and number of effects declared active, as well as the standard errors of these quantities are displayed for all methods and screening rules under the null ($\sigma = 1$) and across alternative hypotheses.

			null hypothesis			**average across alternative hypotheses**				
Method	**Discrepancy Measure**	**Screening Rule**	**IER** (SE)	**EER** (SE)	**ANP** (SE)	**RR** (SE)	**IER** (SE)	**EER** (SE)	**FDR** (SE)	**ANP** (SE)
Permutation	$\lvert\hat{\theta}\rvert_j$	-	0.050 (0.001)	0.660 (0.015)	0.750 (0.019)	0.487 (0.006)	0.004 (<0.001)	0.049 (0.006)	0.042 (0.005)	1.200 (0.014)
Bonferroni	$\lvert\hat{\theta}\rvert_j$	-	0.003 (<0.001)	0.048 (0.007)	0.048 (0.007)	0.251 (0.006)	0.000 (<0.001)	0.002 (0.001)	0.002 (0.001)	0.400 (0.009)
Lenth	$\lvert\hat{\theta}_j\rvert/PSE$	-	0.007 (0.001)	0.057 (0.007)	0.099 (0.016)	0.551 (0.008)	0.004 (0.001)	0.031 (0.005)	0.014 (0.003)	1.862 (0.030)
FDR corrected Lenth	$\lvert\hat{\theta}_j\rvert/PSE$	step-up	0.011 (0.002)	0.059 (0.007)	0.168 (0.024)	0.403 (0.008)	0.012 (0.001)	0.082 (0.009)	0.033 (0.004)	1.778 (0.037)
Step Down Lenth	$\max_{\{j:j\in\mathcal{I}\}}\left\lvert\frac{\hat{\theta}_j}{PSE_{\mathcal{I}}}\right\rvert$	step-down	0.015 (0.002)	0.110 (0.010)	0.228 (0.026)	0.560 (0.008)	0.016 (0.002)	0.095 (0.009)	0.052 (0.006)	2.050 (0.037)
L&N	$\left(\frac{N-1}{N-\lvert\mathcal{I}\rvert}\right)^{1/2}\max_{\{j:j\in\mathcal{I}\}}\left\lvert\frac{\hat{\theta}_j}{PSE_{\mathcal{I}}}\right\rvert$	step-up	0.006 (0.001)	0.053 (0.007)	0.086 (0.015)	0.560 (0.009)	0.002 (0.001)	0.023 (0.005)	0.011 (0.002)	1.744 (0.034)
Single Imputation	$\max_{\{j:j\in\mathcal{I}\}}\left\lvert\frac{\hat{\theta}_j}{PSE_{\mathcal{I}}}\right\rvert$	step-down	0.007 (0.001)	0.052 (0.007)	0.051 (0.017)	0.538 (0.009)	0.005 (0.001)	0.033 (0.006)	0.013 (0.003)	1.839 (0.033)
		step-up	0.016 (0.003)	0.051 (0.007)	0.236 (0.042)	0.480 (0.009)	0.014 (0.002)	0.043 (0.006)	0.024 (0.004)	1.780 (0.047)
S-PPC	$PSE_{\mathcal{I}}$	step-down	0.008 (0.002)	0.053 (0.007)	0.122 (0.029)	0.436 (0.008)	0.013 (0.003)	0.029 (0.005)	0.018 (0.003)	1.664 (0.046)
	$\max_{\{j:j\in\mathcal{I}\}}\left\lvert\hat{\theta}_j\right\rvert$	step-down	0.003 (< 0.001)	0.050 (0.007)	0.050 (0.007)	0.327 (0.008)	0.001 (0.003)	0.009 (0.005)	0.005 (0.003)	0.594 (0.046)
		step-up	0.004 (0.001)	0.049 (0.007)	0.058 (0.009)	0.637 (0.009)	0.003 (0.001)	0.028 (0.005)	0.012 (0.002)	2.146 (0.035)
	$\max_{\{j:j\in\mathcal{I}\}}\left\lvert\frac{\hat{\theta}_j}{PSE_{\mathcal{I}}}\right\rvert$	step-down	0.006 (0.001)	0.050 (0.007)	0.086 (0.015)	0.527 (0.008)	0.003 (0.001)	0.026 (0.005)	0.012 (0.003)	1.768 (0.029)
		step-up	0.006 (0.001)	0.049 (0.007)	0.085 (0.015)	0.531 (0.008)	0.003 (0.001)	0.026 (0.005)	0.012 (0.003)	1.788 (0.029)

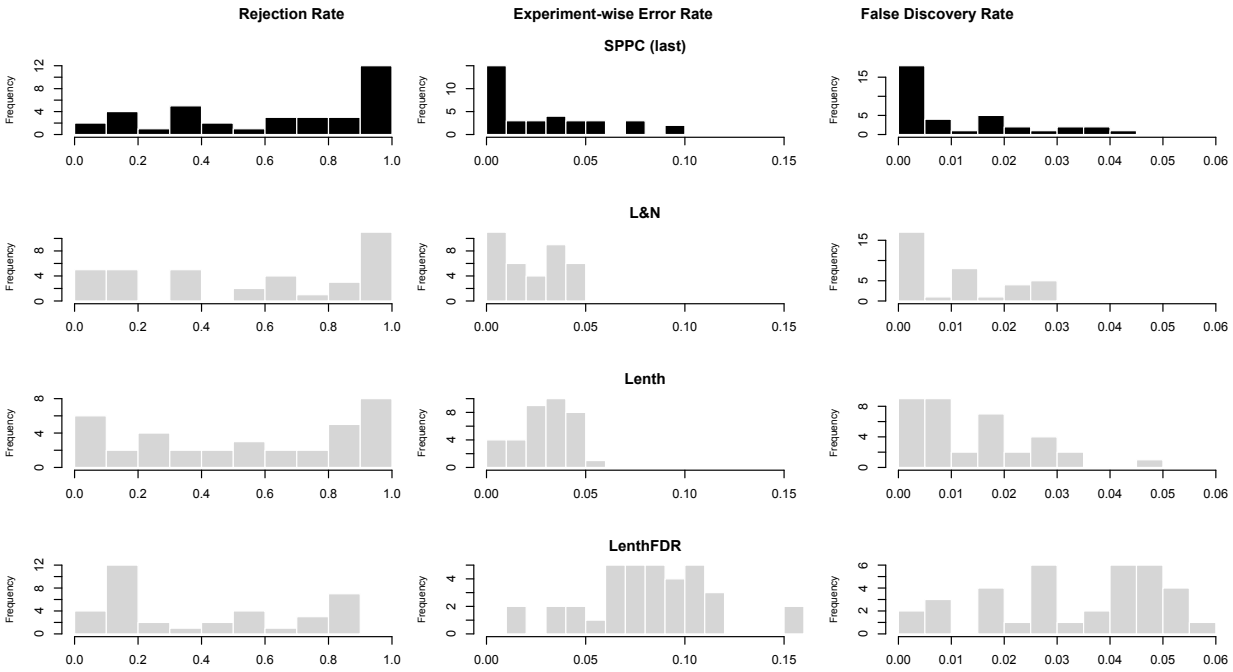


Figure 5.2: Distributions of the Rejection (RR), the False Discovery (FDR), and Experiment-wise Error (EER) Rates across the different alternative hypotheses.

Figures 5.3, 5.4, and 5.5 display the marginal effects of (i) the number of active effects (ii) the standard deviation σ and (iii) the range r of mean active effects respectively, on average RR, EER and FDR. Figure 5.3 shows the average number of effects declared active for each of the true numbers of active effects in the set of alternative hypotheses explored (i.e., $1, 2, 4, 6$). Again, the step-up S-PPC approach with $\max_{\{j:j\in\mathcal{I}\}} \left|\hat{\theta}_j\right|$ as the discrepancy measure is the one with best overall performance. This figure agrees with the finding of Tripolski et al. (2008), that the L&N approach has good performance for a low number of active effects (1,2,4), but that its performance dramatically falls for higher numbers of active effects (this fall reportedly happened for 4 in their simulations). In contrast, the Tripolski et al. (2008) method performs best for higher number of true active effects. Their results are comparable to the Step Down Lenth approach and slightly worse than the S-PPC ones.

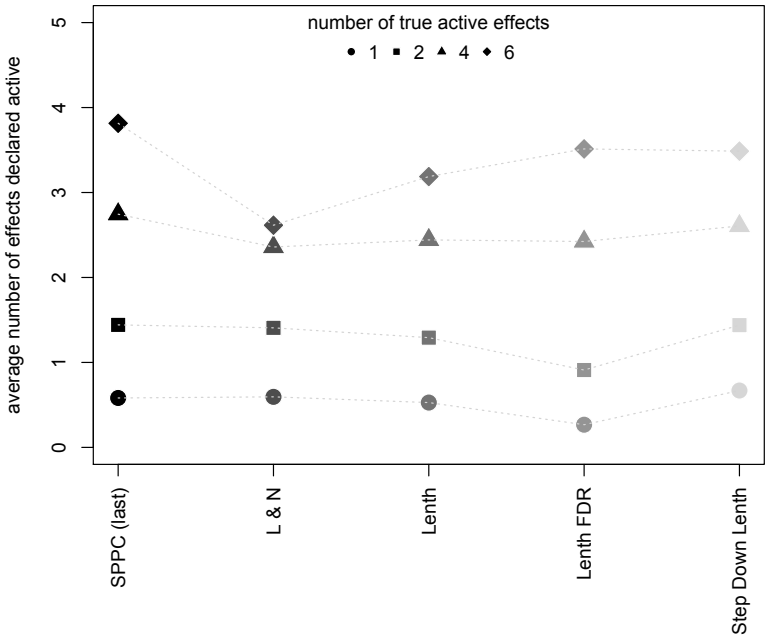


Figure 5.3: Average number of effects declared active for the different values of *true* active effects = 1,2,4,6.

As shown in Figure 5.4, the Lenth FDR controls the FDR better, not surprisingly because it is designed to do so. The false discovery rates across combinations are below, but closer to the 0.05 threshold, and only for this method the average FDR decreases as σ increases. Figure 5.4 reveals that σ has a bigger impact on the BS-PPC than for any other method because of the higher variance it shows for different values of σ. The BS-PPC improvement on the RR relative to the other methods increases with σ.

Figure 5.5 shows that the BS-PPC is less sensitive to variation in ρ than to variation in σ. Nevertheless, the BS-PPC is preferred because its RR is higher for all levels, and its error rates are below 0.05.

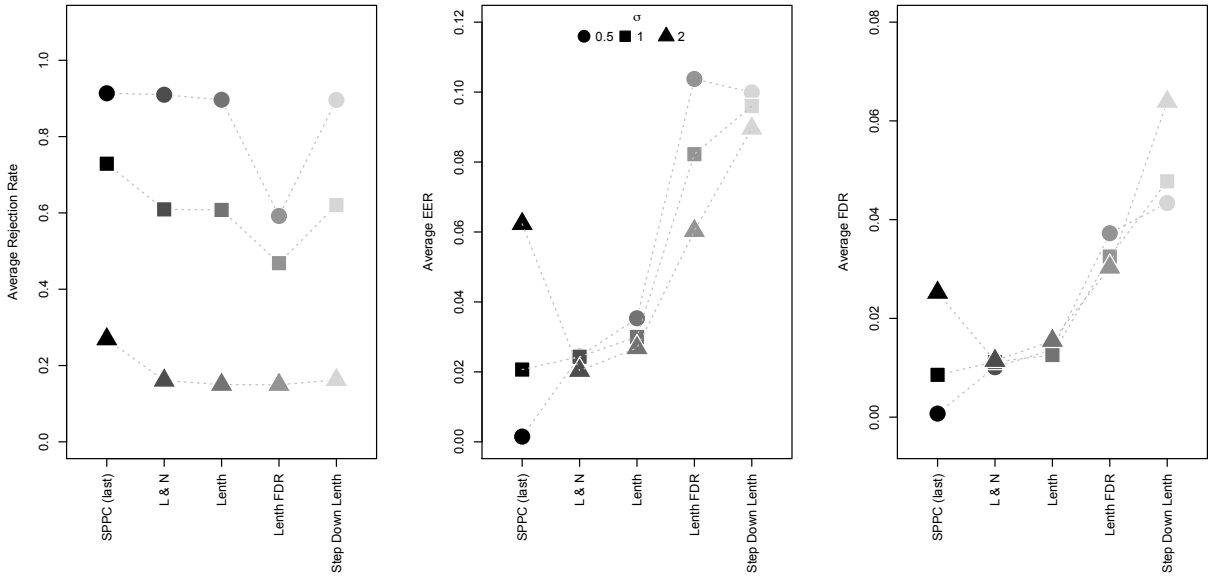


Figure 5.4: Comparison of RR, EER, and FDR across simulations settings with the same value of $\sigma = 0.5, 1, 2$. The standard errors of these values range from 0.005 to 0.011 for the RR, from 0.001 to 0.010 for the EER, and from 0.001 to 0.007 for the FDR.

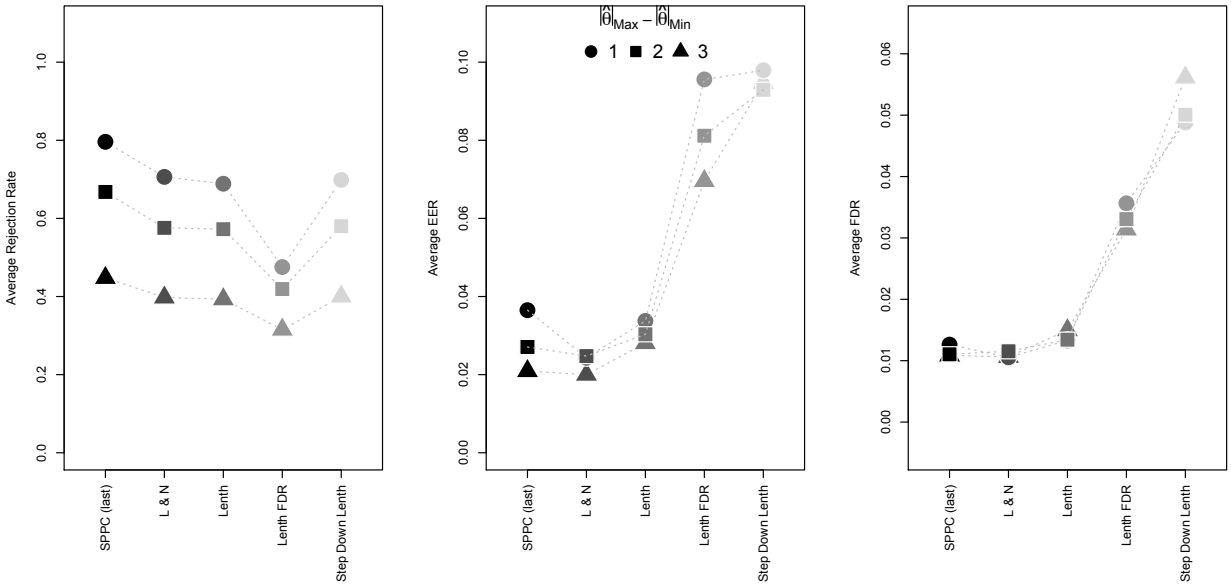


Figure 5.5: Comparison of RR, EER, and FDR across simulations settings with the same range between the absolute values of the factorial effects, $r = 1, 2, 3$. The standard errors of these values range from 0.008 to 0.009 for the RR, from 0.004 to 0.009 for the EER, and from 0.002 to 0.006 for the FDR.

5.3 Additional Graphs of Simulation Results

Figure 5.6 displays the distributions of average rejection, false discovery and experiment-wise error rates for all explored existing methods and optimal versions of our proposals. Note that due to the range of values taken for different methods, this plot does not have the same scales for FDR and EER.

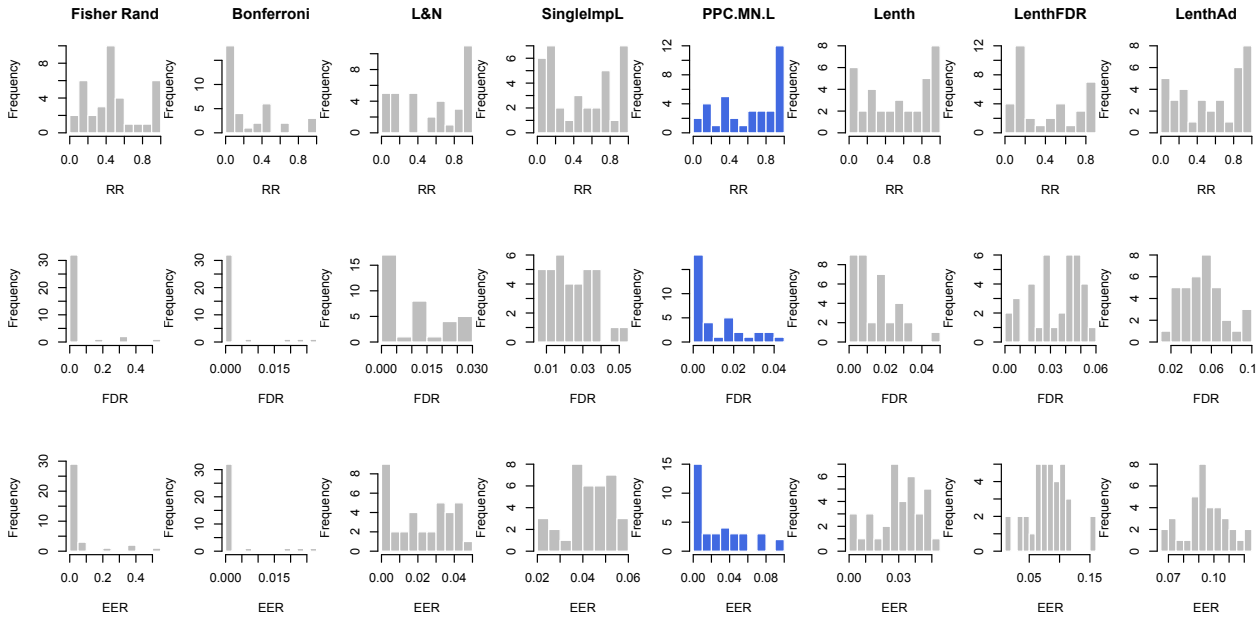


Figure 5.6: Histograms showing the Rejection (RR), the False Discovery (FDR), and Experiment-wise Error Rates (EER), across the different simulation settings. Improve this

Chapter 6

The Replicated Case

In this chapter we explore the replicate case of the two randomization based approaches proposed for the unreplicated case: Single Imputation (**SI**) and Sequential Posterior Predictive Checks (**S-PPC**). The extensions of these methods to the replicated case is straight forward, only slight modifications have to be made.

The exploration of the replicate case is relevant because of its broader applicability, for example in social sciences, when the assumption of small variation between units relative to the magnitude of treatment effects, is not plausible. The unreplicated case is a much more challenging setting where the usual tools are not useful because of the inability to use the traditional estimates of variation. In the replicate case, linear regression is a viable analytical tool, although it does not directly account for multiple comparisons. Most of the ideas of the methods proposed in the previous chapter for the unreplicated case are still applicable for the replicated case. In this chapter, we highlight the few differences that do exist and perform an analogous comprehensive simulation study.

We go over an example, that has been slightly modified from it's original form to be a balanced design, in the simple 2^2 full factorial case. While doing that, we will review and exemplify some concepts introduced in the previous chapter, as well as highlight the differences between the unreplicated and the replicated cases. We continue assuming fixed units, SUTVA and balanced designs. As in the previous chapter, let N denote the total number of experimental units and K the number of treatment factors. Now, let r denote the number of replicates such that $N = 2^K r$.

6.1 An Example: effect of exercise and androgenic steroids on strength.

An experiment was designed to test whether the use of androgenic steroids increases strength. Fourty-four men, between 19 and 40 years, all experienced in weight lifting and weighing between 90 to 115 percent of their ideal weight were recruited through advertisements.[1] There are two treatment factors, exercise and testosterone use. Both factors have two levels: on or off. The subjects were randomly assigned to one of four groups: placebo with no exercise, testosterone with no exercise, placebo plus exercise, testosterone plus no exercise. Strength was measured as the difference of one-repetition maximal weight lifted before and after a 10 week treatment period.

The first step of an experiment is to have a goal in mind. In this case the goal is to *assess the effect of the use of androgenic steroids on strength compared to the effect of exercise.* This goal guides the choice of estimands. That is, the quantities that are functions of the

[1]The actual experiment can be found in Bhasin et al. (1996). There were more participants and due to compliance and scheduling issues seven dropped out.

individual responses and are related to the goal of the experiment. The estimands that are most commonly used are linear combinations of averages and for illustration purposes those will be used. Nevertheless, the *potential outcome* approach allows the use of other functions that are usually overlooked but could be of more interest, such as quantiles.

This is a 2^2 full factorial design, where any subject can be assigned to any one of four treatment combinations. Let exercise be the first factor and steroids the second factor, with levels -1 and 1 denoting that the factor is off and on, respectively. If the i-th subject is assigned to no exercise (first factor) and steroids (second factor), the observed gain in strength for this individual is denoted by $Y_i(-1,1)$. Four *potential outcomes* correspond to this man, one for each treatment combination:

$$\mathbf{Y}_i = (Y_i(1,1), Y_i(1,-1), Y_i(-1,1), Y_i(-1,-1)).$$

The *SCIENCE*, the true table of potential outcomes, for the full experiment is shown in Table 6.1.

Table 6.1: The *science* for the full experiment with four treatment conditions and forty-four men.

Unit	Potential outcome for treatment combination				Unit-level factorial effects			
(i)	$(1,1)$	$(1,-1)$	$(-1,1)$	$(-1,-1)$	$\theta_{i,0}$	$\theta_{i,1}$	$\theta_{i,2}$	$\theta_{i,3}$
1	$Y_1(1,1)$	$Y_1(1,-1)$	$Y_1(-1,1)$	$Y_1(-1,-1)$	$\theta_{1,0}$	$\theta_{1,1}$	$\theta_{1,2}$	$\theta_{1,3}$
2	$Y_2(1,1)$	$Y_2(1,-1)$	$Y_2(-1,1)$	$Y_2(-1,-1)$	$\theta_{2,0}$	$\theta_{2,1}$	$\theta_{2,2}$	$\theta_{2,3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
43	$Y_{43}(1,1)$	$Y_{43}(1,-1)$	$Y_{43}(-1,1)$	$Y_{43}(-1,-1)$	$\theta_{43,0}$	$\theta_{43,1}$	$\theta_{43,2}$	$\theta_{43,3}$
44	$Y_{44}(1,1)$	$Y_{44}(1,-1)$	$Y_{44}(-1,1)$	$Y_{44}(-1,-1)$	$\theta_{44,0}$	$\theta_{44,1}$	$\theta_{44,2}$	$\theta_{44,3}$
Average	$\overline{Y}(1,1)$	$\overline{Y}(1,-1)$	$\overline{Y}(-1,1)$	$\overline{Y}(-1,-1)$	θ_0	θ_1	θ_2	θ_3

The unit level estimands and potential outcomes remain the same as in the unreplicated

case. That is, we have the four potential outcomes ($\boldsymbol{Y}_i$) and the 4-dimensional row vector $\boldsymbol{\theta}_i$ containing the unit level factorial effects, and the same one to one relationship between these:

$$\begin{aligned}
\theta_{i,0} &\equiv \frac{Y_i(1,1)+Y_i(1,-1)+Y_i(-1,1)+Y_i(-1,-1)}{4},\\
\theta_{i,1} &\equiv \frac{Y_i(1,1)+Y_i(1,-1)}{2}-\frac{Y_i(-1,1)+Y_i(-1,-1)}{2},\\
\theta_{i,2} &\equiv \frac{Y_i(1,1)+Y_i(-1,1)}{2}-\frac{Y_i(1,-1)+Y_i(-1,-1)}{2},\\
\theta_{i,3} &\equiv \frac{Y_i(1,1)+Y_i(-1,-1)}{2}-\frac{Y_i(1,-1)+Y_i(-1,1)}{2}.
\end{aligned}$$

represented in a summarized way by the $\boldsymbol{G}$ matrix defined in the previous chapter:

$$\boldsymbol{\theta}_i = \frac{1}{2^K}\boldsymbol{Y}_i\boldsymbol{G}$$

where

$$\mathbf{G} = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

is the *model matrix at the unit level*.

Similarly, the estimands at the *population level* are analogous to those of the unreplicated case: the population level main effects of each of the treatment factors and their interaction, which are denoted by θ_1, θ_2 and θ_3, respectively. That is, the population level vector of estimands is

$$\boldsymbol{\theta} \equiv (\theta_0, \theta_1/2, \theta_2/2, \theta_3/2).$$

Now, assuming complete randomization, the treatment assignment for the replicate case is a generalization of the one defined for the unreplicated case. Analogous to 1.1.1 in Chapter 1, we can define the assignment mechanism for a 2^K full factorial design with r replicates. Let $\boldsymbol{z}$, a K dimensional row vector with entries -1 and 1, denote a treatment combination, then the treatment assignment mechanism can be defined as

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

and it consists of randomly selecting r different units for every treatment combination, such that $Pr(W_i(\boldsymbol{z}) = 1) = r/N$ (implicitly conditional on the science). Also $\sum_i W_i(x) = 1$ for $i = 1, 2, \ldots, N$, and $\sum_{\boldsymbol{z}} W_i(\boldsymbol{z}) = r$ for all $\boldsymbol{z}$. As in the previous chapter, let

$$w_i = \sum_{\boldsymbol{z}} \boldsymbol{z} W_i(\boldsymbol{z})$$

be the treatment combination that the ith subject receives, let $\boldsymbol{W}$ be the generic treatment assignment vector of random variables, and let $\boldsymbol{w}$ be a specific realization of $\boldsymbol{W}$, i.e., a vector that contains all the individual treatment assignments *after* randomization. Hence, for a *given* treatment assignment the table of *observed* potential outcomes looks analogous to the one displayed in Table 6.2 where again the missing potential outcomes are represented by question marks.

The averages of the observed values within each column are *unbiased estimates*[2] of the averages of the columns of potential outcomes (see Tables 6.1 and 6.2). We can use con-

[2]The proof can be found in the appendix of Chapter 1 for a general randomization setting using symmetry arguments, and for the one used in fractional factorial designs in Dasgupta et al. (2012).

Table 6.2: *Observed Outcomes* for the full experiment with $\boldsymbol{w} = \big((1,1),(-1,-1),(-1,1),(1,-1)\big)'$.

Unit	Observed outcome for treatment combination				$\boldsymbol{w}$
(i)	$(1,1)$	$(1,-1)$	$(-1,1)$	$(-1,-1)$	
1	Y_1^{obs}	?	?	?	$(1,1)$
2	?	?	?	Y_2^{obs}	$(-1,-1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
43	?	?	Y_{43}^{obs}	?	$(-1,1)$
44	?	Y_{44}^{obs}	?	?	$(1,-1)$

trasts of these estimates of average potential outcomes of each treatment combination to get unbiased estimates of the factorial effects defined as contrasts of the treatment group means:

$$\begin{aligned}
\hat{\theta}_1 &\equiv \frac{\overline{Y}^{obs}(1,1)+\overline{Y}^{obs}(1,-1)}{2} - \frac{\overline{Y}^{obs}(-1,1)+\overline{Y}^{obs}(-1,-1)}{2}, \\
\hat{\theta}_2 &\equiv \frac{\overline{Y}^{obs}(1,1)+\overline{Y}^{obs}(-1,1)}{2} - \frac{\overline{Y}^{obs}(1,-1)+\overline{Y}^{obs}(-1,-1)}{2}, \\
\hat{\theta}_{12} &\equiv \frac{\overline{Y}^{obs}(1,1)+\overline{Y}^{obs}(-1,-1)}{2} - \frac{\overline{Y}^{obs}(1,-1)+\overline{Y}^{obs}(-1,1)}{2},
\end{aligned}$$

together with the overall average potential outcome,

$$\theta_0 \equiv \frac{\overline{Y}^{obs}(1,1)+\overline{Y}^{obs}(1,-1)+\overline{Y}^{obs}(-1,1)+\overline{Y}^{obs}(-1,-1)}{4},$$

to define the vector of estimands $\boldsymbol{\theta} = (\theta_0, \theta_1/2, \theta_2/2, \theta_3/2)$.

The model matrix for the *estimates* at the *population level*, $\mathbf{X}$, identifies the one to one relationship between the observed averages and the estimates of the factorial effects at the

population level

$$\boldsymbol{X} = \begin{pmatrix} \mathbf{G} \\ \vdots \\ \mathbf{G} \end{pmatrix}, \tag{6.1}$$

where, assuming a balanced design, $\boldsymbol{G}$ is repeated as many times as units in every treatment group. In the exercise and steroid experiment $r = 11$. A fundamental difference with the unreplicated case is that for the replicate case the potential outcomes approaches will require both $\boldsymbol{G}$ and $\boldsymbol{X}$ to run. The matrix $\boldsymbol{G}$ is required to fill in the missing outcomes at the unit level to run the final randomization-based assessment of each model, and the mattrix $\boldsymbol{X}$ is required to get the posterior estimates of the estimands of interest.

As noted in the previous chapter, for this definition of the estimands the corresponding estimates agree with those obtained by ordinary least squares in the classical set up for the additive linear model $\mathbf{Y}^{obs} = \mathbf{X}\boldsymbol{\theta} + \epsilon$, where $\mathbf{Y}^{obs}$ is the vector of *observed* outcomes for each unit. And

$$\hat{\boldsymbol{\theta}}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^{obs} = \frac{1}{r2^2}\mathbf{X}'\mathbf{Y}^{obs}.$$

6.2 Example: Results with Fisher Randomization Test

The randomization test remains practically the same for the replicate case relative to the unreplicated case, except that in this case we use the t like test statistic. In other words, we standardize by the estimate of the variance whereas in the unreplicated case we used the absolute factorial effects (although we could have used the PSE and increase the similarity between these two cases). The symmetry between the factorial effects is preserved.

The results for the steroid and exercise example using the Fisher randomization test are displayed in figure 6.1.

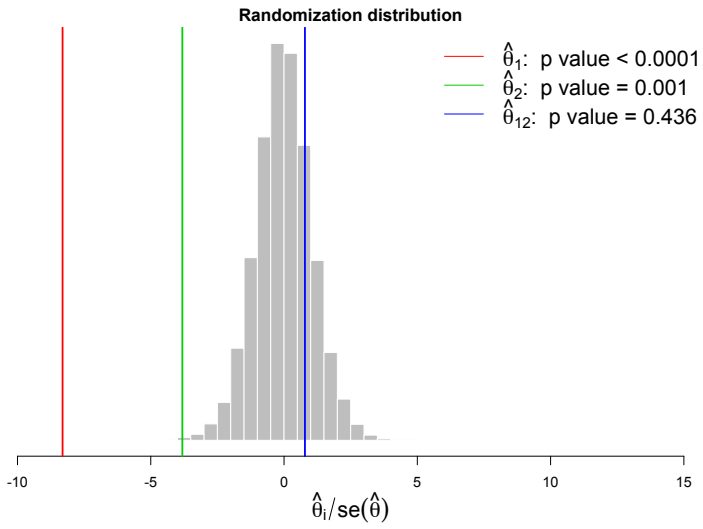


Figure 6.1: Results using the simple randomization test on all factorial effects.

It is worth noting that the main weaknesses of this approach mentioned in Chapter 2 continue to be the main weaknesses here. First, all effects are tested using the same null hypothesis of absolutely no treatment effect. However, if one of the effects is clearly not inactive, what is the point of assessing the importance of the rest of the factorial effects against the null of absolutely no treatment effect? Second, this method does not account for multiple testing. One possible solution is to use the Bonferroni correction, which is known to be too conservative for cases with a high number of factors. Recall from Chapter 2 that sequential methods have been proposed to control for the experiment-wise error rate and account for multiple testing. The proposed sequential methods have also that goal in mind, and they essentially remain unchanged from their unreplicated versions.

6.3 Example: Results with Sequential Single Imputation Randomization Tests

In the steroid and exercise experiment described above we have $|\hat{\theta}_1| \geq |\hat{\theta}_2| \geq |\hat{\theta}_{12}|$[3]. The assessment procedure goes as follows:

1. To assess the largest absolute effect we use the same randomization test described in the previous section, but with a different test statistic $\mathbf{T}$, the largest absolute standardized effect. The null at this step is

$$H_{01} : Y_i(-1,-1) = Y_i(-1,1) = Y_i(1,-1) = Y_i(1,1).$$

2. To test the second largest effect, it makes no sense to assume that the largest one is null. One option is to assume that it has *a constant additive effect* across all units, say $\theta_1^i = \theta_1^*$ for all i. This corresponds to a different sharp null:

$$H_{02} : \theta_1^i = \theta_1^*, \theta_2^i = 0, \theta_{12}^i = 0.$$

3. The last step is analogous to the previous one, however a new effect is added to the set of active effects. For the exercise and steroid example the sequence of null hypothesis is:

$$H_{03} : \theta_1^i = \theta_1^*, \theta_2^i = \theta_2^*, \theta_{12}^i = 0.$$

[3]This ordering is shown in Figure 6.1

We use the point estimates of the factorial effects to determine the values of the θ^*s at each step. In these balanced 2-level full factorial designs, due to the orthogonality of $\boldsymbol{X}$, the estimate of each factorial effect is the same regardless of which other effects are assumed inactive. This is desirable because the θ^* values given in a certain step will be the same for subsequent ones. However, this is not fundamental for the procedure to work.

Furthermore, given that the imputation is done at the unit level, all missing potential outcomes are filled in using the procedure described in the previous chapter. That is, the same $\boldsymbol{G}$ matrix is used for the unreplicated and replicated versions.

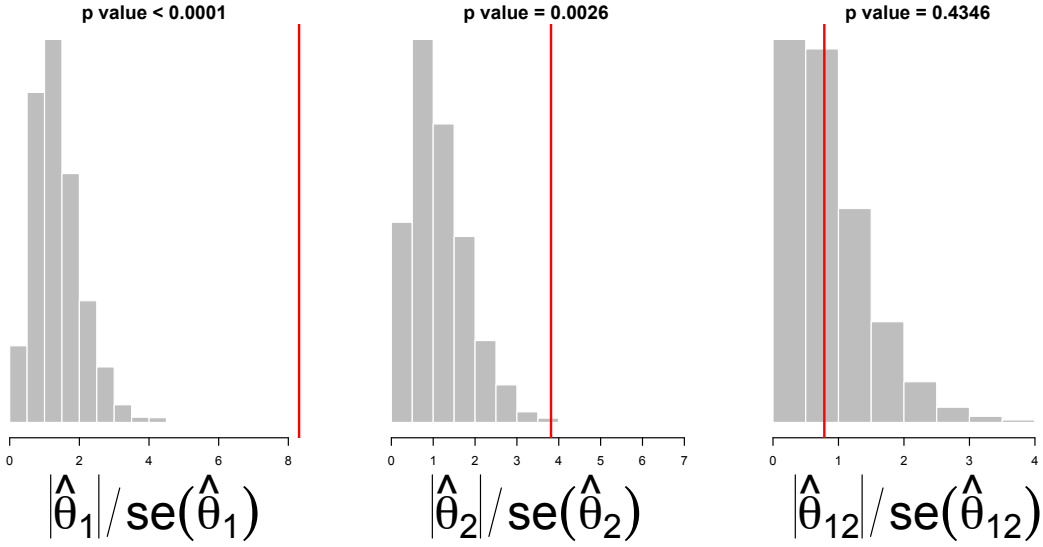
The nature of the sequence of sharp null hypotheses result in distinct randomization distributions for the maximum of all effects assumed null at each step. The randomization distribution corresponding to the first step is identical to the absolute value of the one described in the previous section for the Fisher randomization test.

Applying this to the exercise and steroid experiment we get the results displayed in Figure 6.2(a).

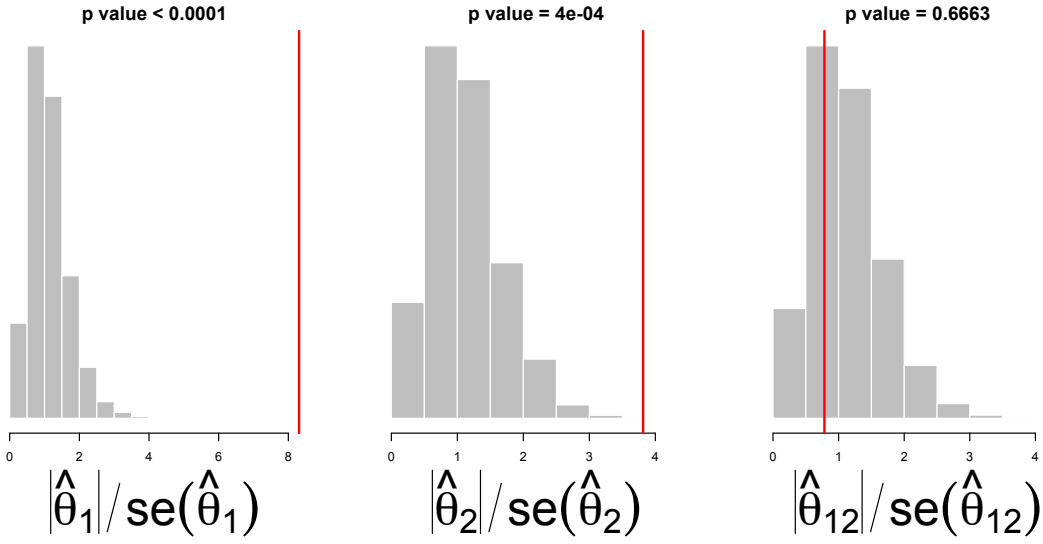
Regardless of the presence of replicates, there are two main concerns with this approach. First, the non null factorial effects at each step are assumed *constant across individuals* and, second, the uncertainty of the estimate used to impute the missing potential outcomes is ignored.

6.3.1 Extending Loughin and Noble (1997) to the Replicate Case

In Chapter 2 we discussed the relationship between this method and that proposed in Loughin and Noble (1997). We showed that although they have a similar motivation behind



(a) Single Imputation Approach



(b) Loughin and Noble Extension

Figure 6.2: Results for the Exercise and Steroid Experiment using the Single Imputation (Sequential Randomization Tests) approach and the Loghin and Noble extension.

them, the use of the potential outcomes framework does result in different methodologies. However, because Loughin and Noble (1997) performed quite well in the unreplicated case it was appealing to explore the performance of a straight forward extension of it to the replicate case. The key idea of projecting the response to a subspace (increasingly large) generated by the effects assumed active at each step and then using a permutation test on the residuals, remains the same. However, the scaling and p value calculation used in their paper are unnecessary here because of the presence of replicates. Therefore, the discrepancy measure used is the same as in the single imputation method: $\max_{\{j:j\in\mathcal{I}\}}\left|\frac{\hat{\theta}_j}{SE_{\mathcal{I}}}\right|$, and the usual randomization based p value calculation can be performed by direct comparison to the re-randomization distribution. The results of analyzing the exercise and testosterone example with this extension are shown in Figure 6.2(b).

6.4 Example: Results with Sequential Posterior Predictive Checks

To overcome the draw backs of the previous approaches we proposed the use of sequential posterior predictive checks (**SPPC**), that work both in the unreplicated and replicated cases. This procedure does makes use of distributional assumptions; in contrast to the randomization tests which are non parametric.

As previously stated, one major difference between the unreplicated and replicated case for this method is that for the replicated case the model matrix that is used to fit the model, $\boldsymbol{X}$ is not the same as the unit level matrix that is used to fill in the missing potential outcomes, $\boldsymbol{G}$.

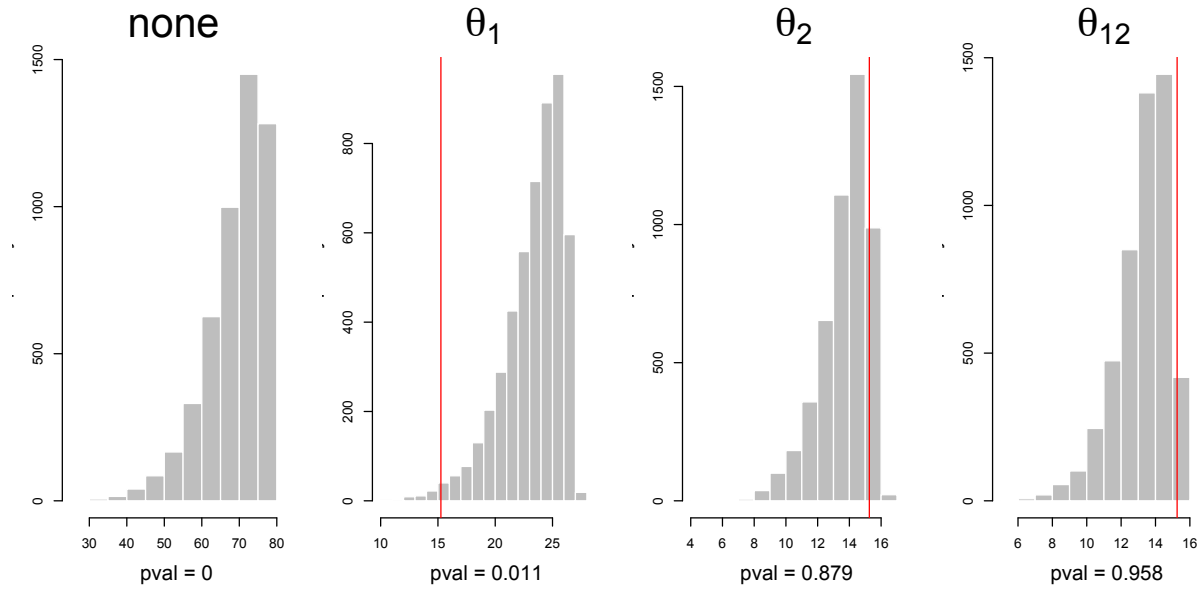
Analogous to the other methods, the choice of discrepancy measure and definition of extremeness are fundamental, and should be determined before the experiment is performed. For the replicate case we explore the use of slightly different discrepancy measures relative to those used in the unreplicated case but their motivation in both cases are vary similar. The two discrepancy measures explored for this example are the mean square residual of the full (or saturated) model, MS_{res}, and the maximum absolute value of the effects assumed inactive in the current prior, $\max_{j\in\mathcal{I}}\left|\frac{\hat{\theta}_j}{SE(\hat{\theta}_j)}\right|$. However, in the bigger simulation and motivated by the results of the unreplicated case we also included the unstandardized maximum absolute effect of $\mathcal{I}$, $\max_{j\in\mathcal{I}}\left|\hat{\theta}_j\right|$. Unlike the PSE in the unreplicated case, one motivation to choose MS_{res} when there are replicates is that it is the same across all models in the sequence of models being assessed. Because the MS_{res} is a measure of what is not explained by the model, at each step we aim to compare the observed value of $\mathbf{T}$ to its posterior predictive distribution where a specific set of effects are active. Hence, it is a one sided test because we are looking for the first model for which the MS_{res} is not "too high" relative to the posterior predictive distribution. One important thing to keep in mind when using the MS_{res} as the discrepancy measure is that it is a sufficient statistic for the model that includes all the effects. Hence, the full model (which includes all factorial effects) will not be rejected with the posterior predictive checks. This discrepancy measure is similar to the $PSE_{\mathcal{I}}$ used for the unreplicated case but they are fundamentally different. The $PSE_{\mathcal{I}}$ for the saturated model is not defined and therefore, for the unreplicated case the discrepancy measure related to unexplained variation from the model did change from one step of the process to the next unlike the MS_{res} in the replicated case. Nevertheless, in neither case this measure displayed the best performance.

The $\max_{j\in\mathcal{I}} \left|\frac{\hat{\theta}_j}{SE(\hat{\theta}_j)}\right|$ option for discrepancy measure was considered because it is a useful statistic to find outliers which is the ultimate goal when screening for effects (i.e., to find the outliers among all factorial effects). Furthermore, the use of this statistic in the methods proposed in the literature for the unreplicated case (which use the PSE to standardize) was very effective. In addition, it is expected to be less sensitive to the normality assumption. However, recall that using the PSE did not improve the performance of the S-PPC in the unreplicated case. Regardless of the presence of replicates, the saturated model can't be tested with this approach because the active set is empty.

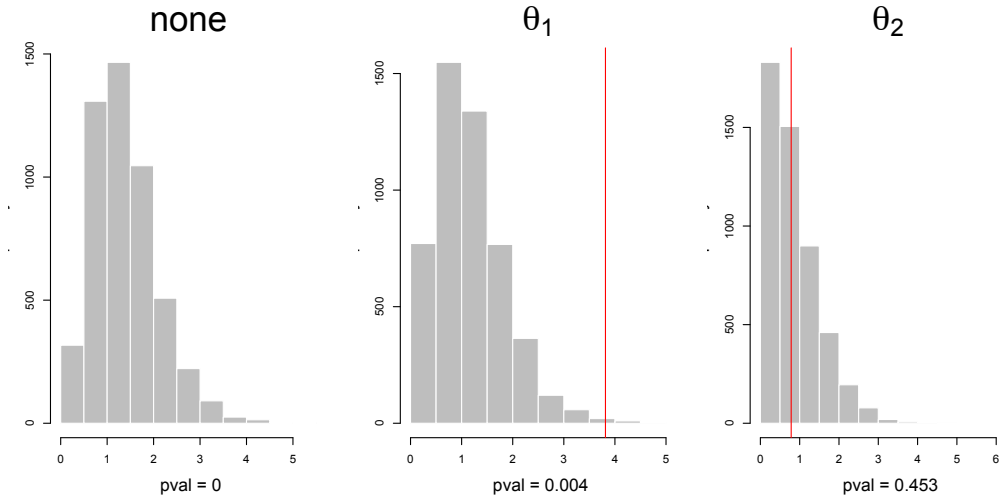
The results for the exercise and steroid experiment with this method and with both discrepancy measures are displayed in figures 6.3(a) and 6.3(b). Using a cut off value of 0.05, the third model is the first one that seems consistent with the data observed. This corresponds to the model with only the two main effects present.

6.5 Simulation Study

Although all methods lead to the same conclusion in the exercise and testosterone example, it is clear that there are differences in the methodologies behind them. Furthermore, the results in the unreplicated case highlight these differences and suggest the further exploration of their relative performances in the replicate case. A simulation study analogous to the one performed for the unreplicated case is now presented.



(a) Results using MS_{res} as the discrepancy measure.

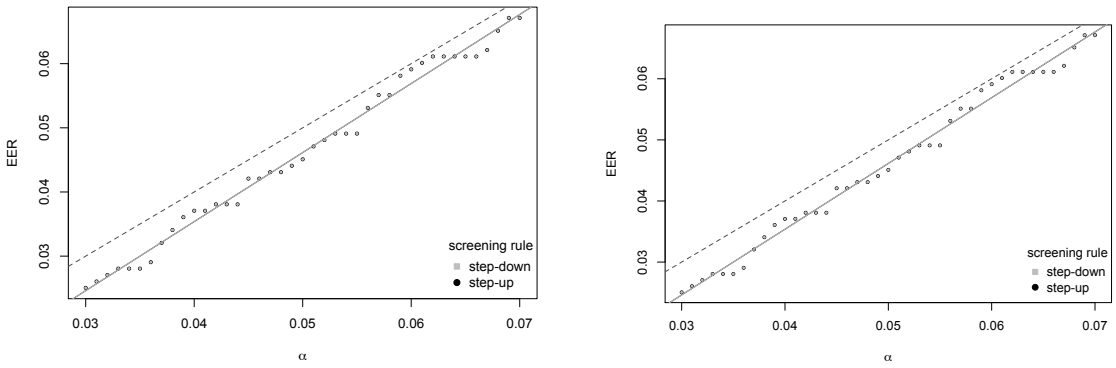


(b) Results using $\max_{j \in \mathcal{I}} \left| \frac{\hat{\theta}_j}{SE(\hat{\theta}_j)} \right|$ as the discrepancy measure.

Figure 6.3: Posterior predictive distributions of discrepancy measures for the steroid and exercise experiment. The title of each subplot reflects the last factorial effect that was identified as active. Therefore all the factorial effects that were previously assessed are also deemed active.

6.5.1 Calibration Study-Under the Null Hypothesis

The calibration study results of the single imputation approach for the replicate case are displayed below, in Figure 6.4. Note that the EER values obtained from different screening rules and the statistics using either estimate of the variance are the same. The OLS fit, $EER = -0.008 + 1.076\alpha$, suggests a cut off of 0.054, whereas the observed EER values suggest a cut off of 0.055. As before, we will continue to choose the one suggested by the observed values. Interestingly, for the single imputation approach the choice of the variance estimate does not seem to have an effect, in contrast to the results for the **S-PPC**.



(a) Unpooled Estimate of the Variance (b) Pooled Estimate of the Variance

Figure 6.4: Scatter plots of the Calibration Study for the Single Imputation Approach. The dashed line represents the identity line. The continuous lines denote the OLS fits of the observed EER on the α level used for each screening rules.

The results of the calibration of the **S-PPC** with all four discrepancy measures and the two screening rules are displayed in Figure 6.5. In this part of the study we explore further whether the different screening rules continue to have a strong effect on the results, and if this depends on the statistic used.

Keeping in mind that these simulations were performed for 2 replicates, it is expected

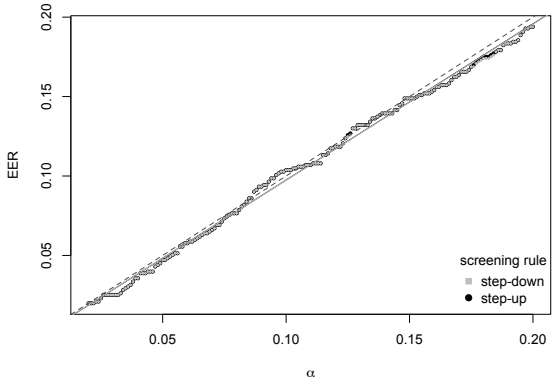
that the unpooled estimate of the variance is very unstable which is confirmed by these results. In Figure 6.5(b) and because of the noise in this estimation, it is clear that neither screening rule has an acceptable performance. It is expected that the performance would improve as the number of replicates increases. However, it was excluded in the subsequent stage of this simulation study.

Figure 6.5(a) shows that for the absolute value of the usual t statistic, $\max_{\{j:j\in\mathcal{I}\}}\left|\frac{\hat{\theta}_j}{SE}\right|$, both screening rules lead to basically the same results, which are very close to the relationship $\alpha = EER$. In fact, the OLS fit for both screening rules is approximately $EER = 0.001 + 0.983(2)\alpha$ leading to a cut off value of 0.052, and the observed EER values in the calibration lead to a same cut off value.

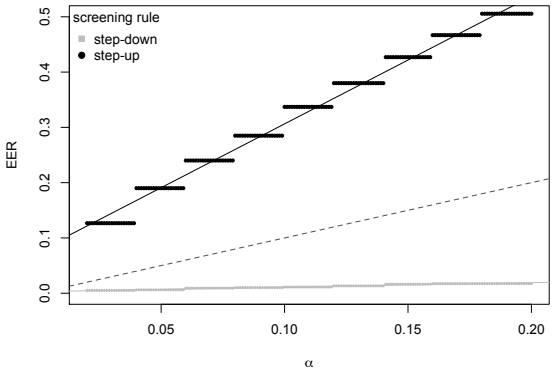
Given the superior performance of the raw scale factorial effect $\max_{\{j:j\in\mathcal{I}\}}\left|\hat{\theta}_j\right|$ over the standardized in the unreplicated case, we decided to include this discrepancy measure in the study of the replicate case. As seen in Figure 6.5(c), the step-up procedure leads to a higher slope, but the difference is negligible for values of α below 0.06. The observed EER values suggest the use of 0.51 as the cut off for both screening rules because it is the last value of α that leads to an EER below 0.05. The OLS fits are $EER = 0.958\alpha$ and $EER = -0.001 + 1.001\alpha$ for the step-down and step-up screening rules respectively. These fits suggest the corresponding cut offs of 0.051 and 0.052. We chose 0.051 for both.

For the MSE discrepancy measure both procedures lead to the same results. The observed EER values and the OLF fits ($EER = -0.004 + 1.035\alpha$) suggest a cut off value of 0.055 for α.

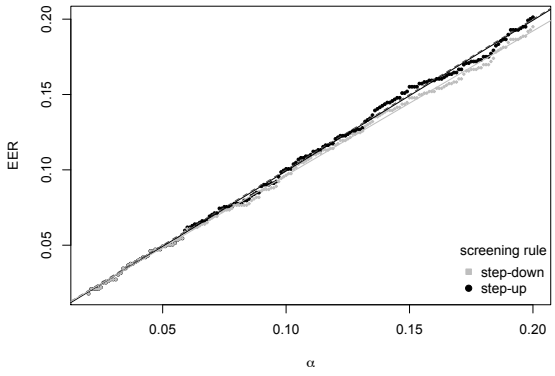
We display the cutoffs used for the simulation across alternative settings in Table 6.3.



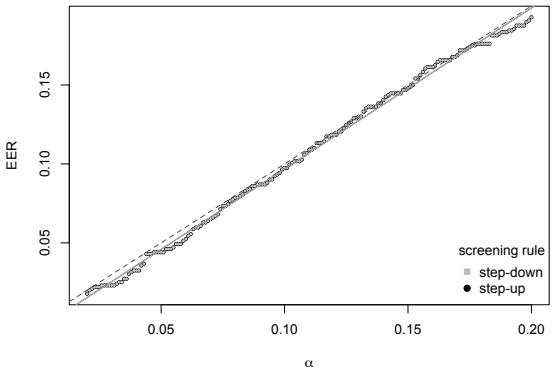
(a) Discrepancy Measure: $\max_{\{j:j\in\mathcal{I}\}}\left|\frac{\hat{\theta}_j}{SE}\right|$



(b) Discrepancy Measure: $\max_{\{j:j\in\mathcal{I}\}}\left|\frac{\hat{\theta}_j}{SE_{Ney}}\right|$



(c) Discrepancy Measure: $\max_{\{j:j\in\mathcal{I}\}}\left|\hat{\theta}_j\right|$



(d) Discrepancy Measure: MSE

Figure 6.5: Scatter plot of the Calibration Study for Sequential Posterior Predictive Checks. The greyscale denotes the two screening rules. The dashed line represents the identity line, and each continuous line denotes the OLS fit of the observed EER on the α level used for the corresponding screening rule. Each subfigure corresponds to a different discrepancy measure.

Table 6.3: Cutoffs obtained in the calibration of the Sequential Posterior Predictive Checks using 2^4 factorial designs for the different methods, discrepancy measures and screening rules proposed.

Method	**Discrepancy Measure**	**Screening Rule**	α
SI	$\max_{\{j:j\in\mathcal{I}\}} \left\lvert \frac{\hat{\theta}_j}{SE_{\mathcal{I}}} \right\rvert$	both	0.055
	$\max_{\{j:j\in\mathcal{I}\}} \left\lvert \frac{\hat{\theta}_j}{SE_{\mathcal{I}}^{Ney}} \right\rvert$	both	0.055
S-PPC	$\max_{\{j:j\in\mathcal{I}\}} \left\lvert \hat{\theta}_j \right\rvert$	both	0.051
	$\max_{\{j:j\in\mathcal{I}\}} \left\lvert \frac{\hat{\theta}_j}{SE_{\mathcal{I}}} \right\rvert$	both	0.052
	MSE_{sat}	step-down	0.055

6.5.2 Simulation across Alternative Hypotheses

Following the same reasoning as Tripolski et al. (2008), we included in the simulation of the alternative hypotheses the Benjamini and Hochberg (1995) method applied to the regression p values as an interesting and widely used method to compare our results to. However, there is a slight difference in the screening rule suggested for the unreplicated case and the general one given in Benjamini and Hochberg (1995), which states that letting "$p_{(1)}, \ldots, p_{(m)}$ be the ordered p-values and denote by $H_{(i)}$ the null hypothesis corresponding to $p_{(i)}$. Define the following Bonferroni-type multiple testing procedure: let k de the largest i for which $p_{(i)} \leq i/mq*$; then reject all $H_{(i)}$ i=1,. . . ,k." Therefore, as summarized in Tripolski et al. (2008), the BH procedure declares the biggest k effects as active. In contrast, for the unreplicated case they suggest "For controlling the FDR, [...] declare as active all effects that have a p-value smaller than the largest p-value that upholds the inequality $p_{(i)} \leq i/mq*$."

For the replicate case, we implemented the Benjamini and Hochberg (1995) procedure.

Similar to the previous chapter, we simulated potential outcomes repeatedly (1000 times) from a total of 108 alternative hypotheses defined by setting $\max_{\{j:j\in\mathcal{A}\}}|\mu_j|$ to 4, and different levels of noise ($\sigma = 0.5, 1, 2$), number of active effects ($a = 1, 2, 4, 6$), the range between active effects (defined as $\rho = \max_{\{j:j\in\mathcal{A}\}}|\mu_j| - \min_{\{j:j\in\mathcal{A}\}}|\mu_j| = 1, 2, 3$), and the number of replicates $r = 3, 5, 10$. Each combination of these simulation factors determines a true value of $\boldsymbol{\mu}$ by selecting a factorial effects at random to be active, and letting the corresponding μ_j's assume the value $4 - \rho$ if $a = 1$, and the values $4 - \rho\left(1 - \frac{t}{a-1}\right)$ for $t = 0, \ldots, a-1$ if $a > 1$. The remaining μ_j's are set at zero. Finally, the potential outcomes for each of the $r2^4$ units are drawn from $\boldsymbol{Y}_i(\boldsymbol{z})|\boldsymbol{\mu}, \sigma^2 \sim N(\boldsymbol{\mu}, \sigma^2)$.

We compare these methods based on five summary measurement based on averages across 1000 simulated data sets: **IER** (average proportion of false positives), **EER** (proportion of data sets with at least one false positive), **FDR** (average proportion of false positives among all the effects declared active), **RR** (average proportion of true positives), and **ANP** (average number of positive effects declared active). In this context, positive effects are the factorial effects declared active, false positives are inactive effects incorrectly identified as active, and true positives are active effects that are correctly identified as such.

The averages across all 108 alternative settings are displayed on Table 6.4. To better understand the **ANP** measure, note that the average number of true active effects across the simulation settings is 3.25. The performance of the methods is a lot closer than it was for the unreplicated case. This is not surprising because of the increase in the information available in the data and the ability to better estimate the variance parameter.

In this case, where we have replicates, the additional information allows us to estimate

the error and mostly eliminate the masking effect that active effect incorrectly identified as inactive can have by blowing up the variance. Therefore the RR for the basic procedures like the permutation test and linear regression are higher that the other methods as would be expected but not observed in the unreplicated case (see 5.2) because of the masking effect that can be so strong. Interestingly, the standardized discrepancy measure, $\max_{\{j:j\in\mathcal{I}\}}\left|\hat{\theta}_j\right|$, does allow for some of this masking to be identified leading to a difference between the step up and step down procedures. However, these procedures lead to the same solution if we use the standardized discrepancy measure. This increase in rejection rate comes at the cost of increased sensitivity to the noise in the data (as seen in Figure 6.9).

As in the unreplicated case, the MSE_{res} underperforms but it is a lot closer to the discrepancy measures involving the maximum absolute factorial effect.

Several figures are now displayed to illustrate the performances of the different methods. For the correct interpretation of these plots it is important to keep in mind the scales on the y-axis because the plots are made to emphasize the differences but, taking in to account the scale, the practical differences between these methods might still be small in the alternative settings explored.

Among all the methods that do account for multiple comparisons, the FDR is best in terms of the RR and the FDR (because it is closest to 0.05). However, the average EER across all alternatives is 0.137, and Figure 6.6 shows that for some settings the EER can exceed 0.2.

In terms of identifying the permutation versions of the commonly used procedures, note that the permutation test agrees with the results obtained by linear regression. Interestingly enough, our extension of the Loughin and Noble procedure to the replicate ca agrees with

Table 6.4: Summary of results of the simulation study. Average rates and number of effects declared active, as well as the standard errors of these quantities are displayed for all methods and screening rules under the null ($\sigma = 1$ and $r = 3$) and across alternative hypotheses. The single imputation method we only report the results obtained using the Neyman - unpooled - estimate of the variance. The calibration for the pooled estimate of the variance has an error which still needs to be resolved (e.g. leads to an EER under the null of 0.018 and, across the alternatives, an RR of 0.907, EER of 0.016 and FDR of 0.006).

			null hypothesis			**average across alternative hypotheses**				
Method	**Discrepancy Measure**	**Screening Rule**	**IER** (SE)	**EER** (SE)	**ANP** (SE)	**RR** (SE)	**IER** (SE)	**EER** (SE)	**FDR** (SE)	**ANP** (SE)
Linear Regression	$\frac{\hat{\theta}_j}{SE}$	-	0.049 (0.002)	0.496 (0.016)	0.739 (0.029)	0.966 (0.003)	0.050 (0.002)	0.425 (0.016)	0.158 (0.006	3.762 0.027)
Permutation	$\lvert\hat{\theta}_j\rvert$	-	0.049 (0.002)	0.494 (0.016)	0.738 (0.029)	0.966 (0.003)	0.050 (0.002)	0.425 (0.016)	0.157 (0.006	3.760 0.027)
Linear Regression - Bonferroni	$\frac{\hat{\theta}_j}{SE}$	-	0.003 (< 0.001)	0.043 (0.006)	0.043 (0.006)	0.918 (0.004)	0.003 (0.001)	0.037 (0.006)	0.013 (0.002)	3.087 (0.012)
Linear Regression - FDR	$\frac{\hat{\theta}_j}{SE}$	-	0.004 (0.001)	0.044 (0.006)	0.06 (0.010)	0.951 (0.004)	0.015 (0.001)	0.137 (0.011)	0.039 (0.004)	3.312 (0.016)
L&N	$\left(\frac{N-1}{N-\lvert\mathcal{I}\rvert}\right)^{1/2} \max_{\{j:j\in\mathcal{I}\}} \left\lvert\frac{\hat{\theta}_j}{SE_{\mathcal{I}}}\right\rvert$	step-up	0.003 (< 0.001)	0.039 (0.006)	0.039 (0.006)	0.908 (0.004)	0.001 (< 0.001)	0.016 (0.004)	0.007 (0.002)	3.020 (0.011)
	$\max_{\{j:j\in\mathcal{I}\}} \left\lvert\frac{\hat{\theta}_j}{SE_{\mathcal{I}}}\right\rvert$	step-up	0.003 (< 0.001)	0.039 (0.006)	0.039 (0.006)	0.917 (0.004)	0.003 (0.001)	0.035 (0.006)	0.012 (0.002)	3.081 (0.012)
Single Imputation	$\max_{\{j:j\in\mathcal{I}\}} \left\lvert\frac{\hat{\theta}_j}{SE_{\mathcal{I}}}\right\rvert$	both	0.004 (< 0.001)	0.051 (0.007)	0.053 (0.007)	0.924 (0.004)	0.005 (0.001)	0.055 (0.007)	0.018 (0.003)	3.128 (0.013)
S-PPC	$\max_{\{j:j\in\mathcal{I}\}} \lvert\hat{\theta}_j\rvert$	step-up	0.003 (< 0.001)	0.049 (0.007)	0.050 (0.007)	0.928 (0.004)	0.005 (0.001)	0.051 (0.007)	0.016 (0.003)	3.132 (0.013)
		step-down	0.003 (< 0.001)	0.048 (0.007)	0.048 (0.007)	0.889 (0.005)	0.007 (0.002)	0.049 (0.007)	0.017 (0.003)	2.934 (0.027)
	$\max_{\{j:j\in\mathcal{I}\}} \left\lvert\frac{\hat{\theta}_j}{SE_{\mathcal{I}}}\right\rvert$	step-up	0.004 (< 0.001)	0.050 (0.007)	0.053 (0.007)	0.923 (0.004)	0.005 (0.001)	0.053 (0.007)	0.017 (0.003)	3.123 (0.013)
		step-down	0.004 (< 0.001)	0.050 (0.007)	0.053 (0.007)	0.923 (0.004)	0.005 (0.001)	0.053 (0.007)	0.017 (0.003)	3.122 (0.013)
	MSE_{sat}	step-up	0.004 (0.001)	0.052 (0.007)	0.063 (0.010)	0.884 (0.005)	0.151 (0.008)	0.174 (0.009)	0.120 (0.006)	4.888 (0.106)
		step-down	0.004 (0.001)	0.052 (0.007)	0.063 (0.010)	0.883 (0.005)	0.007 (0.002)	0.042 (0.006)	0.017 (0.003)	3.038 (0.023)

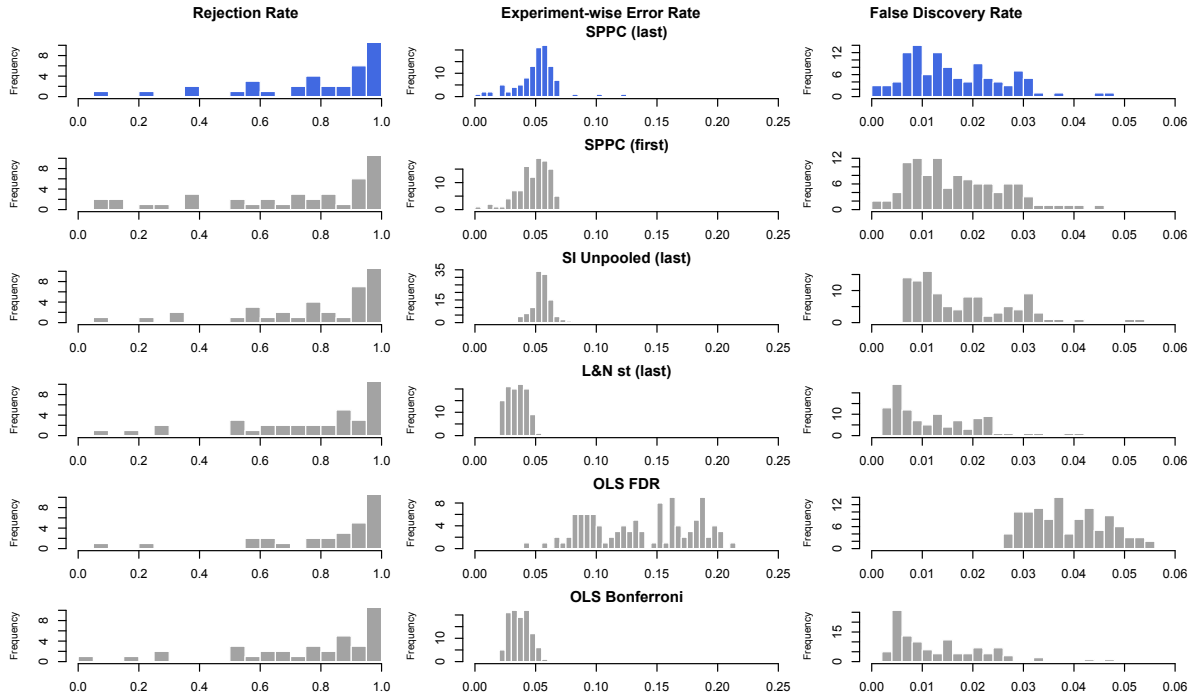


Figure 6.6: Distributions of the Rejection (RR), the False Discovery (FDR), and Experiment-wise Error (EER) Rates across the different alternative hypotheses.

the results obtained with the Bonferroni correction.

In Figure 6.6 we display the RR, FDR, and EER, across the different simulation alternative hypotheses settings for a subset of the previously established methods and compare them to the S-PPC with the best performance. As expected because of the increase in information about the variability in the data due to the presence of replicates, most rejection rates are close to 1 across the 108 simulation settings. Therefore, to better visualize the differences between the methods, we truncated the frequency axis (y-axis) for the rejection rate at 15. This patterns displayed in this figure are very different to what was observed for the unreplicated case in Figure 5.2. Here, the step-up S-PPC with discrepancy measure $\max_{\{j:j\in\mathcal{I}\}}\left|\frac{\hat{\theta}_j}{SE_{\mathcal{I}}}\right|$, the best S-PPC settings (BS-PPC), does have distinct modes around 1 for the RR (like the unreplicated case), but the modes are not at 0 for the FDR and EER. It

is hard to compare the distributions of RRs, but the tail for the FDR correction seems to be slightly lighter. Specifically, the FDR correction hast 2 combinations with RR below 0.4 whereas all other methods have at least 4. The EER distribution for the FDR correction method is all over the place. It is not surprising that it is not close to 0.05 because that is not their goal. The distribution of FDRs is a clear illustration of their goal because their method aims to push the false discovery rates towards 0.05. Note, however, that there are still some combinations that are above that 0.05 threshold. All other methods are skewed towards zero and do not have values above 0.05.

For the replicate case our method is in fact a trade off between controlling the EER and increasing the RR. Regarding the EER, it is not as good as some other methods (our version of replicate L&N and Bonferroni) but much better at it than the FDR correction. Regarding the RR, it is better than our version of replicate L&N and Bonferroni but not as powerful as the FDR correction.

For our proposal, there are no combinations with an FDR above 0.05. However, there are eight combinations with EER above 0.05. In contrast, the L&N and Lenth methods have none above this threshold, at the cost of lower rejection rates. On the other hand, the Lenth method with the FDR correction leads to EER values above 0.05 for 30 of the 36 settings for the alternative hypotheses, which is not surprising because the goal is to control the FDR. Nevertheless, their method has 5 combinations with FDR above 0.05.

Furthermore, Figure 6.7 shows that for larger number of active effects (slightly for $a = 2$ but much more evident for $a = 4, 6$) the average number of effects declared active with the FDR correction is higher than the true number of active effects. An interesting questions is whether it is more worrisome that these quantities are above the truth than below.

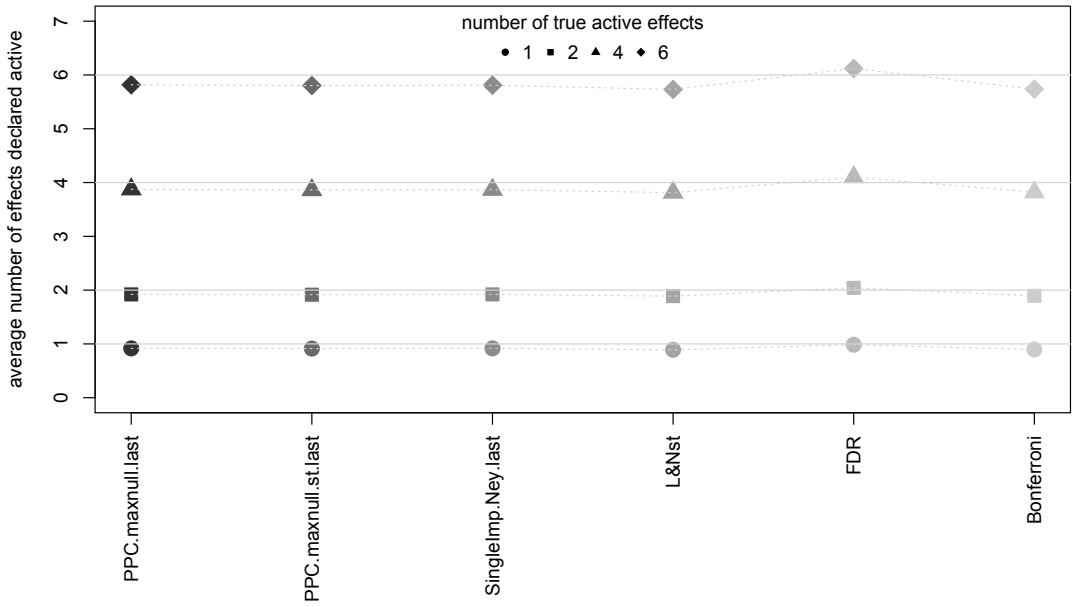


Figure 6.7: Average number of effects declared active for the different values of *true* active effects = 1,2,4,6.

Figure 6.8 compares the RR, EER and FDR average values across different number of active effects. In terms of the proposed methods it is clear that the standardized discrepancy measure, $\max_{\{j:j\in\mathcal{I}\}}\left|\frac{\hat{\theta}_j}{SE_{\mathcal{I}}}\right|$, has the best performance with a slightly higher RR while keeping lower values of EER and FDR. Again, the FDR correction method is shown to have a better overall performance in the RR and FDR. However, it is clear that its behavior in terms of the EER depends heavily on the true number of active effects, Figure 6.7 shows that this method tends to declare more effects active that there actually are. In contrast, for this method the FDR decreases as the number of active effects increases

Like in the unreplicated case, Figure 6.9 shows that the best performing BS-PPC is sensitive to the noise factor unlike most other methods. The FDR corrected method also shows sensitivity to σ in terms of the EER, but lower than that of the BS-PPC. In contrast,

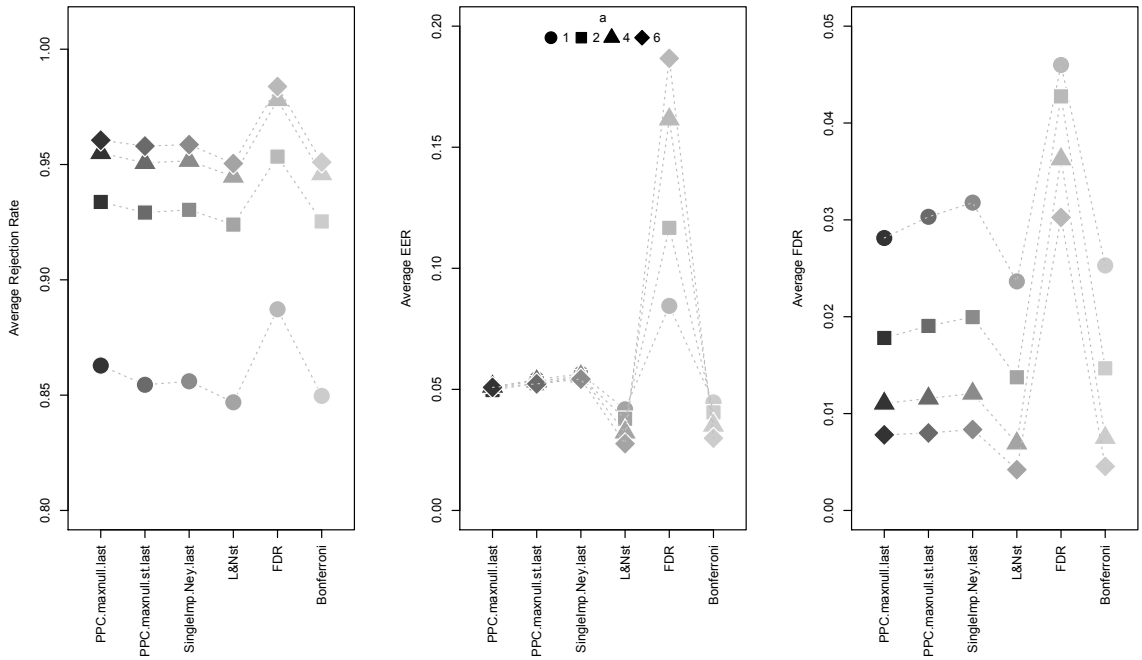


Figure 6.8: Comparison of RR, EER, and FDR across simulations settings with the same value of active effects = 1,2,4,6. The standard errors of these values range from 0.005 to 0.011 for the RR, from 0.001 to 0.010 for the EER, and from 0.001 to 0.007 for the FDR.

all methods, but the FDR corrected method are sensitive to sigma in terms of the FDR, with the BS-PPC being the most sensitive one.

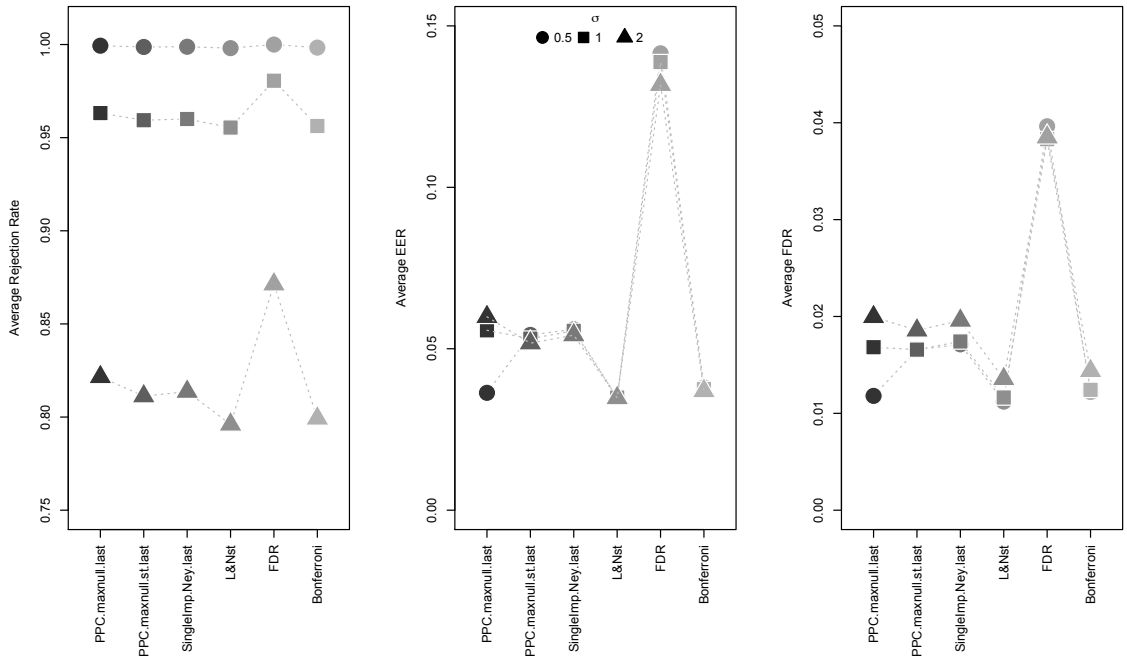


Figure 6.9: Comparison of RR, EER, and FDR across simulations settings with the same value of $\sigma = 0.5, 1, 2$. The standard errors of these values range from 0.0001 to 0.0063 for the RR, from 0.005 to 0.011 for the EER, and from 0.002 to 0.0037 for the FDR.

Figure 6.11 displays the effect of the number of replicates on the performance on the different methods. The comparative advantage of the FDR decreases with the number of replicates although it continues to be higher for all values of r. This simulation factor does not have an effect on the performances on EER and FDR for the BS-PPC and the FDR corrected methods.

The interaction plots presented in Figure 6.12 allow us to take a closer look at what factors play an important role in the different performances of the methods. Only three methods are compared: the OLS with the Bonferroni and FDR corrections, and the BS-PPC.

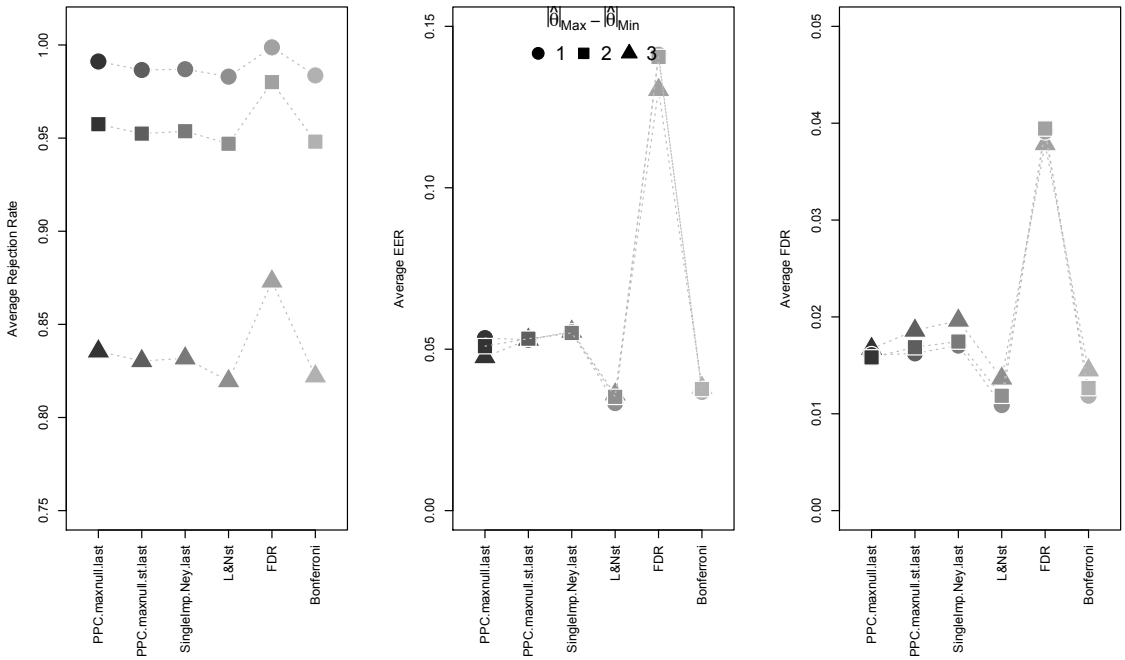


Figure 6.10: Comparison of RR, EER, and FDR across simulations settings with the same value of $\rho = 1, 2, 3$. The standard errors of these values range from 0.001 to 0.0057 for the RR, from 0.0057 to 0.011 for the EER, and from 0.002 to 0.0036 for the FDR.

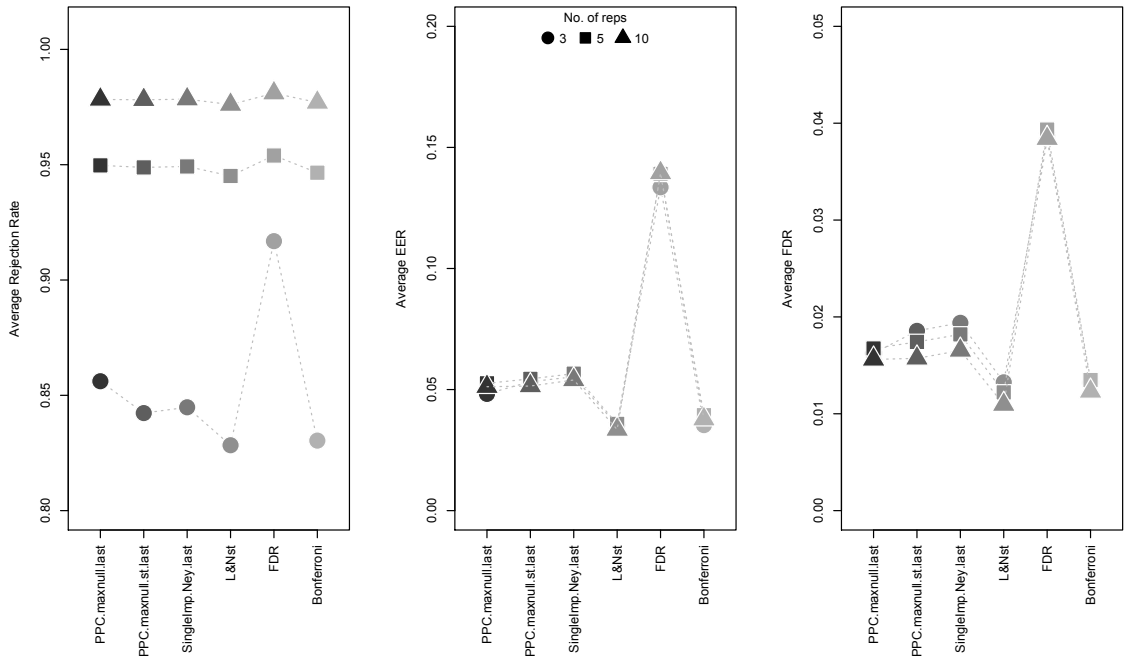
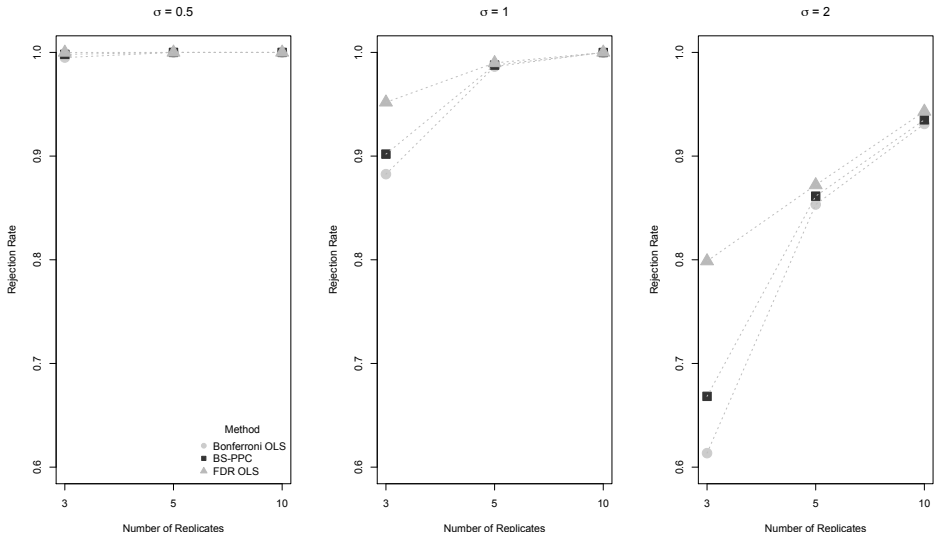
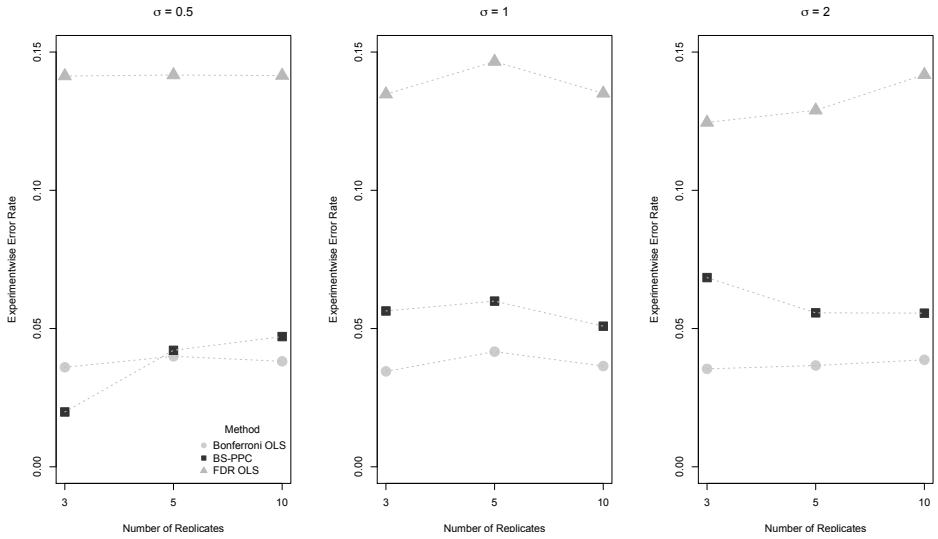


Figure 6.11: Comparison of RR, EER, and FDR across simulations settings with the same value of $r = 3, 5, 10$. The standard errors of these values range from 0.003 to 0.0057 for the RR, from 0.0057 to 0.011 for the EER, and from 0.0021 to 0.0036 for the FDR.



(a) Rejection Rates



(b) Experimentwise Error Rates

Figure 6.12: Interaction Plots showing the relationship between the number of replicates, ρ, and noise, σ for three methods: OLS with the Bonferroni and FDR corrections, and the BS-PPC.

Perhaps the gains in using the unstandardized maximum effect assumed null with the step-up procedure are not enough to justify the loss in efficiency that can be gained by using the standardized version where both procedures lead to the same results. Therefore, it might be appealing to use $\max_{\{j:j\in\mathcal{I}\}}\left|\frac{\hat{\theta}_j}{SE_{\mathcal{I}}}\right|$ as the discrepancy measure for either direction for assessing the sequences.

In preliminary simulation studies we used two replicates $r = 2$. However, that value of r is not included in this more comprehensive simulation study. In our initial exploration of the method we did not include the FDR approach so we had no initial idea of the relative power the FDR to that of the SPPC. However, we did expect it to be more powerful than all the existing methods given the results reported in the literature.

Other interesting steps are to test these procedures when the underlying distribution of the potential outcomes does not satisfy the classical assumptions, such as different variances of potential outcomes across treatment combinations or a nonnormal distribution. For this last point, it is relevant to consider what the classical approaches are in this setting. The use of Box-Cox transformations have withstood the test of time and a closer look into them might be of interest. However, a more appealing approach would be to tackle it directly from a fully Bayesian model, for example, using conjugate priors.

An appealing direction is the search of alternative discrepancy measures or stopping rules regarding the observed patterns of the posterior predictive p values of the different measures, perhaps distinguishing the active and inactive effects in better ways than cutoffs. An initial exploration of bivariate options when the alternative hypothesis is symmetric (i.e., when all active factors have the same magnitude) was performed, something that would be desirable to pursue further.

The simulation results show that, simulating from the setting assumed by the traditional approaches, the most power method is to use the FDR correction on the OLS results. This has the additional benefit of computational efficiency. One major drawback of this approach though is the high experiment-wise error rates that the experimenter can be getting (above 0.1 usually). The advantage of the FDR correction in terms of RR decreases as the number of replicates increase, but its EER remains much higher than other methods.

We still believe that our method adds value to the existing methodologies because, unlike all other methods, it is proposed with the finite population in mind. Further exploration of the performance of other methods when actually simulating from the setting we are assuming is a necessary future step to better understand the properties of our proposal. The **BS-PPC** performs between the Bonferroni and FDR corrections of OLS, in the sense that it sacrifices some of the power that FDR achieves in order to have a better control of the EER. In general, our method is performing well on all three major outcome measures, the FDR, the EER and the RR.

Moreover, to practically interpret the results shown in the previous section one should keep in mind the scales of the y-axis because they were chosen to highlight the differences between the methodologies being compared. For example, the maximum difference on the RR between the FDR corrected OLS and the BS-PPC is in Figure 6.11 and is under 0.07, and in all other plots it is below 0.05. However, the difference between these methods in terms of the EER is consistently higher, usually above 0.05 and in some cases above 0.1 (see Figure 6.9).

Chapter 7

Extensions, Future Steps & Conclusions

Given the work that we have presented up to now, there are two natural extensions: fractional factorial and three level designs. The exploration of these additional settings has also been motivated by an applied project on stem cell research related to direct differentiation of stem cells into pancreatic β cells. Another extension motivated by this project is to unbalanced designs, where the particular question of interest is whether SPPC performs better that existing methods for analysis of unbalanced designs.

7.1 Three level designs

Three level designs are commonly used when there is a belief of a curvature in the relationship between the response and quantitative factors, because such a question cannot be addressed using a two level design. These designs are also employed when a treatment

factor is qualitative and can take one of three possible levels (for example, three machines, three flavors or three experimenters). Furthermore, when there is an interest in studying the effect of modifying a current dose or setting in a manufacturing process, it is common to explore two additional values around the current one using three level designs.

7.1.1 RCM for three-level factorial designs

As in the two-level case, an unreplicated 3^2 full factorial design is used to introduce the concepts. In this design, each of the two treatment factors can take one of *three* levels, typically denoted by 0, 1 and 2. Thus, there are nine treatment combinations denoted by $\boldsymbol{z} = (0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1)$ and $(2,2)$ and nine experimental units. Let $Y_i(\boldsymbol{z}), i = 1, \ldots, 9$, denote the potential outcome of the ith unit if exposed to treatment combination $\boldsymbol{z}$. The ith unit has nine potential outcomes, which comprise the 1×9 row vector $\boldsymbol{Y}_i$. That is,

$$\boldsymbol{Y}_i = (Y_1(0,0), Y_1(0,1), Y_1(0,2), Y_1(1,0), Y_1(1,1), Y_1(1,2), Y_1(2,0), Y_1(2,1), Y_1(2,2)).$$

Finally, we define *the Science* as the 9×9 matrix $\boldsymbol{Y}$ of potential outcomes in which the each row corresponds to the 9-component row vector of each unit's potential outcomes $\boldsymbol{Y}_i$, $i = 1, \ldots, 9$ as shown in Table 7.1, and like 1.1 this table could be extended to include all the unit level factorial effects $\boldsymbol{\theta}_i$. Again, only one potential outcome in each row of the Science is actually observed from an experiment, and the remaining eight are missing, making the causal inference problem a missing data problem.

For each unit, all levels of every factor appear on a third of its potential outcomes.

Table 7.1: The *Science* for the full experiment.

Unit	Potential outcome for treatment combination								
(i)	$(0,0)$	$(0,1)$	$(0,2)$	$(1,0)$	$(1,1)$	$(1,2)$	$(2,0)$	$(2,1)$	$(2,2)$
1	$Y_1(0,0)$	$Y_1(0,1)$	$Y_1(0,2)$	$Y_1(1,0)$	$Y_1(1,1)$	$Y_1(1,2)$	$Y_1(2,0)$	$Y_1(2,1)$	$Y_1(2,2)$
2	$Y_2(0,0)$	$Y_2(0,1)$	$Y_2(0,2)$	$Y_2(1,0)$	$Y_2(1,1)$	$Y_2(1,2)$	$Y_2(2,0)$	$Y_2(2,1)$	$Y_2(2,2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
9	$Y_9(0,0)$	$Y_9(0,1)$	$Y_9(0,2)$	$Y_9(1,0)$	$Y_9(1,1)$	$Y_9(1,2)$	$Y_9(2,0)$	$Y_9(2,1)$	$Y_9(2,2)$
Average	$\overline{Y}(0,0)$	$\overline{Y}(0,1)$	$\overline{Y}(0,2)$	$\overline{Y}(1,0)$	$\overline{Y}(1,1)$	$\overline{Y}(1,2)$	$\overline{Y}(2,0)$	$\overline{Y}(2,1)$	$\overline{Y}(2,2)$

Therefore, at the unit-level we are generally interested in contrasts of these thirds of the unit's potential outcomes. Each factor has two degrees of freedom associated with it, and hence, it would be desirable to split each treatment factor with levels 0, 1 and 2 into two orthogonal contrasts, each associated with one degree of freedom because in such a case the methods proposed for the two level designs can be easily applied to the three level designs. Furthermore, the use of such a breakup might ease the interpretation of the results. Splitting the treatment factor into two orthogonal contrasts makes the most sense when we are dealing with quantitative factors. This is commonly referred to as the linear and quadratic (**LQ**) system (Wu and Hamada 2009, Chapter 5).

7.1.2 Linear and Quadratic Contrasts

We now describe a way of splitting a quantitative treatment factor with levels 0, 1 and 2 (which ideally are equally spaced) into one linear and one quadratic contrasts presented as an alternative analysis method in Wu and Hamada (2009). The linear effect is defined as $Y_i(2) - Y_i(0)$ and the quadratic effect as $(Y_i(2) + Y_i(0)) - 2Y_i(1)$, which can be re-expressed as the difference between two consecutive linear effects $(Y_i(2) - Y_i(1)) - (Y_i(1) - Y_i(0))$.

Mathematically these contrasts correspond to the following vectors:

$$l = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \quad q = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

However, we need to relate this decomposition to the $\boldsymbol{g}$ vectors which make up the $\boldsymbol{G}$ matrix for the two level case described in previous chapters. Unlike those, these orthogonal vectors do not have the same norm. The squared norm of the linear component is 2, whereas the squared norm of the quadratic component is 6. Therefore, to ensure a fair comparison of these contrasts, they need to be scaled to have the same norm (i.e. equal variance, which is an assumption in all of the methods), something that we get automatically in the two level case.

We now show how to define the interaction factorial effects in a 3-level factorial design using the 3^2 design as an example. We define the linear main effects of factors 1 and 2 to be normalized contrasts (i.e. $||\boldsymbol{g}_{\cdot,\cdot}|| = 1$) of the unit level potential outcomes as follows:

$$\begin{aligned} \theta_{i1,L} &= \frac{Y_i(2,0) + Y_i(2,1) + Y_i(2,2)}{\sqrt{6}} - \frac{Y_i(0,0) + Y_i(0,1) + Y_i(0,2)}{\sqrt{6}} = \boldsymbol{Y}_i \boldsymbol{g}_{1,L} \\ &= \sqrt{\frac{2}{3}} \left(\bar{Y}_i(2,\cdot) - \bar{Y}_i(0,\cdot) \right) \propto \bar{Y}_i(2,\cdot) - \bar{Y}_i(0,\cdot) \end{aligned}$$

and

$$\begin{aligned} \theta_{i2,L} &= \frac{Y_i(0,2) + Y_i(1,2) + Y_i(2,2)}{\sqrt{6}} - \frac{Y_i(0,0) + Y_i(1,0) + Y_i(2,0)}{\sqrt{6}} = \boldsymbol{Y}_i \boldsymbol{g}_{2,L} \\ &= \sqrt{\frac{2}{3}} \left(\bar{Y}_i(\cdot,2) - \bar{Y}_i(\cdot,0) \right) \propto \bar{Y}_i(\cdot,2) - \bar{Y}_i(\cdot,0). \end{aligned}$$

Likewise, the quadratic main effects of factors 1 and 2 are

$$\begin{aligned}\theta_{i1,Q} &= \frac{Y_i(0,0)+Y_i(0,1)+Y_i(0,2)+Y_i(2,0)+Y_i(2,1)+Y_i(2,2)}{\sqrt{18}} \\ &\qquad -\frac{Y_i(1,0)+Y_i(1,1)+Y_i(1,2)}{\sqrt{18}} = \boldsymbol{Y}_i\boldsymbol{g}_{1,Q} \\ &\propto \bar{Y}_i(2,\cdot)-2\bar{Y}_i(1,\cdot)+\bar{Y}_i(0,\cdot),\end{aligned}$$

and

$$\begin{aligned}\theta_{i2,Q} &= \frac{Y_i(0,0)+Y_i(1,0)+Y_i(2,0)+Y_i(0,2)+Y_i(1,2)+Y_i(2,2)}{\sqrt{18}} \\ &\qquad -\frac{Y_i(0,1)+Y_i(1,1)+Y_i(2,1)}{\sqrt{18}} = \boldsymbol{Y}_i\boldsymbol{g}_{2,Q} \\ &\propto \bar{Y}_i(\cdot,2)-2\bar{Y}_i(\cdot,1)+\bar{Y}_i(\cdot,0).\end{aligned}$$

Finally, the two factor interaction can be split up into the two way interaction of the linear and quadratic terms of factor 1 with the linear and quadratic terms of factor 2, and re-normalizing each of these. This corresponds to defining the following 4 components of the two factor interaction:

$$\begin{aligned}\theta_{i3,LL} &= \frac{Y_i(2,2)+Y_i(0,0)}{2}-\frac{Y_i(0,2)+Y_i(2,0)}{2} = \boldsymbol{Y}_i\boldsymbol{g}_{3,LL}, \\ \theta_{i3,LQ} &= \frac{Y_i(2,2)-2Y_i(2,1)+Y_i(2,0)}{\sqrt{12}}-\frac{Y_i(0,0)-2Y_i(0,1)+Y_i(0,0)}{\sqrt{12}} = \boldsymbol{Y}_i\boldsymbol{g}_{3,LQ}, \\ \theta_{i3,QL} &= \frac{Y_i(2,2)-2Y_i(1,2)+Y_i(0,2)}{\sqrt{12}}-\frac{Y_i(0,0)-2Y_i(1,0)+Y_i(2,0)}{\sqrt{12}} = \boldsymbol{Y}_i\boldsymbol{g}_{3,QL}, \\ \theta_{i3,QQ} &= \frac{Y_i(2,2)-2Y_i(2,1)+Y_i(2,0)}{6}-2\frac{Y_i(1,0)-2Y_i(1,1)+Y_i(1,0)}{6} \\ &\qquad +\frac{Y_i(0,0)-2Y_i(0,1)+Y_i(0,0)}{6} = \boldsymbol{Y}_i\boldsymbol{g}_{3,QQ}.\end{aligned}$$

Specifically, for a 3^2 full factorial design the eight $\boldsymbol{g}$ vectors that correspond to the linear and quadratic terms of the main effects of factors 1 and 2 respectively are:

$$\boldsymbol{g}_{1,L} = \frac{1}{\sqrt{6}} \begin{pmatrix} -1 \\ -1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \boldsymbol{g}_{1,Q} = \frac{1}{\sqrt{18}} \begin{pmatrix} 1 \\ 1 \\ 1 \\ -2 \\ -2 \\ -2 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \boldsymbol{g}_{2,L} = \frac{1}{\sqrt{6}} \begin{pmatrix} -1 \\ 0 \\ 1 \\ -1 \\ 0 \\ 1 \\ -1 \\ 0 \\ 1 \end{pmatrix}, \boldsymbol{g}_{2,Q} = \frac{1}{\sqrt{18}} \begin{pmatrix} 1 \\ -2 \\ 1 \\ 1 \\ -2 \\ 1 \\ 1 \\ -2 \\ 1 \end{pmatrix},$$

and the two factor interaction can be split into:

$$\boldsymbol{g}_{j,ll} = \frac{1}{2} \begin{pmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 1 \end{pmatrix}, \boldsymbol{g}_{j,lq} = \frac{1}{\sqrt{12}} \begin{pmatrix} -1 \\ 2 \\ -1 \\ 0 \\ 0 \\ 0 \\ 1 \\ -2 \\ 1 \end{pmatrix}, \boldsymbol{g}_{2,ql} = \frac{1}{\sqrt{12}} \begin{pmatrix} -1 \\ 0 \\ 1 \\ 2 \\ 0 \\ -2 \\ -1 \\ 0 \\ 1 \end{pmatrix}, \boldsymbol{g}_{j,qq} = \frac{1}{\sqrt{36}} \begin{pmatrix} 1 \\ -2 \\ 1 \\ -2 \\ 4 \\ -2 \\ 1 \\ -2 \\ 1 \end{pmatrix}.$$

Following the format presented in 1.1, the potential outcomes table 7.1 can be written up in terms of the unit level factorial effects of interest, broken up into their linear and quadratic terms. For every unit we have nine (3^2) unit level estimands: an average and eight factorial effects expressed in the LQ system. That is, let $\theta_{i0} = \frac{1}{3}\mathbf{1}_9$, where $\mathbf{1}_9$ is the 9 dimensional column vector of ones, we define $\boldsymbol{\theta}$ to be the vector of normalized factorial effects to be

$$\boldsymbol{\theta}_i = (\theta_{i0}, \theta_{i1,l}, \theta_{i1,q}, \theta_{i2,l}, \theta_{i2,q}, \theta_{i3,ll}, \theta_{i2,lq}, \theta_{i2,ql}, \theta_{i2,qq})\,.$$

Then the one to one relationship between these is again summarized by $\boldsymbol{G}$ by

$$\boldsymbol{\theta}_i = \boldsymbol{Y}_i\boldsymbol{G},$$

where $\boldsymbol{G}$ is an orthogonal matrix (i.e., $\boldsymbol{G}'\boldsymbol{G} = Diag(9,1,1,1,1,1,1,1,1)$). Specifically,

$$\boldsymbol{G} = \begin{pmatrix}
1 & -1/\sqrt{6} & 1/\sqrt{18} & -1/\sqrt{6} & 1/\sqrt{18} & 1/2 & -1/\sqrt{12} & -1\sqrt{12} & 1\sqrt{36} \\
1 & -1/\sqrt{6} & 1/\sqrt{18} & 0 & -2/\sqrt{18} & 0 & 2\sqrt{12} & 0 & -2\sqrt{36} \\
1 & -1/\sqrt{6} & 1/\sqrt{18} & 1/\sqrt{6} & 1/\sqrt{18} & -1/2 & -1\sqrt{12} & 1\sqrt{12} & 1\sqrt{36} \\
1 & 0 & -2/\sqrt{18} & -1/\sqrt{6} & 1/\sqrt{18} & 0 & 0 & 2\sqrt{12} & -2\sqrt{36} \\
1 & 0 & -2/\sqrt{18} & 0 & -2/\sqrt{18} & 0 & 0 & 0 & 4\sqrt{36} \\
1 & 0 & -2/\sqrt{18} & 1/\sqrt{6} & 1/\sqrt{18} & 0 & 0 & -2\sqrt{12} & -2\sqrt{36} \\
1 & 1/\sqrt{6} & 1/\sqrt{18} & -1/\sqrt{6} & 1/\sqrt{18} & -1/2 & 1\sqrt{12} & -1\sqrt{12} & 1\sqrt{36} \\
1 & 1/\sqrt{6} & 1/\sqrt{18} & 0 & -2/\sqrt{18} & 0 & -2\sqrt{12} & 0 & -2\sqrt{36} \\
1 & 1/\sqrt{6} & 1/\sqrt{18} & 1/\sqrt{6} & 1/\sqrt{18} & 1/2 & 1\sqrt{12} & 1\sqrt{12} & 1\sqrt{36}
\end{pmatrix}.$$

In general, when referring to a 3^K design with K factors, all factorial effects involving $m \in \{1,2,\ldots,K\}$ factors have 2^m degrees of freedom associated to them. Using the LQ

system each factorial effect can be broken up into 2^m orthogonal contrasts (depending on the number of factors involved in it) that correspond to the products of some combination of linear and quadratic effects, one for each of the m factors involved. For example, each two factor interaction is associated to $4 = 2^2$ degrees of freedom, the four mutually orthogonal contrasts are related to the four two way interactions between the linear and quadratic contrasts of both factors involved like in the 3^2 design used to illustrate above. In the r replicate case the norm of the linear and quadratic components of the main effects are $\sqrt{2 \cdot (3^{k-1}) \cdot r}$ and $\sqrt{6 \cdot (3^{k-1}) \cdot r}$, respectively. For any such m factor interaction, let l and q denote the number of linear and quadratic factors involved (such that $l + q = m$), then the norm of this interaction is $\sqrt{2^l \cdot 6^q \cdot 3^{K-m} \cdot r}$ for every $m \leq K$.

Again, consistent with the traditional definition of causal effects in the factorial design literature, the causal estimands at the *population level* could be the averages of the unit level factorial effects. These quantities, denoted by $\theta_{1,L}, \theta_{1,Q}, \theta_{2,L}, \theta_{2,Q}, \theta_{3,LL}, \theta_{3,LQ}, \theta_{3,QL}$ and $\theta_{3,QQ}$, are the population level main effects of each treatment factor decomposed into its linear and quadratic components and the interactions of these across factors, and can be expressed in terms of potential outcomes as:

$$\theta_{j,b} = \frac{\sum_{i=1}^{N} \theta_{i,j,b}}{N} = \overline{\boldsymbol{Y}} \boldsymbol{g}_j, \quad j = 1, 2, 3, \quad b \in \mathcal{B}, \tag{7.1}$$

where

$$\bar{\boldsymbol{Y}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{Y}_i, \tag{7.2}$$

and $\mathcal{B}$ is the set of all interactions expressed in the LQ system format. For example, for the 3^2 case

$$\mathcal{B} = \{L, Q, LL, LQ, QL, QQ\}.$$

After setting up the problem this way we are practically in the same scenario to that discussed in previous chapters: the two level factorial designs. Now, in the implementation of the SI and SPPC methods the only practical difference is that we now have no further scaling in the definition of θ, unlike the 2-level case when all the factorial effects were scaled by 1/2. However, the symmetry is lost in the sense that in the 2 level case all factorial effects were differences between two halves of the potential outcomes, and that ceases to be true for the LQ system in three level designs. This does not affect the procedure for filling in the missing outcomes stated in Section 3.1.2 and the proposed methods still work in this setting. But, this loss of symmetry in the problem is reflected in the fact that the reference distribution for the randomization test will no longer be the same for all estimated factorial effects.

7.1.3 Seat Belt Experiment (Wu and Hamada, 2009)

We now illustrate the method in an example that is a simplification of a 3^{4-1} fractional factorial experiment (see Section 7.2) found in Wu and Hamada (2009). Consider an experiment that aims to study the effect of four factors on the pull strength of seat belts. Each of these four factors are at three levels. For illustration purposes, we focus on three factors and pretend we have a 3^3 full factorial design with three replicates. Table 7.1.3 displays the treatment factors and the three levels defined for each of these.

Table 7.2: Definition of treatment factors and levels

	Level		
Factor	0	1	2
1. Pressure(psi)	1100	1400	1700
2. Die flat (mm)	10.0	10.2	10.4
3. Crimp length (mm)	18	23	27

The OLS results are shown in Table 7.3 for all factorial effects. The corrections for multiple comparisons are also made. In this case we display the factorial effects identified as active using Bonferroni and FDR corrected OLS, as well as S-PPC. In this case, Bonferroni and S-PPC lead to the identification of the same factors as active, namely the linear components of factors 1 and 3, together with the linear-linear interaction between factors 1 and 3.

Table 7.3: Results displayed for the simplified Seat Belt experiment (Wu and Hamada, 2009). The results using procedures that control for multiple comparisons are displayed in the last three columns and mark with a "yes" those rows that correspond to effects identified as active by each method.

	Estimate	Std. Error	t value	Pr(>|t|)	Bonferroni	S-PPC	FDR
(Intercept)	6223.4074	50.0434	124.3602	< 0.0001	-	-	-
X1.L	3346.7364	260.0331	12.8704	< 0.0001	yes	yes	yes
X3.L	-1768.6312	260.0331	-6.8016	< 0.0001	yes	yes	yes
X5.LL	1087.6317	260.0331	4.1827	0.0001	yes	yes	yes
X4.QQ	762.1986	260.0331	2.9312	0.0049			yes
X7.LQQ	595.0304	260.0331	2.2883	0.0261			
X7.LLL	-573.8172	260.0331	-2.2067	0.0316			
X1.Q	-569.5064	260.0331	-2.1901	0.0329			
X7.QQL	566.1961	260.0331	2.1774	0.0338			
X2.L	536.0655	260.0331	2.0615	0.0441			
X4.QL	509.7222	260.0331	1.9602	0.0551			
X4.LL	-503.5457	260.0331	-1.9365	0.0581			
X7.QLQ	497.3711	260.0331	1.9127	0.0611			
X7.QLL	407.7720	260.0331	1.5682	0.1227			
X7.LQL	384.5018	260.0331	1.4787	0.1450			
X7.QQQ	-367.4915	260.0331	-1.4132	0.1633			
X5.QQ	-323.2199	260.0331	-1.2430	0.2192			
X6.LL	292.3317	260.0331	1.1242	0.2659			
X6.QQ	-237.4834	260.0331	-0.9133	0.3652			
X3.Q	-237.1923	260.0331	-0.9122	0.3657			
X7.LLQ	-168.2663	260.0331	-0.6471	0.5203			
X2.Q	-162.0746	260.0331	-0.6233	0.5357			
X6.LQ	-63.7778	260.0331	-0.2453	0.8072			
X5.LQ	-63.5000	260.0331	-0.2442	0.8080			
X4.LQ	-46.5000	260.0331	-0.1788	0.8587			
X6.QL	43.2222	260.0331	0.1662	0.8686			
X5.QL	-2.6111	260.0331	-0.0100	0.9920			

7.2 Fractional Factorial Designs

Fractional Factorial Designs are used when there are limitations to the number of runs in an experiment (for example, for economical reasons) and there are reasons to believe that the higher order effects are not important (effect hierarchy and effect heredity principles). In these cases, fractional factorials allow us to cut down the number of runs and still test those effects which are believed to be important. However, this comes at a price. Fractional factorial designs split all the effects that would be relevant in a full factorial design into groups that cannot be disentangled (i.e., they are indistinguishable) in the analysis. Each such group is called an aliasing group, and the properties of the fractional factorial design are determined by its aliasing structure. The factorial effects of interest are made estimable by making strong assumptions about the higher order effects in each aliasing group.

A 2^{K-p} fractional factorial design has K factors, each with 2 levels, 2^{K-p} runs, and is a $1/2^p$-th fraction of a 2^K full factorial design. The fraction is defined by p independent *defining words*. The group formed by these p words is called the *defining contrast subgroup*. The defining contrast subgroup has 2^{p-1} words plus the identity element I.

7.2.1 How is this different from what we've previously done?

Again, for illustration purposes let us think of the simplest fractional factorial design: a 2^{3-1} unreplicated design. In this case, we have 4 units and 3 treatment factors, each at 2 levels. Hence, only a 1/2 fraction of the treatment combinations in the full 2^3 design (eight) are assigned to the experimental units. In a 2^{3-1} design there is one generator, for example $3 = 1*2$ which means that the main effect of factor 3 is aliased with the two factor interaction

of factors 1 and 2. Another consequence of this design is that the three factor interaction is aliased with the mean. The aliasing structure of this design is:

$$1 = 2 * 3, 2 = 1 * 3, 3 = 1 * 2.$$

Meaning, for example, that in the design matrix the column that correspond to factor 1 is the same as the one that corresponds to the interaction of factors 2 and 3, such that $\hat{\theta}_1 = \hat{\theta}_{2*3}$ where $2 * 3 = 6$ in the notation we used in previous chapters.

Note that an implicit assumption being made when using fractional factorial designs is that the factorial effects aliased within a group will not cancel each other out. If that were not the case then such a design will be unreliable for the assessment of the treatment factors.

An undeniable feature of fractional factorials is that, because of the perfect aliasing, there is no information that can be obtained from the data that allows us to disentangle the effects within an aliasing group. A way to think about this problem is that each aliased group is related to a parameter in the model, whose prior distribution is a mixture of the effects in the group. However, without any additional prior information there is nothing we can say about the individual effects in the group. That is why traditionally, the assumption of inert higher order interactions is made. Specifically, in the 2^{3-1} design discussed earlier this assumption implies that only the main effects are believed to be active. Without such an assumption, this design can still be useful if it is reasonable to assume that the effects within an aliased group are not canceling each other out and there are resources to run follow up experiments to disentangle the effects. Some options for follow up experiments are adding orthogonal runs or to use the fold over technique.

Details of fractional factorial designs, their properties, construction and analysis can

be found in most standard textbooks on experimental design. These include Kempthorne (1952), Kempthorne and Hinkelmann (1984), Wu and Hamada (2009),Box et al. (2005), Montgomery et al. (1984) and Mukerjee and Wu (2006).

An interesting issue that arises in this context is the specification of the potential outcomes that we are interested in. That is, how do we visualize the science table for this problem and define the estimands of interest? Our initial thoughts on this matter have lead us to believe that there are two main options, which we describe in more detail below.

7.2.2 Random Selection of Fraction

Maybe we are interested in all eight potential outcomes. Therefore we should consider that there is positive probability of observing any of each unit's eight potential outcomes. This is only possible if we can think of the problem as a two level randomization in which we first randomly choose the fraction of treatment combinations to apply in the experiment from the set of plausible fractions, and then randomly assign the subset of treatment combinations present in the selected fraction to the units at hand. A sketch of the science table for this setting is shown in Table 7.4.

Table 7.4: The *science* for the full experiment.

Unit	Potential outcome for treatment combination							
(i)	$(1,1,1)$	$(1,1,-1)$	$(1,-1,1)$	$(1,-1,-1)$	$(-1,1,1)$	$(-1,1,-1)$	$(-1,-1,1)$	$(-1,-1,-1)$
1	$Y_1(1,1,1)$	$Y_1(1,1,-1)$	$Y_1(1,-1,1)$	$Y_1(1,-1,-1)$	$Y_1(-1,1,1)$	$Y_1(-1,1,-1)$	$Y_1(-1,-1,1)$	$Y_1(-1,-1,-1)$
2	$Y_2(1,1,1)$	$Y_2(1,1,-1)$	$Y_2(1,-1,1)$	$Y_2(1,-1,-1)$	$Y_2(-1,1,1)$	$Y_2(-1,1,-1)$	$Y_2(-1,-1,1)$	$Y_2(-1,-1,-1)$
3	$Y_3(1,1,1)$	$Y_3(1,1,-1)$	$Y_3(1,-1,1)$	$Y_3(1,-1,-1)$	$Y_3(-1,1,1)$	$Y_3(-1,1,-1)$	$Y_3(-1,-1,1)$	$Y_3(-1,-1,-1)$
4	$Y_3(1,1,1)$	$Y_4(1,1,-1)$	$Y_4(1,-1,1)$	$Y_4(1,-1,-1)$	$Y_4(-1,1,1)$	$Y_4(-1,1,-1)$	$Y_4(-1,-1,1)$	$Y_4(-1,-1,-1)$
Average	$\overline{Y}(1,1,1)$	$\overline{Y}(1,1,-1)$	$\overline{Y}(1,-1,1)$	$\overline{Y}(1,-1,-1)$	$\overline{Y}(-1,1,1)$	$\overline{Y}(-1,1,-1)$	$\overline{Y}(-1,-1,1)$	$\overline{Y}(-1,-1,-1)$

Now, we only get to observe an outcome for four of the eight potential treatment combinations. Clearly, stronger prior information or assumptions have to be made to be able to

fill in all the missing outcomes by only observing outcomes on a fraction of the treatment combinations. We should keep in mind that Chipman et al. (1997) (see Section 7.2.4) points out the inadequacy of the vague prior used in the S-PPC when the number of factorial effects (or potential outcomes as we've seen previously) is higher than the number of units because it will likely result in too many factors identified as active in the model.

7.2.3 Deterministic Selection of Fraction

Alternatively we can deterministically select the fraction of treatment combinations that can be possibly assigned to the experimental units. This occurs when there are certain combinations of factors that are intentionally avoided, for example having all factors on level 1 might be too aggressive for the experimental units, or receiving none of the treatment factors (i.e., all are at level -1) might be unacceptable for the units. Therefore we restrict the potential outcomes that can possibly be observed for any unit to only the subset of plausible treatment combinations in the selected fraction. In this case, the science table consists of fewer rows because for each unit there are only 2^{k-p} potential outcomes that could possibly be observed. For the 2^{3-1} design these tables are shown in Tables 7.5 and 7.6.

Table 7.5: The fraction of the potential outcomes that corresponds to the defining relation $1 = 2 * 3$.

Unit	Potential outcome for treatment combination				Unit-level factorial effects			
(i)	$(1,1,1)$	$(1,-1,-1)$	$(-1,1,-1)$	$(-1,-1,1)$	$\theta_{i,0}$	$\theta_{i,1}$	$\theta_{i,2}$	$\theta_{i,3}$
1	$Y_1(1,1,1)$	$Y_1(1,-1,-1)$	$Y_1(-1,1,-1)$	$Y_1(-1,-1,1)$	$\theta_{1,0}$	$\theta_{1,1}$	$\theta_{1,2}$	$\theta_{1,3}$
2	$Y_2(1,1,1)$	$Y_2(1,-1,-1)$	$Y_2(-1,1,-1)$	$Y_2(-1,-1,1)$	$\theta_{2,0}$	$\theta_{2,1}$	$\theta_{2,2}$	$\theta_{2,3}$
3	$Y_3(1,1,1)$	$Y_3(1,-1,-1)$	$Y_3(-1,1,-1)$	$Y_3(-1,-1,1)$	$\theta_{3,0}$	$\theta_{3,1}$	$\theta_{3,2}$	$\theta_{3,3}$
4	$Y_4(1,1,1)$	$Y_4(1,-1,-1)$	$Y_4(-1,1,-1)$	$Y_4(-1,-1,1)$	$\theta_{4,0}$	$\theta_{4,1}$	$\theta_{4,2}$	$\theta_{4,3}$

Perhaps in many cases, once that fraction is deterministically chosen, we can think of the

Table 7.6: The fraction of the potential outcomes that corresponds to the defining relation $1 = -2 * 3$.

Unit	Potential outcome for treatment combination				Unit-level factorial effects			
(i)	$(1,1,-1)$	$(1,-1,1)$	$(-1,1,1)$	$(-1,-1,-1)$	$\theta_{i,0}$	$\theta_{i,1}$	$\theta_{i,2}$	$\theta_{i,3}$
1	$Y_1(1,1,-1)$	$Y_1(1,-1,1)$	$Y_1(-1,1,1)$	$Y_1(-1,-1,-1)$	$\theta_{1,0}$	$\theta_{1,1}$	$\theta_{1,2}$	$\theta_{1,3}$
2	$Y_2(1,1,-1)$	$Y_2(1,-1,1)$	$Y_2(-1,1,1)$	$Y_2(-1,-1,-1)$	$\theta_{2,0}$	$\theta_{2,1}$	$\theta_{2,2}$	$\theta_{2,3}$
3	$Y_3(1,1,-1)$	$Y_3(1,-1,1)$	$Y_3(-1,1,1)$	$Y_3(-1,-1,-1)$	$\theta_{3,0}$	$\theta_{3,1}$	$\theta_{3,2}$	$\theta_{3,3}$
4	$Y_4(1,1,-1)$	$Y_4(1,-1,1)$	$Y_4(-1,1,1)$	$Y_4(-1,-1,-1)$	$\theta_{4,0}$	$\theta_{4,1}$	$\theta_{4,2}$	$\theta_{4,3}$

full potential outcome table displayed in 7.4 as the result of a combination of an observational and an experimental study where the assignment mechanism to chose a fraction is unknown but the assignment mechanism for the second level randomization is completely known.

Now that we have reviewed the concept of aliasing, we take a deeper plunge into the Bayesian method proposed in Chipman et al. (1997). This procedure takes advantage of the hierarchical structure that is assumed in the effect hierarchy and effect heredity principles in the design of experiments literature. Going forward, it would be appealing to explore the use of this modeling approach to fill in the potential outcomes table and perform a randomization based posterior predictive check. It would also be relevant for the considerations in follow up experiments.

7.2.4 A closer look at Chipman et al. (1997)

Chipman et al. (1997) proposed a model search procedure done in a Bayesian framework when complex aliasing is present. Some of the advantages of using the Bayesian framework stated by the authors (Wu and Hamada, 2009, page 363) are that the Gibbs sampler necessary to search across the models is easy to implement and its computational efficiency (they wrote: "the search moves from one model to another in the model space and visits the most likely

models the most often"), as well as the fact that the effect sparsity and effect heredity principles can be easily incorporated through the priors.

They consider the Bayesian variable selection procedure proposed by George and McCulloch (1993) in the classical general linear model setting with

$$\boldsymbol{Y}^{\boldsymbol{obs}} = \boldsymbol{X}\boldsymbol{\theta}' + \boldsymbol{\epsilon},$$

where $\boldsymbol{X}$ is the model matrix and $\boldsymbol{\theta} \in \mathbb{R}^p$ is the population level row vector of factorial effects, as we defined in the previous chapters, and $\boldsymbol{\eta} \sim MN(\mathbf{0}, \sigma^2 \boldsymbol{I}_{N\times N})$. Their contribution was to define a vector $\boldsymbol{\delta}$ of 0's and 1's to indicate the *significance* of the effects. The j-th entry $\delta_j = 0$ indicates that θ_j is small and therefore not significant. Alternatively, if $\delta_j = 1$ that indicates that θ_j is large and therefore significant. For variable selection, the parameter vector is $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma)$. The posterior of $\boldsymbol{\delta}$ is of particular interest because it specifies the model. In this context, the models with the highest posterior probabilities are identified as important. The full Bayesian formulation is obtained by specifying the priors. A normal mixture prior for $\boldsymbol{\theta}$ is used (note that our proposal is a degenerate version of this that imposes point mass prior if $\delta_j = 0$):

$$p(\theta_j) = \begin{cases} N(0, \sigma^2\tau_j^2) & \text{if } \delta_j = 0, \\ N(0, \sigma^2(c_j\tau_j)^2) & \text{if } \delta_j = 1. \end{cases}$$

As it was mentioned in Chapter 4, it is through the variance that this method distinguishes between active and inactive effects. In that regard, the constants τ_j and c_j are chosen to represent a "small" effect and how large a "large" effect should be. Specifically, when $\delta_j = 0$, the constant τ_j helps tighten θ_j around zero and therefore does not have a large effect. The constants c_j are chosen to be much larger than 1 to indicate the possibility of a large effect

when $\delta_j = 1$. Moreover, an inverse gamma distribution is used for the prior of σ^2, that is

$$p(\sigma^2) \sim Inv - gamma(\nu/2, \nu\lambda/2).$$

Note that this is different from the prior that we assumed in the S-PPC methodology. An independence prior on $\boldsymbol{\delta}$, where p_j is the probability that $\delta_j = 1$, looks like this

$$p(\boldsymbol{\delta}) = \prod_{j=1}^{p+1} p_j^{\delta_j}(1-p_j)^{1-\delta_j}.$$

However, this prior does not take into account the effect heredity principle therefore the hierarchical priors defined in Chipman (1996) were used. This kind of prior breaks up the the probability $p(\boldsymbol{\delta})$ into the conditionals allowing for the incorporation of the desired DOE principles. For example, recalling that in a 2^2 full factorial design $j = 3$ corresponds to the interaction effect, then we could break $p(\boldsymbol{\delta})$ up as follows

$$p(\boldsymbol{\delta}) = p(\delta_1)p(\delta_2)p(\delta_3|\delta_1, \delta_2).$$

There is an implicit inheritance assumption because the significance of the term depends only on those terms from which it is formed. In the general case, there is another assumption being made which refers to the conditional independence of interactions of m factors given

all lower order terms. The inheritance assumption is explicitly made in the prior given to

$$p(\delta_3|\delta_1,\delta_2) = \begin{cases} p_{00} & \text{if } (\delta_1,\delta_2) = (0,0), \\ p_{01} & \text{if } (\delta_1,\delta_2) = (0,1), \\ p_{10} & \text{if } (\delta_1,\delta_2) = (1,0), \\ p_{11} & \text{if } (\delta_1,\delta_2) = (1,1). \end{cases}$$

In Wu and Hamada (2009), the authors provide examples of how to choose these values. They suggest $0.001 \approx p_{00} < p_{01} = p_{10} \approx 0.01 < p_{11} \approx 0.25$ representing weak heredity. Related to infusing the prior with the effect heredity principle, they assume the significance of a term depends only on its parents, defined as "those terms of the next smallest order which can form the original term when multiplied by a main effect". This assumption is called the *immediate inheritance principle*.

The traditional approach to evaluate the posterior distribution of $\boldsymbol{\eta}$ is to implement a Gibbs sampler using full conditional distributions of the parameters. In this problem the joint distribution of $\boldsymbol{Y^{obs}}$ and η is

$$p(\boldsymbol{\eta}, \boldsymbol{Y^{obs}}) = p(\boldsymbol{Y^{obs}}|\boldsymbol{\theta}, \boldsymbol{\sigma^2})p(\boldsymbol{\theta}|\sigma^2, \boldsymbol{\delta})p(\boldsymbol{\sigma^2})p(\boldsymbol{\delta}).$$

From here the full conditionals can be derived and consist of a multivariate normal draw for $\boldsymbol{\theta}|\sigma^2, \boldsymbol{\delta}$, an inverse gamma draw for $\boldsymbol{\sigma^2}|\boldsymbol{\theta}, \boldsymbol{\delta}$ and $p+1$ Bernoulli draws for $\delta_j|\boldsymbol{\theta}, \sigma^2, \{\delta_v\}_{v\neq j}$. One of the attractive features of this approach, as the authors mention, is that it has a straightforward extension to nonnormal data.

Another feature of this approach is that it involves constants $\boldsymbol{c}, \boldsymbol{\tau}, \nu, \lambda$ which the authors

view mainly as tuning constants. In agreement with Box and Meyer (1986), they suggest $c_j = 10$. They report high sensitivity of the posterior to the choice of τ and argue it is an advantage of the procedure because "the appropriateness of a given value may be judged by the models (or posterior model probabilities) it generates" and the experimenter is more likely to have good prior knowledge on the number of truly active effects than the values of $\boldsymbol{\tau}$. Their suggestion for τ_j follows the observations of George and McCulloch (1993) and is

$$\tau_j = \frac{SE(\boldsymbol{Y}^{\boldsymbol{obs}})/5}{3(max(\boldsymbol{X}_j) - min(\boldsymbol{X}_j))}.$$

Here, $\boldsymbol{X}_j$ refers to the column in the model matrix that is associated to the factorial effect j (see equation 6.1). We are using $\boldsymbol{X}$ and not $\boldsymbol{G}$ to use the notation defined in Chapter 6 for the replicate case that still works for the unreplicated case if we consider $r = 1$. Recall that the S-PPC procedure uses an improper prior for σ. However, the author argue that in this context it is inappropriate because it allows it to be too close to zero. They recommend a proper prior specially for cases where the number of covariates (in our case factorial effects) is higher than the number of units which will likely result in a lot factors in the model. Whereas the choice of such a proper prior may play in important role in unbalanced designs where partial aliasing is present, it may not be very relevant for the designs that we're considering in this thesis.

7.2.5 Implementing S-PPC for fractional factorial designs

It is possible to apply the S-PPC methodology to fractional factorial designs by thinking about the science in this second context. Everything that was mentioned for both the

replicated and unreplicated cases would follow where each of the factorial effects in $\boldsymbol{\theta}$ is a hidden mixture of the factorial effects defined in a full factorial design. To be able to draw conclusions about the particular effects in each aliased group we would need to use either prior information, design of experiments principles or follow up runs to disentangle between them. The example discussed in the next section uses this approach for the analysis of fractional factorial designs. The procedure is basically the same as analyzing a full factorial design for $f = K - p$ treatment "factors", but where instead of individual factors we are assessing aliased groups of factorial effects.

7.3 A Case Study: Directed Differentiation of Stem Cells to Pancreatic β Cells

We now exemplify the use of S-PPC in a large fractional factorial design that is part of an ongoing collaborative project. The objective of this project is to design and analyze multifactor experiments to identify novel pathway interactions in the directed differentiation of embryonic stem cells into pancreatic β cells, a goal that is relevant for potential treatment alternatives for type 1 diabetes.

7.4 Background

Stem cells can differentiate into diverse specialized cell types, and can self-renew to produce more stem cells. There are two main types of stem cells: embryonic and adult stem cells. ***Embryonic*** stem cells are formed in early gestation, as opposed to *adult* stem cells

which are found in various tissues. Also, embryonic stem cells have the ability to form nearly all cell types in the body, including the β cells that diabetics lack. The main function of β cells (the target cells), which are found in the pancreas, is to store and release insulin. Insulin is a hormone that brings about effects which reduce blood glucose concentration.

Type 1 diabetes is an autoimmune disease that consists in the destruction of insulin-producing β cells of the pancreas. The subsequent lack of insulin leads to increased blood and urine glucose. Eventually, type 1 diabetes is fatal unless treated. The current treatment is insulin replacement therapy, which *does not* recapitulate β cell function. The success of this directed differentiation would potentially help in finding an alternative treatment for type 1 diabetes. However, the exploration of these molecular pathways leading to effective differentiation of stem cells into β cells is very challenging. There are 30 small molecule pathway modulators that could be used (in some order) in these multi-stage differentiation process to produce the desired outcome.

The first experiment, reported in Zemplenyi (2013), was exploring eight of these compounds. It was run following a 3^{8-3} fractional factorial design suggested by Hongquan Xu in Xu (2005). Three protein related responses were the focus of this experiment: C-peptide, Nkx6.1 and their co-expression. These responses were chosen from a wide range of proteins that pancreatic β cells express to determine a small subset that define the β cell type proteins in the sense that they enable β cells to produce insulin in response to high glucose levels. C-peptide is a component of the proinsulin protein from which mature insulin is produced. Nkx6.1 is a member of a class of proteins called *transcription factors*, which help transcribe genes. As a transcription factor, Nkx6.1 helps regulate which genes are being expressed in a cell and can thereby direct and maintain β cell type. Because cells with high levels of Nkx6.1

are more likely to differentiate toward β cell lineage, one of the goals of our experiment was to find compounds that induce Nkx6.1 expression. Of the eight compounds used in this experiment, the main effects of compounds ALK 5 inhibitor, DAPT, and DEAB indicated that higher concentrations of these lead to low levels of Nkx6.1 expression and high levels of C-peptide expression. Whereas the main effect of ISX9 indicates that a higher concentration leads to low levels of both Nkx6.1 and C-peptide levels. Other higher order factorial effects involving these modulators were found significant. However, no treatment combination led to particularly high levels of Nkx6.1 and C-peptide co-expression. Nevertheless, the results of this experiment suggested that further investigation with cyclopamine may increase levels of co-expression. Detailed explanations and results can be found in Harvard's undergraduate thesis Zemplenyi (2013).

7.4.1 The Second Design

The goal for this second round of experiments was to test a higher number of compounds. As in the first design, we are interested in the third order and lower interactions as well as the main effects. In that regard, the advantages of 2 level design relative to the previously used 3 level designs was discussed. A 2-level design was preferred because no benefit was obtained in the first experiment by using three levels and it would allow the testing of more treatment factors. The number of runs (wells) was restricted to be 2000 ± 200. There was a need for replication, 2 or 3 replicates were thought to be acceptable. Our collaborators expressed interest in testing between 10 and 30 compounds. Hence, the design chosen was a 2^{24-14} fractional factorial design of resolution IV, with two replicates for each treatment combination. Let the modulators be denoted by A,B,..., Z (excluding X and I to avoid

notational ambiguity), then the generators of the design are:

$$
\begin{aligned}
L &= ABCDEFG \qquad M = ABCGHJK \\
N &= ACDGH \qquad O = ACEGJ \\
P &= ADEHJ \qquad Q = -ADEFG \\
R &= -ABFHJ \qquad S = -ABDEF \\
T &= -ABCDH \qquad U = -ADEFK \\
V &= -ABCDEJK \qquad W = -ABCEG \\
Y &= -ACFGH \qquad Z = -ADGHK.
\end{aligned}
$$

The defining contrast subgroup has 2^{13} words (here 16,383) plus the identity element I.Noting that there are no words of odd number length, the word length pattern of this design is displayed in Table 7.7.

Table 7.7: Word length pattern of design used. Note that there are no words of odd number length.

word length	4	6	8	10	12	14	16	18	20
number of words	18	279	1397	3859	5283	3845	1406	273	23

As stated previously, this is a resolution IV design, which means that some two factor interactions are aliased with other two factor interactions. However, the main effects are clear (not aliased with two factor interactions or other main effects), and some are strongly clear (in addition to other main effects and two factor interactions, not aliased with three factor interactions).

A concern was that large numbers of treatment factors in the "on" setting could have a detrimental effect on the cell causing the death of more cells. To address this question we asses whether the average cell count of the wells changes (e.g. decreases) as the number of active factors increases. Figure 7.1 shows that that is not the case. The overlaid scatter plot corresponds to the total cell counts for each treatment combination associated with a certain number of factors present (i.e., the number of factors set at level 1). The black line corresponds to the OLS fit of the total number of cells on the number of factors present (set at level 1), for which a slope of zero is quite feasible.

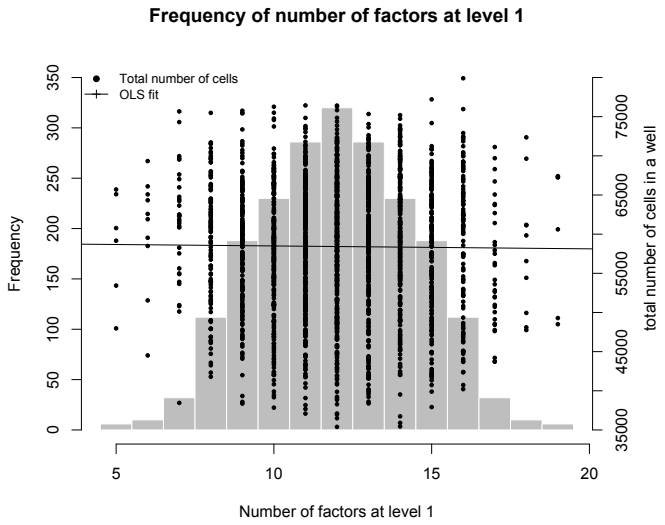


Figure 7.1: The grey histogram displayed in the background corresponds to the distribution of the number of factors at level 1 across the 1024 treatment combinations. The overlaid scatter plot corresponds to the total cell counts for each treatment combination associated with a certain number of factors present (i.e., the number of factors set at level 1 for each treatment combination). The black line corresponds to the OLS fit of the total number of cells on the number of factors present (set at level 1).

There were multiple responses of interest but, for illustration purposes, in this chapter we focus on one: the proportion of the total number of imaged cells that expressed C-Peptide. The clarification of the "imaged" cells is given because, although the care put into

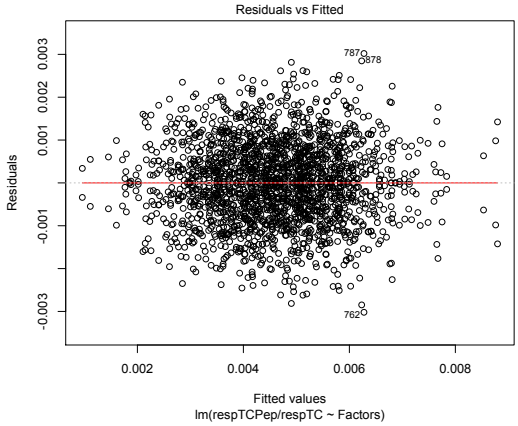
the experimental process makes it reasonable to assume that the initial number of cells in each well is close to 200,000, the imaging does not cover the entire well, and we were told that it is not reasonable to assume that the distribution of cells within each well is spatially homogenous. Nevertheless, the area covered by the imaging is comparable for each well. Therefore, the response we are focusing on is relative to the imaged portion of the well defined as:

$$Y_{C-pep} = \frac{\text{total number of imaged cells expressing C-Peptide}}{\text{total imaged cell count}}.$$

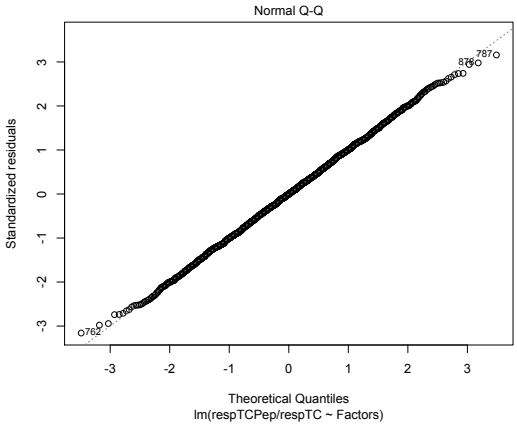
Recall that running the S-PPC for the replicated case requires the use of the matrices $\boldsymbol{G}$ and $\boldsymbol{X}$ defined in Chapter 2. Hence, in this example as well as the seat belt one (7.1.3) we use both matrices. The $\boldsymbol{X}$ matrix is used to get the posterior draws of the hyperparameters and the $\boldsymbol{G}$ is used to impute all missing outcomes at the unit level.

Figure 7.2 and Table 7.8 display the results of the experiment obtained through OLS. Table 7.8 includes all the effects deemed significant if using a 0.05 cutoff. However, this problem is a natural candidate for multiple comparisons correction given the large number of factorial effects that we are dealing with. This table also displays which factorial effects would be deemed as active using the FDR and Bonferroni multiple comparisons corrections on the OLS results. The values with "yes" denote those factorial effects that would still be significant after using either Bonferroni (a cut off of $p_{value} < 0.05/1023 \approx 0.00005$, because the number of aliased groups being tested is 1023) or FDR (using $q = 0.05$).

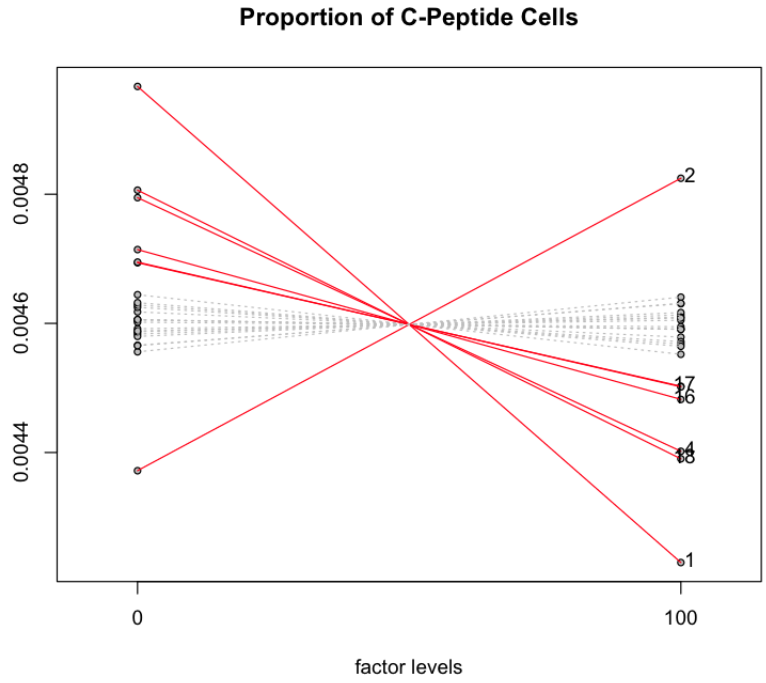
To increase efficiency when analyzing this experiment using the SPPC we decreased the number of models to be tested by reducing the set of factorial effects that form the active and inactive sets, $\mathcal{A}$ and $\mathcal{I}$ respectively. In other words, we eliminated most columns of $\boldsymbol{G}$ and $\boldsymbol{X}$. Specifically, knowing that the usual OLS does not account for multiple comparisons



(a) Fitted values vs standardized residuals



(b) QQ plot of residuals



(c) Main Effect plot

Figure 7.2: Plots for the model on the response variable: Proportion of cells that express C-Peptide.

we know that our method is more conservative than just using the α cutoff on the direct p values. Therefore, we reduced the number of models to be assessed from 1023 to 73 because using $\alpha = 0.05$ leads to only 73 factorial effects found to be significant for an $IER = 0.05$.

Note that the S-PPC identified more effects than the other two multiple comparisons methods. Together with the results shown for the seat belt experiment in Section 7.1.3 these particular examples show that there is no specific ordering in the effects found by the S-PPC and the FDR corrected OLS.

Figure 7.3 show the posterior predictive distributions and p values of the results obtained using the SPPC in this example. A couple of three factor interactions are identified as active agreeing with the the belief expressed by our collaborator that higher order effects are relevant in this problem.

The fact that the S-PPC finds more active effects than the FDR method is surprising given the results in the previous chapter, but it indicates that there is no ordering between the SPPC and the FDR methods in terms of which one identifies more active effects for a particular setting. Further exploration of what gives rise to these differences is left for future work.

Table 7.8: Results of the model on the Y_{C-Pep} response (proportion of cells that express C-Peptide). The adjusted R^2 is 0.192. For practical reasons, only effects with p values < 0.02 are displayed.

Factorial Effect	Estimate	Std. Error	t value	Pr(>|t|)	Bonferroni	S-PPC	FDR
(Intercept)	0.00460	0.00003	153.90509	< 0.00001	-	-	
X1	-0.00037	0.00003	-12.33970	< 0.00001	yes	yes	yes
X2	0.00023	0.00003	7.58014	< 0.00001	yes	yes	yes
X18	-0.00021	0.00003	-6.96266	< 0.00001	yes	yes	yes
X4	-0.00020	0.00003	-6.57477	< 0.00001	yes	yes	yes
X1.2.14	-0.00012	0.00003	-3.97882	0.00007		yes	yes
X16	-0.00012	0.00003	-3.88315	0.00011		yes	yes
X9.13	-0.00010	0.00003	-3.39746	0.00071		yes	
X1.4.11	-0.00010	0.00003	-3.32140	0.00093		yes	
X21	-0.00010	0.00003	-3.23298	0.00126			
X17	-0.00010	0.00003	-3.20214	0.00141			
X1.4.8	0.00009	0.00003	3.12049	0.00186			
X9.24	0.00009	0.00003	3.08579	0.00208			
X2.5.6	0.00009	0.00003	3.07147	0.00219			
X1.9.12	-0.00009	0.00003	-2.95438	0.00320			
X2.12	0.00009	0.00003	2.94253	0.00333			
X2.5.20	-0.00009	0.00003	-2.92037	0.00357			
X3.12	-0.00009	0.00003	-2.89905	0.00382			
X7.8.9.11	-0.00009	0.00003	-2.88235	0.00403			
X4.6	0.00008	0.00003	2.75610	0.00595			
X1.2.23	-0.00008	0.00003	-2.74926	0.00608			
X2.3.6.10	-0.00008	0.00003	-2.70231	0.00700			
X4.7.12	0.00008	0.00003	2.67504	0.00759			
X1.5.10	-0.00008	0.00003	-2.60480	0.00933			
X12.17	0.00008	0.00003	2.59163	0.00969			
X7.11.24	0.00008	0.00003	2.58620	0.00984			
X2.5.10	-0.00008	0.00003	-2.53745	0.01131			
X7.8.10.11	0.00008	0.00003	2.53372	0.01143			
X1.8.10.11	-0.00008	0.00003	-2.51929	0.01191			
X10.23	-0.00008	0.00003	-2.51410	0.01209			
X5.9.24	0.00007	0.00003	2.48497	0.01311			
X2.14	0.00007	0.00003	2.47102	0.01363			
X3.11	0.00007	0.00003	2.46002	0.01406			
X2.3.6	0.00007	0.00003	2.44755	0.01455			
X14.24	0.00007	0.00003	2.43520	0.01505			
X1.4.7.12	-0.00007	0.00003	-2.42605	0.01544			
X1.2.10.11	0.00007	0.00003	2.39775	0.01667			
X1.3.6.12	-0.00007	0.00003	-2.38053	0.01747			
X4.5.6.9	-0.00007	0.00003	-2.37791	0.01759			
X3.5.6.12	0.00007	0.00003	2.33348	0.01982			

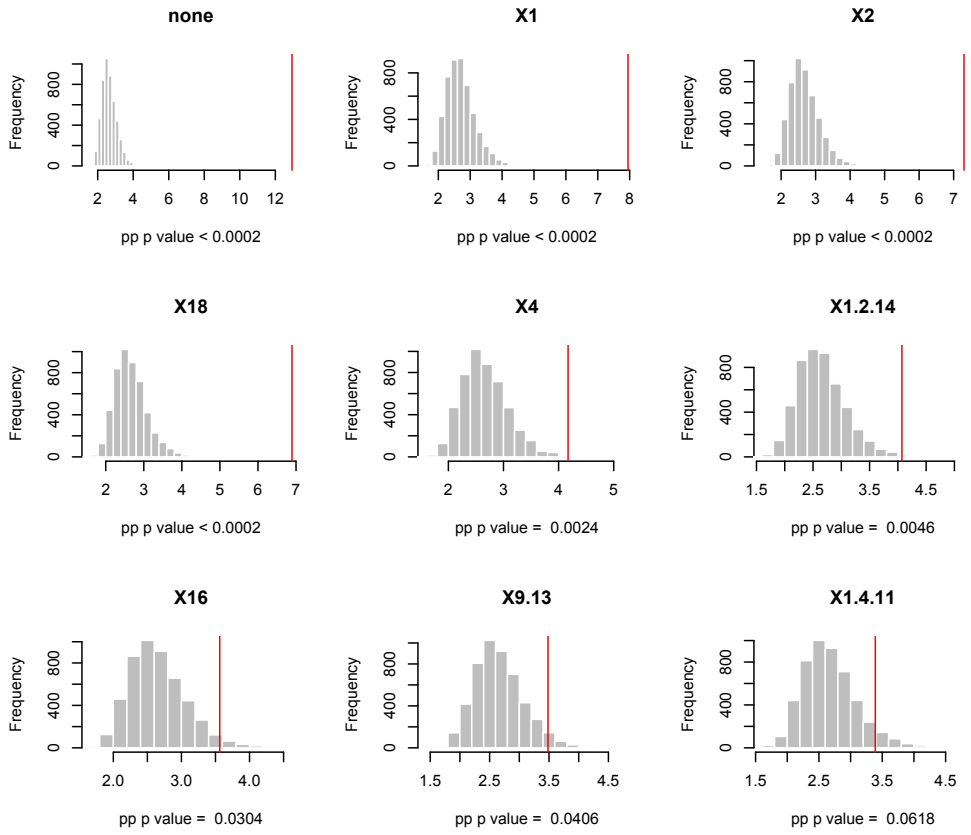


Figure 7.3: Posterior predictive distributions and p values obtained for the $\max_{\{j:j\in\mathcal{I}\}}\left|\frac{\hat{\theta}_j}{SE_{\mathcal{I}}}\right|$ discrepancy measure and step down procedure.

7.5 Conclusions

The potential outcomes framework gives a more flexible and goal-oriented approach to the design and analysis of experiments. This set up of the problem allows different definitions of estimands compared to the restriction to contrasts of means in the traditional approach, and the inclusion of prior information in the assessment of factorial effects. Our approaches are the only ones that have been proposed with the finite population in mind. However, further study of the relative performance of these methods simulating from the finite population setting is left for future work.

In order to account for the exploration of multiple factors, two sequential procedures are proposed. The single imputation approach is non parametric, assumes additivity of treatment effects and ignores the uncertainty of the point estimates of the factorial effects for subsequent steps in the sequential procedure. In contrast, the sequential posterior predictive checks require the modeling of the response variable, here assumed Normal to fit the traditional assumptions of classical screening methods.

Without any doubt the best proposal was the Bayesian sequential method based on posterior predictive checks. We believe that this Bayesian procedure is the correct way to approach the analysis of such experiments because it accounts for the uncertainty of estimation and does not assume a constant treatment effect. Furthermore, relative to all methods compared, we are convinced that the S-PPC give a more intuitive approach into screening by selecting a parsimonious model that is consistent with the data, instead of finding a null model that is not consistent with the observed data, as does the traditional p value approach. For screening in the unreplicated case we recommend the use of step up posterior predictive checks with $\max_{\{j:j\in\mathcal{I}\}} \left|\hat{\theta}_j\right|$ as the discrepancy measure. It is an effective

and principled Bayesian procedure with good frequency operating characteristics. At the moment, that suggestion holds for the replicate case too, but might change after rerunning the simulation study.

In the unreplicated case our simulation results show that our BS-PPC approach balances rejection rates (the finding of true positives) and the error rates in a very appealing way. It was also the most powerful of all methods compared. For the replicate case our method seems to be a trade off between controlling the EER and increasing the RR. Focusing on the EER, it is not as good at controlling it as some other methods (our version of replicate L&N and Bonferroni) but much better at it than the FDR correction. Regarding the RR, it is better than Bonferroni and our extension of the L&N method for the replicate case, but not as powerful as the FDR corrected OLS (the examples presented in this chapter suggest that this might change with the). Overall, the comparison of the methods is more complicated for the replicate case because the advantages of one versus the other can be very small. However for the replicated case, the FDR corrected OLS has the best performance in terms of RR (highest value) and FDR (closes to 0.05). An additional advantage of this method is that it is easy to scale to larger number of factors, whereas the sequential posterior predictive checks require much longer computational time. After running these simulation studies, an appealing idea is to compare the performance of the FDR correction on the permutation test for the unreplicated test using the traditional Benjamini and Hochberg (1995) procedure to the Tripolski et al. (2008) proposed approach.

Interestingly, for the single imputation method in the unreplicated case, the unstandardized test statistic did not perform well, but standardizing using the pseudo standard error showed significant improvement in the method's performance. Perhaps surprisingly, the op-

posite occurs for the S-PPC method although the impact was relatively small. We believe that this behavior is due to the fact that the S-PPC is already accounting for the uncertainty in the estimates and including the PSE in the discrepancy measure only increases the noise in the posterior predictive distribution. Whereas, in the single imputation approach this standardizing is, in some way, including the magnitudes of the other effects that are being assumed inactive by throwing them into the inactive poll and hence in the error and therefore allowing the procedure to better identify outliers. This phenomenon is related to the observation of Ding and Miratrix's paper in preparation, in a context where the design matrix is not orthogonal. They've shown that the permutation test only works if the randomization distribution of the factor tested is obtained by recording the estimated coefficients that result of fitting the full model. In our case, because of the orthogonality the estimate for a factorial effect does not change regardless of what other factors are in the model. However, we believe that when standardizing by the PSE we are including the additional information in the other factors making the randomization distribution useful.

Further theoretical exploration and simulation studies for the 3-level and fractional factorial designs is left for future work. We cannot conclude this work without mentioning nonnormal data. For this case we can consider two paths to explore. First, a common approach to deal with nonnormal data is to use the Box-Cox family of transformations. It would be appealing to dig a little deeper into this idea. Is there a related two parameter transformation that works well to transform the data? The motivation for this exploration is the fact that there are two objectives when attempting to achieve normality: symmetry and tail behavior. In this direction, an extensive literature review is necessary. A second approach, that is more aligned with the Bayesian motivation of this thesis is to approach it from a hierarchical perspective for other kinds of data, such as counts. The relationship

between factorial effects and potential outcomes fitted beautifully in the simplicity of the normal-normal mixture model. Initial attempts on this regard, using an example where the potential outcomes are counts, suggest that setting up the relationship between model parameters, factorial effects (as usually defined - i.e., contrasts of treatment group means) and potential outcomes to fill in the missing potential outcomes will not be as simple for nonnormal data.

Studying the effect of design imbalances in the performance of our methodology is something we are interesting in pursuing. We are curious about whether the potential outcome framework will make the method more robust or more sensitive to this situation. Related to this notion, the main advantage of these factorial designs is the orthogonality of the model matrices, making the estimates of any particular factorial effect the same regardless of which other factorial effects are included in the model. This property is lost in the general setting. For a more general applicability of the S-PPC we can explore its viability for non orthogonal treatment factors. In general, to be able to order the sequence of models to test, we have to be able to get these data into a framework similar to that presented in this thesis. Specifically, scaling considerations should be made. All factors should be normalized (like in the three level design case) to make them all comparable. How far would it take us to chose to create the sequence of models based on some sort of orthogonalization of the treatment factors by using the residuals of the previous model in the sequence?

Another interesting direction is to explore the scalability of our method. However, one major drawback is the computational inefficiency of our method because it requires many posterior predictive draws from the distribution of the discrepancy measure. Each of these need the posterior draws of $\boldsymbol{\mu}, \boldsymbol{\sigma}$ and $\boldsymbol{\theta}_i$ for all the units in the experiment to fill in the

missing outcomes for every draw of the randomization distribution. A way of making this procedure more efficient is highly desirable. In the context of big data, given the computational burden that the randomization based posterior predictive checks can have perhaps a particular screening rule would be more desirable over the other. For example, if the step-down approach were more appropriate the scaled versions of the discrepancy measures would be preferred both in the replicate and unreplicated cases. Moreover, it would be unfeasible to test all of the models in the sequence perhaps tracking a drop on the posterior predictive p values would be a better indication of when to stop the procedure. Even in the stem cell example discussed earlier in this chapter, we used the fact that the 0.05 cut off for OLS does not account for multiple comparisons to limit the sequence of models to test.

The S-PPC methodology uses a natural framework that allows for the use of covariates. Although it is true that in randomized settings "controlling for" background covariates by including them in the regression generally performs well because of the initial balance on background covariates, this framework opens the door to use more sophisticated matching methods to adequately control for covariates when this balance is not achieved for the particular randomization at hand.

Chapter 8

Appendix

8.1 Unbiased estimates of averages of potential outcomes in randomized experiments using symmetry arguments

$$E(\overline{y}(z)) = E\left(\frac{\sum_i W_i Y_i(z)}{\sum_k W_k}\right) = \overline{Y}(z).$$

Proof:

Assume that $\sum_k W_k > 0$, then the ratio is well defined (which is not the case for the regular Bernoulli case). Let $V_i = \frac{W_i}{\sum_k W_k}$. Hence, the V_i's are not independent, but they are identically distributed because of the symmetry. Therefore

$$E\left(\frac{\sum_i W_i Y_i(z)}{\sum_j W_j}\right) = \sum_i E\left(\frac{W_i}{\sum_j W_j}\right) Y_i(z) = \sum_i E\left(V_i\right) Y_i(z).$$

From the symmetry and the linearity of the expectation, we know that $E(V_i) = 1/n$ for all i. Therefore

$$E\left(\frac{\sum_i W_i Y_i(z)}{\sum_s W_s}\right) = \sum_i E\left(\frac{W_i}{\sum_s W_s}\right) Y_i(x) = \sum_i Y_i(z)/n = \overline{Y}(z).$$

Hence, the unbiasedness holds if $\sum_s W_s > 0$. In other words, the unbiasedness of any linear combination of averages holds as long as there is at least one unit in every treatment group, which is a reasonable assumption because a randomization that violated this is unacceptable. It is possible to make it fully true by defining the estimate of linear combinations of mean potential outcomes when a treatment has no units assigned to it such that the average with those allocations equals the population treatment effect. For example, in some cases with two treatments defining $\overline{(Y)}^{obs}(j) = 0$ when no units were assigned to treatment j will make the estimate of the difference unbiased. However, this seems to be a very artificial patch because the assumption made in the above proof is quite reasonable, especially when designing an experiment.

Let $\mathbf{C} = \sum_z c(z)\overline{Y}(z)$ is a contrast of interest where $c(z)$ is a constant corresponding to the z combination. Note that $\mathbf{C}_i = \sum_z c(z)Y_i(z)$ is the same contrast at the unit level. By linearity of the expectation, the unbiasedness extends to any contrast.

Bibliography

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Bhasin, S., Storer, T. W., Berman, N., Callegari, C., Clevenger, B., Phillips, J., Bunnell, T. J., Tricker, R., Shirazi, A., and Casaburi, R. (1996). The effects of supraphysiologic doses of testosterone on muscle size and strength in normal men. *New England Journal of Medicine*, 335(1):1–7.

Box, G. E., Hunter, J. S., and Hunter, W. G. (2005). *Statistics for experimenters: design, innovation, and discovery*, volume 2. Wiley Online Library.

Box, G. E. and Meyer, R. D. (1986). An analysis for unreplicated fractional factorials. *Technometrics*, 28(1):11–18.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36.

Chipman, H., Hamada, M., and Wu, C. (1997). A bayesian variable-selection approach for analyzing designed experiments with complex aliasing. *Technometrics*, 39(4):372–381.

Cox, D. R. (1958). *Planning of Experiments*. London: Chapman & Hall.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1(4).

Dasgupta, T., Ma, C., Joseph, V. R., L., W. Z., and J., W. C. F. (2008). Statistical modeling and analysis for robust synthesis of nanostructures. *Journal of the American Statistical Association*, 103:594–603.

Dasgupta, T., Pillai, N. S., and Rubin, D. B. (2012). Causal inference from 2^k factorial designs using the potential outcomes model. *arXiv preprint arXiv:1211.2481*.

Dong, F. (1993). On the identification of active contrasts in unreplicated fractional factorials. *Statistica Sinica*, 3:209–217.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing.

Fisher, R. A. (1935). The design of experiments.

Gelman, A., Meng, X. L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, pages 733–807.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Hamada, M. and Balakrishnan, N. (1998). Analyzing unreplicated factorial experiments: a review with some new proposals. *Statistica Sinica*, 8(1).

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Kempthorne, O. (1952). The design and analysis of experiments.

Kempthorne, O. (1955). The randomization theory of experimental inference. *Journal of the American Statistical Association*, 50(271):946–967.

Kempthorne, O. and Hinkelmann, K. (1984). *Experimental design, statistical models, and genetic statistics: essays in honor of Oscar Kempthorne*, volume 50. CRC Press.

Lenth, R. V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics*, 31(4):469–473.

Loughin, T. M. and Noble, W. (1997). A permutation test for effects in an unreplicated factorial design. *Technometrics*, 39(2):180–190.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, pages 1142–1160.

Montgomery, D. C., Montgomery, D. C., and Montgomery, D. C. (1984). *Design and analysis of experiments*, volume 7. Wiley New York.

Mukerjee, R. and Wu, C. J. (2006). *A modern theory of factorial design*. Springer.

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translation of excerpts by D. Dabrowska and T. Speed. *Statistical Science*, 6:462–47.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. In *The Proceedings of the social statistics section of the American Statistical Association*, pages 233–239.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational and Behavioral statistics*, 2(1):1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test: Comment. *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, pages 1151–1172.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480.

Rubin, D. B. (1998). More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in medicine*, 17(3):371–385.

Rubin, D. B. (2008). Statistical inference for causal effects, with emphasis on applications in epidemiology and medical statistics. *Handbook of statistics*, 27:28–58.

Rubin, D. B. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010).

Tripolski, M., Benjamini, Y., and Steinberg, D. M. (2008). The false discovery rate for multiple testing in factorial experiments. *Technometrics*, 50(1).

Wu, C. and Hamada, M. S. (2009). Experiments: Planning, analysis and optimization, hoboken.

Xu, H. (2005). A catalogue of three-level regular fractional factorial designs. *Metrika*, 62(2-3):259–281.

Ye, K. Q. and Hamada, M. (2000). Critical values of the lenth method for unreplicated factorial designs. *J. Quality Technology*, 32(1):57–66.

Ye, K. Q., Hamada, M., and Wu, C. (2001). A step-down lenth method for analyzing unreplicated factorial designs. *Journal of Quality Technology*, 33(2):140–152.

Zahn, D. A. (1975). Modifications of and revised critical values for the half-normal plot. *Technometrics*, 17(2):189–200.

Zemplenyi, M. (2013). Design and analysis of a fractional factorial screening experiment to identify small molecule inducers of pancreatic beta cells, undergraduate thesis.