This is the First Manuscript of the article published by Taylor & Francis in *Journal of Applied Statistics* on 2018, available at https://www.tandfonline.com/doi/abs/10.1080/02664763.2018.1548584

Selecting significant effects in factorial designs: Lenth's method versus the Box-Meyer approach

Rafel Xampeny, Pere Grima, Xavier Tort-Martorell
Department of Statistics and Operations Research
Universitat Politècnica de Catalunya – BarcelonaTech, Spain

ABSTRACT

The Lenth method is conceptually simple and probably the most common approach to analyzing the significance of the effects in factorial designs. Here, we compare it with a Bayesian approach proposed by Box and Meyer and which does not appear in the usual software packages. The comparison is made by simulating the results of 4, 8 and 16 run designs in a set of scenarios that mirror practical situations and analyzing the results provided by both methods. Although the results depend on the number of runs and the scenario considered, the use of the Box and Meyer method generally produces better results.

KEYWORDS: Factorial design, significant effects, Lenth method, Box-Meyer method, four-run experiments.

1. Introduction

Through experimentation, two-level factorial designs provide a great number of possibilities for efficiently analyzing how a set of variables affect a response – particularly in industrial environments. This influence is quantified by calculating the effects, which are orthogonal contrasts of the response vector. Since the effects are affected by random variability – which is inherited from the variability of the response –it is necessary to analyze whether its value is significantly different from zero.

When there are replicas, that is to say, when the experiment has been conducted several times at each experimental condition, we can estimate the experimental error and from it we can get an estimate of the variance of the effects. This estimate can be used to perform significance tests for each effect in the usual way. However, given that the resources for experimentation are usually limited, replicas are typically lacking. In cases it is necessary to

analyze the significance of the effects using other methods, which can be graphical or analytical.

Among the graphical methods is the Pareto diagram of effects – where the value of the significant effects is expected to stand out from the rest – and the representation of the effects on a Normal Probability Plot (NPP) [5]. When the effects are represented in NPP, it is expected that the non-significant ones (which belong to a Normal distribution with average $\mu = 0$) will fall on a line that passes through the point $(0, 0.5)$. A variant of NPP is the Half Normal Plot; and in this case the line goes through the point $(0, 0)$.

Representing the effects with NPP is very useful. However, it is not always easy to interpret, especially when there are few effects, as in designs with 8 or fewer runs (a study on the topic can be seen in [6]). Furthermore, it cannot be used for making automatic decisions in statistical software packages. The impossibility of automating its use prevents comparison of its effectiveness with other methods and thus the method is not included in this study.

There are many analytical methods for testing the significance of effects in the absence of replicas. Hamada and Balakrishnan [9] analyze the advantages and disadvantages of a wide selection of them. The one that appears in the most typical textbooks (such as Box, Hunter and Hunter [4] and Montgomery [11]) as well as in the most usual statistical software packages for industrial applications (see [7]) is the Lenth method [10], which is conceptually simple and provides good results.

Box and Meyer published a method using a Bayesian approach [2] [3]. However, due probably to its greater complexity, it did not become widely used and is not among those usually considered when analyzing the significance of effects; nor does it appear as an option in the statistical software packages that are most commonly used by practitioners [7].

In this article, we defend the Box-Meyer method, showing its effectiveness in a wide variety of scenarios that endeavor to represent practical situations. The article is organized as follows. First, the Lenth and Box-Meyer methods are described. Next, we present the situations in which the two methods are compared and the comparison criteria are described. Next, the results obtained are analyzed, showing that the Box-Meyer method performs best in most situations.

2. Lenth and Box-Meyer Methods

Lenth's method consists of estimating the standard deviation of the effects based on the fact that if $X \sim N(0, \sigma)$, the median of $|X|$ is equal to 0.645σ and therefore $1.5 \cdot$

$\text{median}|X| = 1.01\sigma \cong \sigma$. Supposing that κ_i $(i = 1, \dots, n)$ are the values of the effects of interest and that their estimators c_i are distributed according to $N(\kappa_i, \sigma_{ef})$, then s_0 Is defined as $1.5 \cdot \text{median}|c_i|$and this value is used to calculate a new median by excluding the estimates of the effects with the value $|c_i| > 2.5s_0$ in order to exclude those with $\kappa > 0$. In this way you get the so-called *Pseudo Standard Error*:

$$PSE = 1.5 \cdot \underset{|c_i|<2.5s_0}{\text{median}}|c_i|$$

From the *PSE* you can calculate a margin of error, *ME*, which, for a confidence level of 95% will be $ME = t_{0.975,\nu} \times PSE$. If $|c_i| > ME$, then the effect $c_i$is considered significant.

Lenth [10] includes a table with the values of $t_{0.975}$ for designs 2^{k-p} with values of $k - p$ that are understood to be between 3 and 8, that is, designs with between 8 and 256 runs. No examples or references to designs with $k - p = 2$ (4 runs) are included; but some software packages also use it in this case (perhaps because the original article does not explicitly discourage its use). On the other hand, Lenth proposes using $\nu = n/3$, with n being the number of effects considered; and this is the value that has been used in some known software packages [6], although it has been shown that it produces type I error probabilities below 5%, which is counterbalanced by higher probabilities of type II error. Ye and Hamada [14] and Fontdecaba et al. [8] have proposed values of t that deliver better results (Table 1).

Table 1: *Proposed values for the value of $t_{0.975}$ that must be applied together with the PSE*

	Proposed values for $t_{0.975}$		
Estimated effects	Lenth	Ye and Hamada	Fontdecaba *et al.*
7	3. 76	2.297	2
15	2. 57	2. 156	2

The Box and Meyer method [2], [3] considers the set of all possible models that can be proposed: $M_0, M_1, \cdots, M_m$. The value of m is equal to $2^n - 1$, with n being the number of effects that are going to be analyzed. So, for example, in a 2^3 design with factors A, B and C, we will have $m = 127$, with M_0 being a model that does not include any significant effect until M_{127}, which includes the 7 effects considered: A, B, C, AB, AC, BC and ABC. This is to determine – by means of Bayes' theorem – the probability of each model M_i given the response vector $\boldsymbol{y}$, that is to say:

$$p(M_i|\boldsymbol{y}) = \frac{p(M_i)f(\boldsymbol{y}|M_i)}{\sum_{h=0}^{m} p(M_h)f(\boldsymbol{y}|M_h)}$$

Calculating $p(M_i)$ is easy. If the total number of effects considered is N, the probability that an effect is active is π and f_i is the number of active effects in model M_i; then $p(M_i) = \pi^{f_i}(1-\pi)^{N-f_i}$. The value of π must be previously fixed. Box and Meyer use the value of 0.25 in the examples they present.

For the calculation of $f(\boldsymbol{y}|M_i)$, it is necessary to assign an a priori distribution to the effects values. Box and Meyer propose using $N(0, \gamma^2\sigma^2)$. Where the mean is 0 due to the lack of a priori knowledge regarding the direction of each effect, and the parameter γ captures the magnitude of the effect relative to the experimental noise. It is suggested to assign to γ the value that minimizes the probability that all the effects are null. The expression of $f(\boldsymbol{y}|M_i)$ and the details for deducing it can be seen in the Appendix of Box and Meyer's second article [3].

3. Test scenarios

To study the probabilities of error in the effects significance analysis, we have proposed a series of scenarios that try to represent situations that the experimenter can find in practice. These scenarios consider part of the effects to be null, that is, that their values belong to a distribution $N(\mu = 0;\ \sigma_{ef})$. The rest have an average that is equal to Δ or a multiple of this value. With no loss of generality, $\sigma_{ef} = 1$ is taken and, following the criteria of Ye *et al*. [15], the values of Δ are designated Spacing and they vary between 0.5 and 8 in increments of 0.5.

We perform simulations for designs with 4, 8 and 16 runs and omit designs with more runs since they are not widely used. What is more, this designs allow estimating a lot of effects, many of which – according to the effect sparsity principle – will be zero. In this circumstance, identifying those that are significant is an easy task with any procedure.

At the opposite end are the designs with 4 runs. Although they are not usually considered in articles that deal with effects significance analysis, it is not unusual to have two factors remaining for study in the last steps of a sequential experimentation process. What is certain is that with only three effects (those obtained from a design with 4 runs) it is difficult to select those that should be considered significant when no information is available on the experimental error. In this circumstances the usual methods are totally ineffective. We have seen practitioners and students surprised to see in the information provided by the software they are using that none of the two factors they are considering have an influence on the response – even though everything indicated that at least one should. Thus, we have included in our analysis the case of designs with only 4 runs. We will see that also in this

case and in spite of not presenting extraordinary results the Box-Meyer method improves the Lenth method.

In the three cases, 4, 8 and 16 run designs we propose 6 scenarios.

For the 4 run designs the scenarios cover from the case when the three effects are null up to when all three effects are active. In this last case, the effects can have the same average value, Δ, or average values of Δ, 2Δ, 3Δ (Table 2)

Table 2: *Effect values in scenarios considered in 4-run designs*

Scenarios	Effects		
	1	2	3
$S4_1$	0	0	0
$S4_2$	**Δ**	0	0
$S4_3$	**Δ**	**Δ**	0
$S4_4$	**Δ**	**2Δ**	0
$S4_5$	**Δ**	**Δ**	**Δ**
$S4_6$	**Δ**	**2Δ**	**3Δ**

For 8 run designs, we first considered scenarios 1, 2, 3 and 5 that were used by Fontdecaba *et al*. [8] to analyze the performance of Lenth's method. And then we added scenarios 4 and 6, in which there exists the possibility that 4 significant effects also exist (Table 3).

Table 3: *Effect values in scenarios considered in 8-run designs*

Scenarios	Effects						
	1	2	3	4	5	6	7
$S8_1$	**Δ**	0	0	0	0	0	0
$S8_2$	**Δ**	**Δ**	0	0	0	0	0
$S8_3$	**Δ**	**Δ**	**Δ**	0	0	0	0
$S8_4$	**Δ**	**Δ**	**Δ**	**Δ**	0	0	0
$S8_5$	**Δ**	**2Δ**	**3Δ**	0	0	0	0
$S8_6$	**Δ**	**2Δ**	**3Δ**	**4Δ**	0	0	0

For 16 run designs, we use the same scenarios that were used for the first time by Venter and Steel [13], and later by Ye *et al*. [15] and Fontdecaba *et al*. [8]. From 1 to 7 significant effects are considered (Table 4):

Table 4: *Effect values in scenarios considered in 16-run designs*

Scenarios	Effects														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$S16_1$	**Δ**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$S16_2$	**Δ**	**Δ**	**Δ**	0	0	0	0	0	0	0	0	0	0	0	0
$S16_3$	**Δ**	**Δ**	**Δ**	**Δ**	**Δ**	0	0	0	0	0	0	0	0	0	0
$S16_4$	**Δ**	**Δ**	**Δ**	**Δ**	**Δ**	**Δ**	**Δ**	0	0	0	0	0	0	0	0
$S16_5$	**Δ**	**2Δ**	**3Δ**	0	0	0	0	0	0	0	0	0	0	0	0
$S16_6$	**Δ**	**2Δ**	**3Δ**	**4Δ**	**5Δ**	0	0	0	0	0	0	0	0	0	0

4. Simulation

For each scenario, and within each scenario for each Spacing value we have simulated 10,000 situations. Each of them has been analyzed using Lenth's (with $t = 3.76$ and $t = 2.297$) and Box-Meyer's methods.

To apply the Lenth method in designs with 8 or 16 runs, we perform the analysis using values of $t_{0.975}$ (which were proposed in the original article [6]) and also those proposed by Ye and Hamada [14]. When using this in 4-run designs the value of $t_{0.975}$ with a single degree of freedom ($t = 12.71$) – which would be the one obtained by following Lenth's proposed general rule and which is still used by some statistical software packages – gives very bad results, practically never detects the active effects[1]. We have studied how the probabilities of type I and type II errors vary according to the value of t in the scenarios considered. It is observed (Figure 1) that even when dropping down to a value of $t = 2$, the active effects are barely detected. For $t = 2/3$, the type I error ratios are similar to those obtained with the Box-Meyer method, so we present the comparison with the values obtained for this value of t.

When we apply the Box-Meyer method in designs with 8 or 16 runs, we followed the authors' recommendation both for the a priori proportion of significant effects (π = 0.25), and for the estimation of the parameter γ (the value that minimizes the probability of the model having all null effects). In designs with 4 runs, it is reasonable to consider that the 3 effects may be null or may be active; thus, in this case we have taken the value of $\pi = 0.50$. Regarding the value of γ, we have analyzed the type I and type II error proportions in all scenarios and for all Spacing values (Figure 2). Some values give good results in some

[1]R.V. Lenth was aware that his method could not be applied to 4 run designs and in his paper never tries to do that. Unfortunately several statistical packages apply it to all two level designs independently of the number of runs.

scenarios but bad in others. We have chosen $\gamma = 2$, since this value reasonably balances the goodness of the results in all scenarios.

To determine the probabilities $p(M_i|\boldsymbol{y})$, we have used the BsMD package developed by Barrios [1] for the statistical software R [12]. Introducing the design matrix, the response vector and the values for π and γ delivers a list of models that are ordered according to the posterior probability of being correct of each of them. The effects considered significant are those contained in the model with the greatest probability. The package also includes a function to identify the value of γ that minimizes the probability that all effects are null. We have used this value in 8 and 16 run designs.

To illustrate the procedure followed, let us take as an example the results from one of the 10 000 simulations performed in scenario $S8_2$ with a Spacing value $\Delta = 3$. The values of the effects are those indicated in Table 5 (Effects, c_i). Applying the Lenth method delivers a $PSE = 0.5625$, if we use the value of $t = 3.76$, the effects that present $|c_i| > 2.115$ must be considered significant. In this case effect 1. Since those that are actually active are effects 1 and 2, a type II error is committed because 2 is not considered significant. If we apply the Box-Meyer method, we first determine the value of γ that minimizes $p(M_0|\boldsymbol{y})$, it is $\gamma = 2.5$. Using this value, the model with the greatest a posteriori probability is the one that includes the effects 1, 2 and 4. As only effects 1 and 2 are really active, the Box-Meyer method succeeds in identifying them as such; but it is also mistaken in considering effect number 4 to be significant and thus commits a type I error in this case. Table 5 summarizes the results obtained.

Table 5: *Results with the values of the effects obtained by simulation for a design with 8 experiments, scenario S82, $\Delta = 3$.*

Effects		Actual	Effects significance analyzed by:			
#	c_i	Fact	Lenth Method ($t = 3.76$)		Box and Meyer Method	
1	4.44	Active	SIGNIFICANT	(Correct)	SIGNIFICANT	(Correct)
2	1.75	Active	Not significant	(Type II error)	SIGNIFICANT	(Correct)
3	-0.13	Inert	Not significant	(Correct)	Not significant	(Correct)
4	1.18	Inert	Not significant	(Correct)	SIGNIFICANT	(Type I Error)
5	-0.48	Inert	Not significant	(Correct)	Not significant	(Correct)
6	0.27	Inert	Not significant	(Correct)	Not significant	(Correct)
7	-0.08	Inert	Not significant	(Correct)	Not significant	(Correct)

After performing 10 000 simulations, the errors of each type and for each method are added together and their percentage of all total possibilities is calculated. Thus, in scenario $S8_2$ you

can commit up to 50 000 type I errors (in each simulation there are 5 inert effects that, erroneously, can be considered active), and you have 20 000 options for a type II error (in each simulation there are 2 active effects that may not be identified). The results obtained in the case of our example ($S8_2$, $\Delta = 3$) are indicated in Table 6.

Table 6: *Types of error produced in the 10 000 simulations of the values of the effects in scenario* $S8_2$ *with* $\Delta = 3$

	Type I error		Type II error	
	Absolute value	Percentage	Absolute value	Percentage
Lenth Method (t=3.76)	356	$\frac{356}{50000}100 = 0.712$	15544	$\frac{15544}{20000}100 = 77.72$
Box-Meyer Method	1619	$\frac{1619}{50000}100 = 3.238$	10090	$\frac{10090}{20000}100 = 50.45$

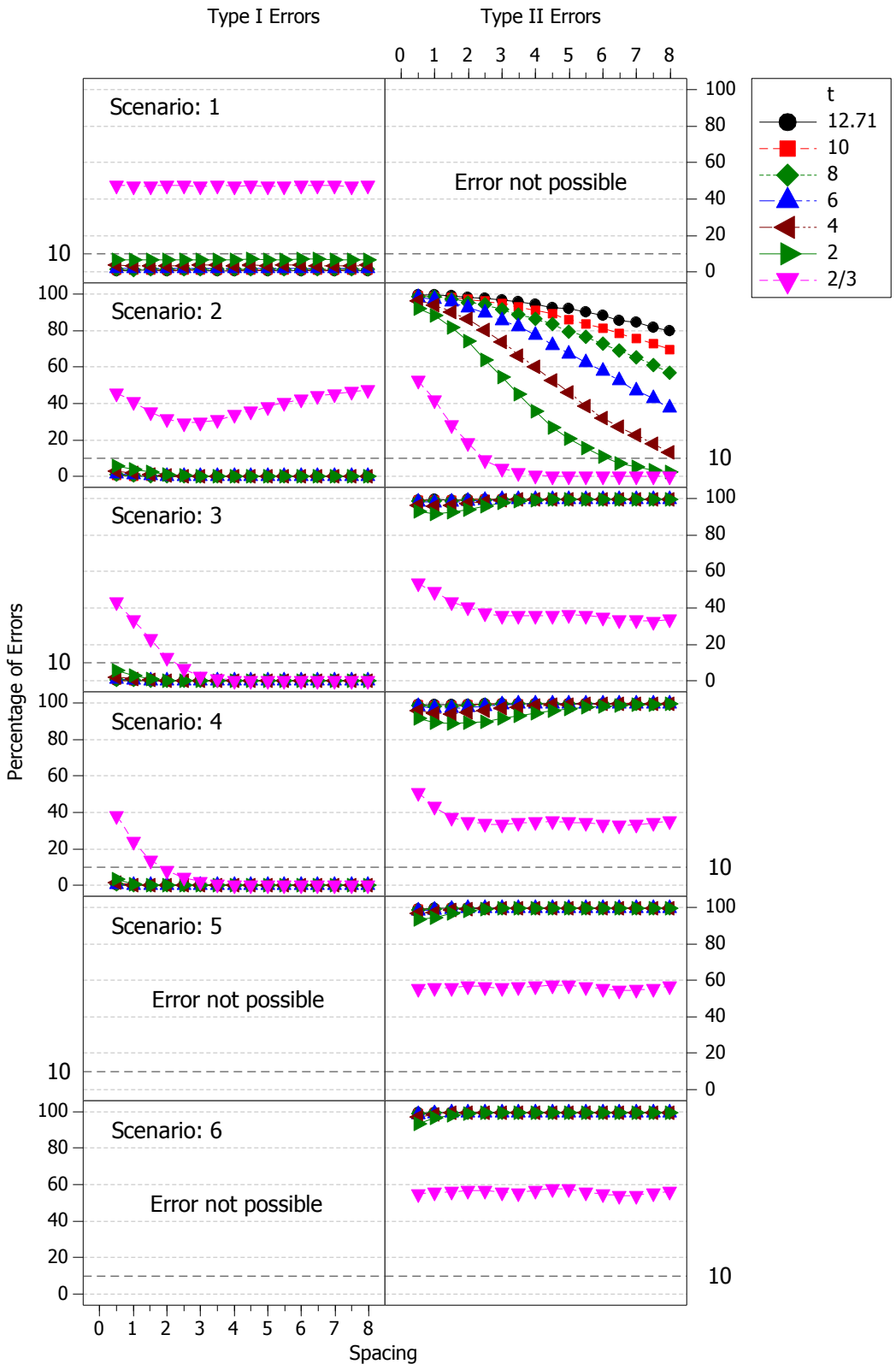

Figure 1: *Lenth method. Proportion of errors in designs with 4 runs depending on the value of t chosen.*

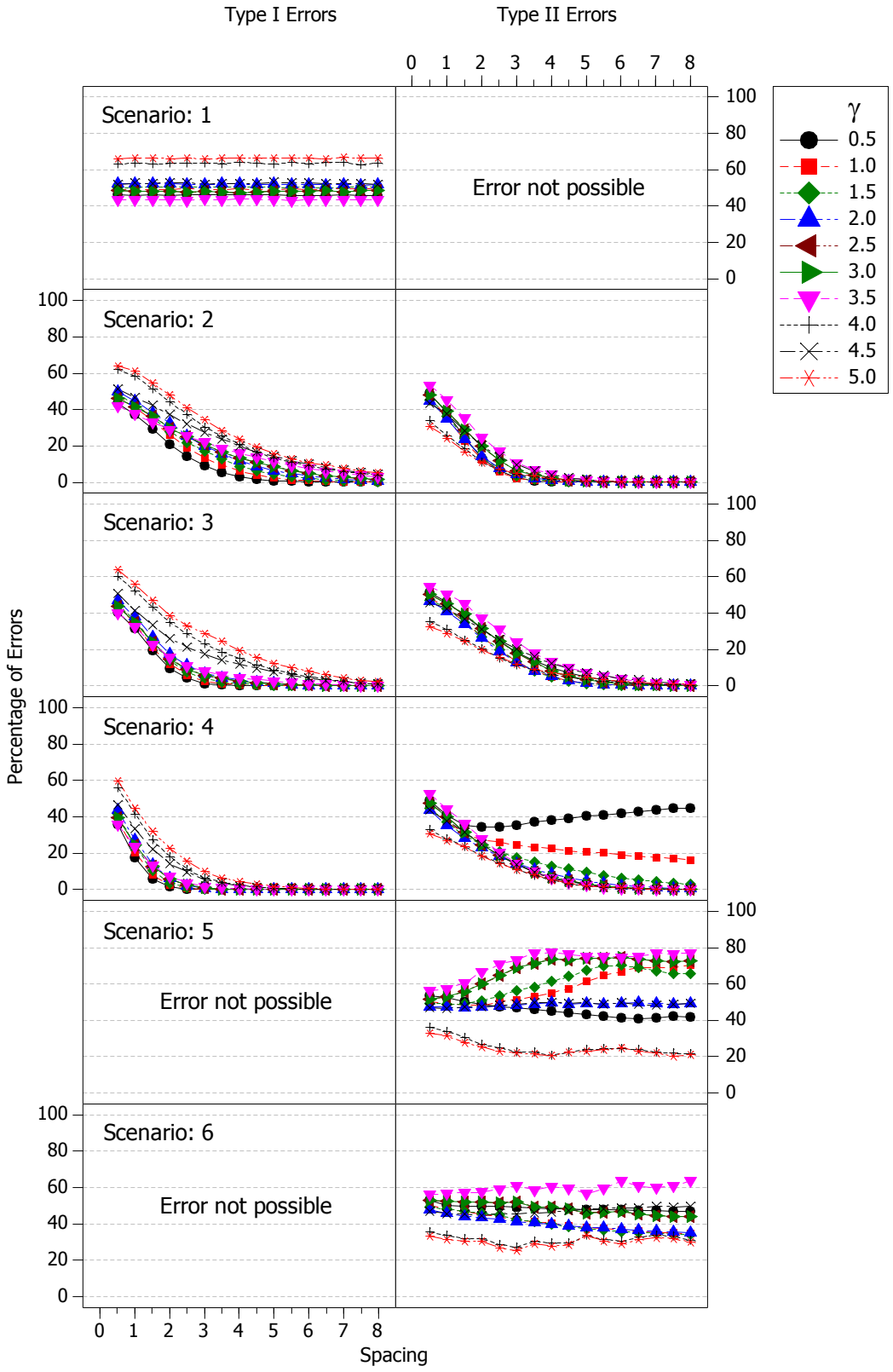

Figure 2: *Box-Meyer method. Proportion of errors in designs with 4 runs depending on the value of* γ

5. Results

Figure 3 shows the comparison of the results obtained for 4 run designs. In scenario 1 there is no probability of type II error since no effect is active. Nor can there be any type I error in scenarios 5 and 6, since all effects are active. Lenth's method always gives a lower proportion of type I error, but at the expense of systematically ignoring type II errors in all scenarios except for 2. The Box-Meyer method produces a greater proportion of type I errors, especially in scenario 1, but type II errors fall significantly in all scenarios. We cannot say that the Box-Meyer method is excellent in this case, but the results are clearly better than with Lenth's method. In any case, it seems important to us that the experimenter is aware of the shape of these error curves.

For 8 run designs, the results are summarized in Figure 4. Regarding type I errors, the differences are small and in all cases reasonable values are presented. Regarding type II errors, the greater probability of error in the 4 scenarios emerges when using the value of $t = 3.76$, as already shown in [8]. The Box-Meyer method has lower values of type II error in all scenarios and for all Spacing values.

In 16 run designs the results are presented in Figure 5. In this case the number of type I errors are also reasonable in all cases. Regarding the proportion of type II errors, the worst performance of the Lenth method occurs with $t = 2.57$, especially in scenarios 1, 2 and 3; and the highest proportion of type II errors with the Box-Meyer method occurs in scenario 4, especially with Spacing values above 3. In this case, the problem lies in having 46.7% of active effects, a value that is far from the $\pi = 0.25$ that is generally assumed. If $\pi = 0.50$ is considered, the results are practically identical to those from the Lenth method.

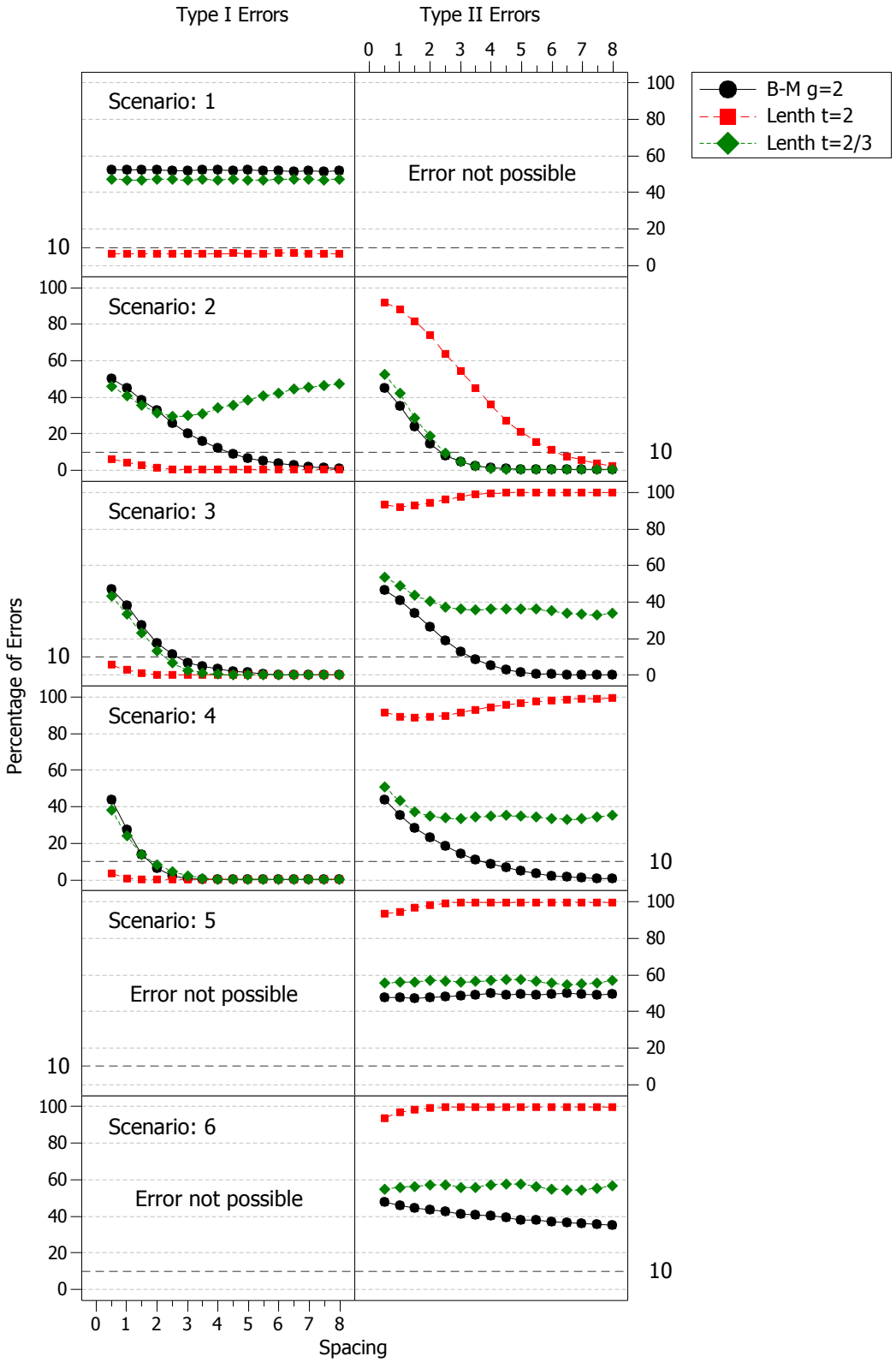

Figure 3: *Designs with 4 runs. Comparison of the Lenth and Box-Meyer methods*

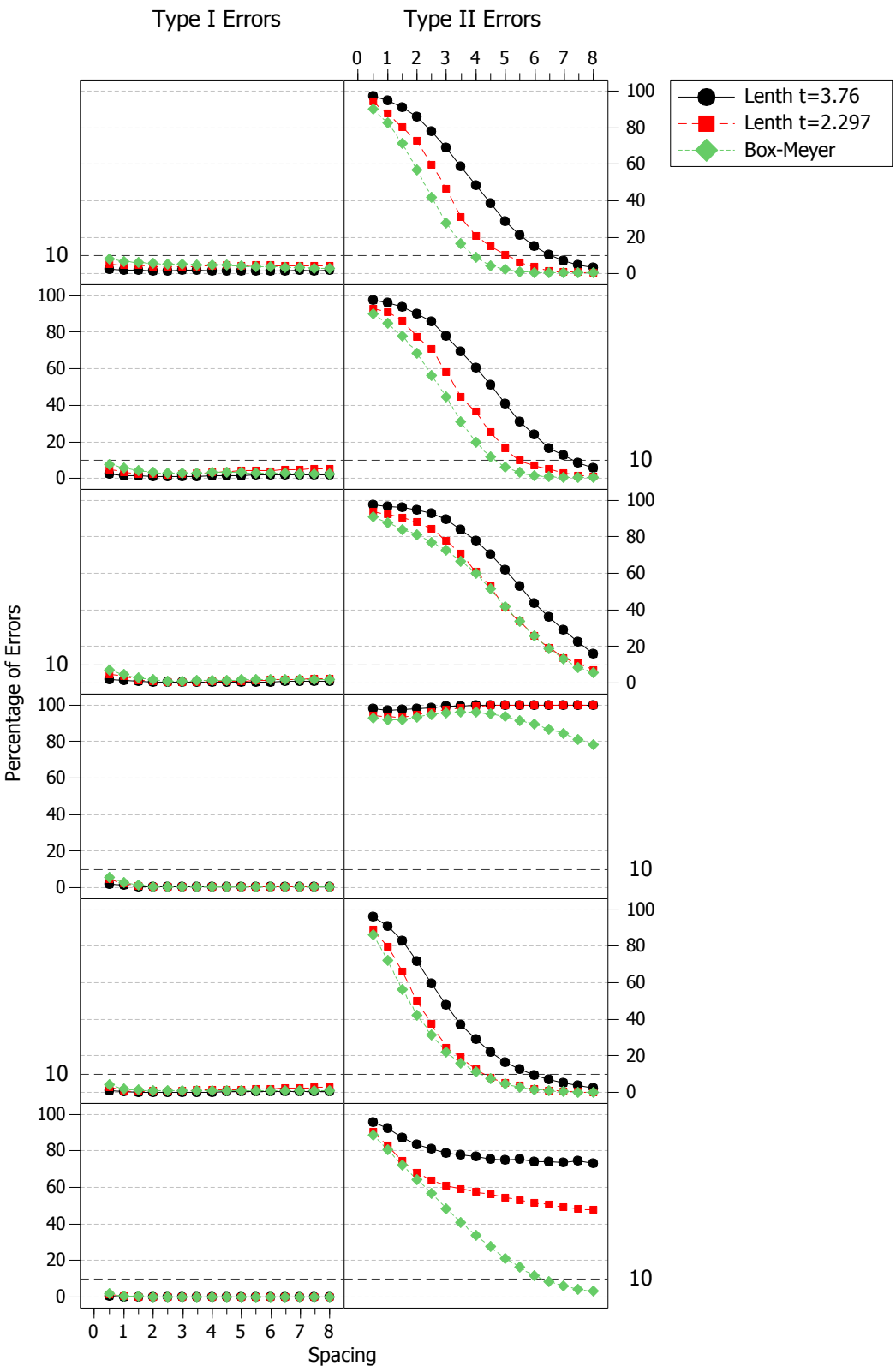

Figure 4: Designs with 8 runs. Comparison of the Lenth and Box-Meyer methods. The values obtained in the example of Table 6 have been circled

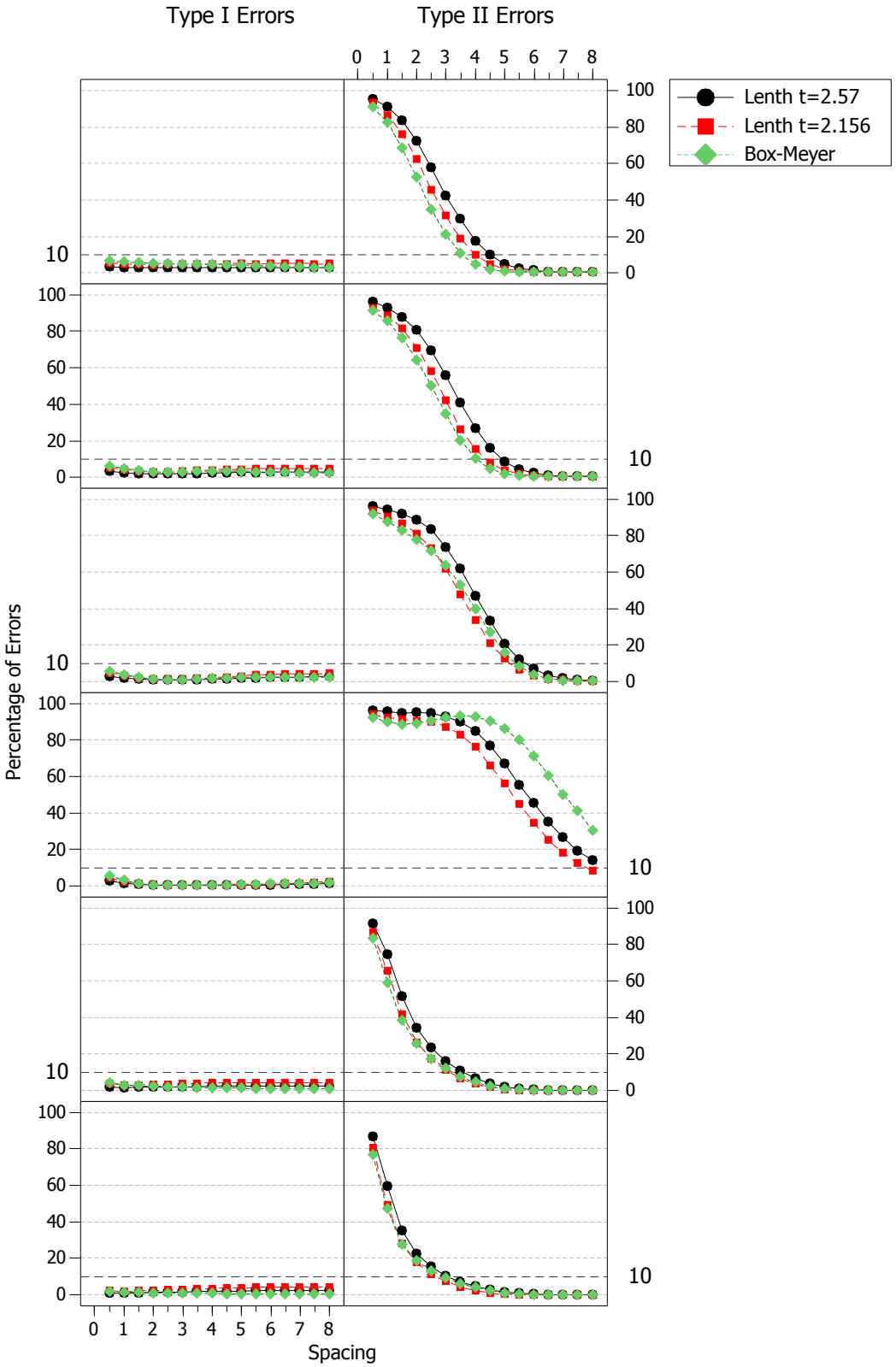

Figure 5: *Designs with 16 runs. Comparison of the Lenth and Box-Meyer methods*

6. Summary and recommendations

Our conclusions and recommendations after the thorough comparison between Lenth and Box-Meyer methods are:

- 4 run designs: This is a neglected situation in the literature on factorial designs. In this case, both the representation of effects in NPP and Lenth's method are – by their very nature – ineffective at identifying which effects should be considered significant. With only 3 effects, it is not possible to discriminate "those that separate from the line" when the NPP is used. Also, the Lenth method is not reliable and – especially if the recommended value of t is used – practically in no case does it detect the active effects. Naturally, miracles cannot be expected with only 3 effects, and the Box-Meyer method does not deliver excellent results either, but they are – in all scenarios considered – better than those delivered by the Lenth method, even when using the value of t that favors it more.

  It is important to be aware that if a design with 4 runs is carried out without prior information about the variability of the response, it is not possible to analyze the significance of the effects with reasonable error probabilities. If the experiment is carried out at the end of a process of sequential experimentation, the best option is to estimate the experimental error from the values of the non-significant effects obtained in the previous experiments, and estimate the variance of effects from it.

- 8 run designs: Of the two most usual designs (8 and 16 runs), these are the most difficult to analyze. The smaller number of effects makes it difficult to discriminate between those that are significant and those that are not. In this case the Box-Meyer method performs better than the Lenth method (better than when using the original value of $t = 3.76$, of course, but also when using the $t = 2.297$ value proposed by Ye and Hamada), in all scenarios and for all Spacing values.

- 16 run designs: In this case the differences are barely noticeable, except in scenarios 1 and 2, in which the Box-Meyer method is slightly better (lower proportion of type II errors); but it is slightly worse in scenario 3 and notably worse in scenario 4. In scenario 4 the proportion of active effects is close to 50%, a value that is far from the 25% assumed a priori. In both Scenario 3 and Scenario 4, if a proportion of significant effects is considered at around 50%, the results are similar to those obtained with Lenth's method.

This study clearly shows that the Box-Meyer method gives– in general – better results than the widely adopted Lenth one. Therefore, we strongly advocate for the incorporation of the Box-Meyer method to statistical packages. Having it available as an alternative or even

complementary to another method will help the experimenter make better informed decisions.

A last point, worth mentioning, is that the simulation carried out confirms what other authors have already shown (see, for example [14], [8]), namely: the value of t that appears in the original article on the Lenth method and that is still used in the most widely distributed packages of statistical software [6] produces, on the one hand, a probability of type I errors smaller than the intended 5%; causing, as a counterpart, a high probability of type II error, that is, it does not consider effects to be active when they actually are. In all the designs and in all the scenarios considered, the value of t proposed by Ye and Hamada produces a type I error probability that is closer to 5% and a lower probability of type II error.

References

[1] E. Barrios (based on Daniel Meyer's code). BsMD: Bayes Screening and Model Discrimination. R package version 2013.0718, 2013; software available at https://CRAN.R-project.org/package=BsMD

[2] G.E.P. Box and R.D. Meyer, *An analysis for unreplicated fractional factorials*. *Technometrics* 28 (1986), pp. 11-18.

[3] G.E.P. Box and R.D. Meyer, *Finding the Active Factor in Fractionated Screening Experiments*. Journal of Quality Technology 25 (1993), pp. 94-105.

[4] G.E.P. Box, J.S. Hunter and W.G. Hunter, *Statistics for experimenters: Design, innovation, and discovery*. Wiley, New Jersey, 2005.

[5] C. Daniel, *Use of half-normal plots in interpreting factorial two-level experiments*, *Technometrics* 1 (1959), pp. 311-341.

[6] G. De León, P. Grima and X. Tort-Martorell, *Comparison of normal probability plots and dot plots in judging the significance of effects in two level factorial designs*, J. Appl. Stat. 38 (2011), pp. 161-174.

[7] S. Fontdecaba, P. Grima and X. Tort-Martorell, *Analyzing DOE with statistical software packages: Controversies and proposals.* Amer. Statist. 68 (2014), pp. 205-211.

[8] S. Fontdecaba, P. Grima and X. Tort-Martorell, *Proposal of a single critical value for the Lenth method*. Quality Technology and Quantitative Management 12 (2015), pp. 41–51.

[9] M. Hamada and N. Balakrishnan, *Analyzing unreplicated factorial experiments: A review with some new proposal.* Statistica Sinica 8 (1998), pp. 1–41.

[10] R.V. Lenth, *Quick and easy analysis of unreplicated factorials*, Technometrics 31 (1989), pp. 469-473.

[11] D.C. Montgomery. *Design and analysis of experiments*. Willey, Singapore, 2013.

[12] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2016; software available at https://www.R-project.org/

[13] J.H. Venter and S.J. Steel, *Identifying active contrasts by stepwise testing*. Technometrics 40 (1998), pp. 304-313.

[14] K.Q. Ye and M. Hamada, *Critical values of the Lenth method for unreplicated factorial designs.* Journal of Quality Technology 32 (2000), pp. 57–66.

[15] K.Q. Ye, M .Hamada and C.F.J. Wu, *A step-down Lenth method for analyzing unreplicated factorial designs*. Journal of Quality Technology 33 (2001), pp. 140-152.