Consequences of using estimated response values from negligible interactions in factorial designs

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

ABSTRACT

This article analyzes the increase in the probability of committing type I and type II errors in assessing the significance of the effects when some properly selected runs have not been carried out and their responses have been estimated from the interactions considered null from scratch. This is done by simulating the responses from known models that represent a wide variety of practical situations that the experimenter will encounter; the responses considered to be missing are then estimated and the significance of the effects is assessed. Through comparison with the parameters of the model, the errors are then identified. To assess the significance of the effects when there are missing values, the Box-Meyer method has been used. The conclusions are that 1 missing value in 8 run designs and up to 3 missing values in 16 run designs experiments can be estimated without hardly any notable increase in the probability of error when assessing the significance of the effects.

KEYWORDS: Factorial design, missing values, negligible interactions, Lenth method, significant effects.

1. Introduction

In a factorial design, it is possible to estimate as many missing response values as there are interactions that can be considered negligible[1,2]. Take, for example, a 2^3 design with a table of contrasts such as Table 1.

Table 1: *Contrasts and responses for a* 2^3 *design*

A	B	C	AB	AC	BC	ABC	Y
-1	-1	-1	1	1	1	-1	y_1
1	-1	-1	-1	-1	1	1	y_2
-1	1	-1	-1	1	-1	1	y_3
1	1	-1	1	-1	-1	-1	y_4
-1	-1	1	1	-1	-1	1	y_5
1	-1	1	-1	1	-1	-1	y_6
-1	1	1	-1	-1	1	-1	y_7
1	1	1	1	1	1	1	y_8

If the ABC interaction is negligible we have:

$$-y_1 + y_2 + y_3 - y_4 + y_5 - y_6 - y_7 + y_8 = 0$$

And from this expression we can deduce any response value depending on the remainder.

This procedure can be very useful when it is not possible to perform all the runs required by the chosen design, but it also has undesired consequences. It is straightforward to see that if σ_y^2 is the variance of the responses obtained from the experimentation, the variance of the estimated response will be $7\sigma_y^2$. We will discuss later how this fact affects the analysis of the significance of the effects.

Another problem with this procedure is that the estimation of missing values is not always possible. For example, if in a 2^3 design there were two missing values and the interactions BC and ABC could be considered negligible, we would have 28 possible pairs of missing values and only the values of 16 of them could be estimated. Table 2 shows the contrasts associated with interactions BC and ABC. Their expressions can provide a system of two equations with two unknowns to deduce, for example, the values of y_1 and y_2; however, this cannot be done to deduct y_1 and y_3 since the system of equations is inconsistent.

Table 2: Contrasts associated with the BC and ABC interactions in a 2^3 *design*

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
BC	1	1	-1	-1	-1	-1	1	1
ABC	-1	1	1	-1	1	-1	-1	1

In addition, when there is more than one missing response, the variances of the estimated values depend on which those responses are and also on the interactions used for their estimation. In Xampeny *et al.* [3], it is shown that if in a 2^4 design the five three or more factors interactions can be considered negligible, there will be 4368 possible quintets of missing responses, of which it is impossible to estimate the values of 1360 of them due to their systems of equations being inconsistent. For the combinations that can be estimated, there are notable differences in the variances of the estimated values, depending on the missing responses. For example, the combination of missing values y_1, y_2, y_3, y_8 and y_{12} is one of the 480 that lead to estimates with maximum variances, namely: $31\sigma_y^2$, $15\sigma_y^2$, $15\sigma_y^2$, $7\sigma_y^2$ and $7\sigma_y^2$, respectively; while the combination y_1, y_4, y_6, y_{10} and y_{15} is one of the 16 that present lower values in the variance of the estimates, precisely $2.56\sigma_y^2$ for all of them. Naturally, the bigger the variance of the estimated values, the bigger the variance of the effects.

An additional problem is that since some values of the response are deduced from others, the effects are correlated among them. For example, if in a 2^3 design we have experimentally obtained the eight values of the response, the main effect of, let us say factor A, will be:

$$A = \frac{1}{4}(-y_1 + y_2 - y_3 + y_4 + y_5 + y_6 - y_7 + y_8)$$

From which we get that the variance of A, $V(A)$ is equal to $\sigma_y^2/2$. However, if we have a missing value, for instance y_1, we have:

$$y_1 = +y_2 + y_3 - y_4 + y_5 - y_6 - y_7 + y_8$$

Then the main effect of A will be:

$$\begin{aligned} A = \frac{1}{4}(- \ & y_2 - y_3 + y_4 - y_5 + y_6 + y_7 - y_8 \\ + \ & y_2 - y_3 + y_4 - y_5 + y_6 - y_7 + y_8) = \\ = \frac{1}{4}(& -2y_3 + 2y_4 - 2y_5 + 2y_6) \end{aligned}$$

From which it follows that in this case $V(A)=\sigma_y^2$, which is double that obtained when all the response values have been obtained experimentally.

Additionally, as said above, when some response values have been estimated, the effects are not independent. Following with the previous example we have:

$$B\ =\frac{1}{4}(-2y_2+2y_4-2y_5+2y_7)$$

And therefore:

$$V(A+B)=V\left[\frac{1}{4}(-2y_2-2y_3+4y_4-4y_5+2y_6+2y_7)\right]=$$

$$=3\sigma_y^2$$

As $V(A+B)=V(A)+V(B)+2\mathrm{Cov}(A,B)$ it immediately follows that $\mathrm{Cov}(A,B)=0.5\sigma_y^2$.

Xampeny et al.[4] provide recommendations on which runs to omit and how to estimate them when not all of them can be done for all two level 8 and 16 runs factorial designs containing contrasts formed only by interactions of three or more factors. When these recommendations, detailed in table 3, are followed, the effects are estimated with the following properties: 1) same variance for all of them, 2) minimum increase in variance compared to what would occur without missing values, and 3) minimum value of the correlation between effects.

This approach for saving runs has also disadvantages, and the objective of this article is to quantify them. This is done by simulation: a series of scenarios are proposed (varying the numbers and values of the significant effects) and, in each of them, we compare the number of errors made in the analysis of the importance of the effects when all the runs are available with the number of errors when some runs have been estimated following the recommendations in Table 3.

Below are detailed which scenarios these are, the methods followed for the assessing the significance of the effects, the results obtained and, finally, the conclusions that can be drawn from this work.

Table 3: Recommended runs to skip for obtaining effects with minimum variance and the same for all of them[4].

Design	Results to estimate	Missing runs recommended	Example of missing runs and interactions used	How to estimate the missing responses
2^3	1	Unimportant	Either ABC	From equating to zero the null interaction.
2^4	1	Unimportant	Either ABC, ABD, ACD, BCD, ABCD	Mean of the 5 values obtained from equating to zero the 5 null interactions.
	2	Pairs that can be estimated with 2 systems of 2 equations using 4 null interactions	y_6, y_{12} First System: ABC, ACD Second System: ABD, BCD	For each missing value: Mean of the two results obtained with two systems of equations
	3	Trios that can be estimated with 4 systems of 3 equations using only 4 interactions	y_1, y_4, y_5 First System: ACD, BCD, ABCD Second System: ABD, BCD, ABCD Third system: ABD, ACD, ABCD Fourth system: ABD, ACD, BCD	For each missing value: Mean of the four results obtained solving four systems of equations
	4	Subset of the quartets that can be estimated using only 4 null interactions	y_1, y_4, y_6, y_7 ABD, ACD, BCD, ABCD	Results obtained from a single system of four equations
	5	Subset of the quintets that can be estimated with a system of 5 equations	$y_1, y_4, y_6, y_{10}, y_{15}$ ABC, ABD, ACD, BCD, ABCD	Results obtained from a single system of five equations
2^{6-2}	1	Unimportant	Anyone The two negligible contrasts	Mean of the two results obtained from each null contrast
	2	Any of the 64 pairs of missing values that can be estimated with two null interactions.	y_1, y_3* The two negligible contrasts	Results obtained from a system of two equations
2^{7-3}	1	Unimportant	Anyone The negligible contrast	From equating the null contrast to zero

* With generators E = ABC and F = BCD

2. Simulation scenarios

To study the probabilities of error in the analysis of the significance of the effects, we have proposed a series of scenarios that aim to represent the most common situations that the experimenter can encounter. These scenarios consider that part of the effects are null: that is, that their values belong to a distribution of $N(\mu = 0;\ \sigma_{ef})$. The rest have an average 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.*[5], the values of Δ are called Spacing and they vary from 0.5 to 8 in increments of 0.5.

For 8 run designs, we consider the 4 scenarios that were already used by Fontdecaba *et al.*[6] to analyze the behavior of Lenth's[7] method.

$S8_1$: $\mu_1 = \Delta,\ \mu_2 = \cdots = \mu_7 = 0$
$S8_2$: $\mu_1 = \mu_2 = \Delta,\ \mu_3 = \cdots = \mu_7 = 0$
$S8_3$: $\mu_1 = \mu_2 = \mu_3 = \Delta,\ \mu_4 = \cdots = \mu_7 = 0$
$S8_4$: $\mu_1 = \Delta,\ \mu_2 = 2\Delta,\ \mu_3 = 3\Delta,\ \mu_4 = \cdots = \mu_7 = 0$

And for 16 run designs we consider those that were used for the first time by Venter and Steel[8], then later also by Ye *et al.*[5] and by Fontdecaba *et al.*[6]:

$S16_1$: $\mu_1 = \Delta, \mu_2 = \cdots = \mu_{15} = 0,$
$S16_2$: $\mu_1 = \mu_2 = \mu_3 = \Delta,\ \mu_4 = \cdots = \mu_{15} = 0$
$S16_3$: $\mu_1 = \cdots = \mu_5 = \Delta,\ \mu_6 = \cdots = \mu_{15} = 0$
$S16_4$: $\mu_1 = \cdots = \mu_7 = \Delta,\ \mu_8 = \cdots = \mu_{15}\ = 0$
$S16_5$: $\mu_1 = \Delta, \mu_2 = 2\Delta,\ \mu_3 = 3\Delta,\ \ \mu_4 = \cdots = \mu_{15} = 0$
$S16_6$: $\mu_1 = \Delta,\ \mu_2 = 2\Delta,\ \mu_3 = 3\Delta,\ \mu_4 = 4\Delta,\ \mu_5 = 5\Delta,\ \mu_6 = \cdots = \mu_{15} = 0,$

From the model provided by each scenario, the factors' effects are obtained by simulation. They are analyzed below to identify those that are considered significant. By comparing the results of this analysis with the coefficients of the model, the errors committed are identified.

For the missing values, we proceed as follows: From the values generated for the effects and an arbitrary value for the mean we calculate the response values. Then, the response values that are considered missing are replaced by their estimates – which are calculated through the established procedure in each case. Finally, we calculate the effects again and analyze their significance.

For example, if the values of the randomly generated effects in scenario S8$_3$ with a Spacing value of $\Delta = 5$ are:

A	B	C	AB	AC	BC	ABC
5.25	-4.32	6.07	-0.50	-0.68	-0.27	1.39

then, by assessing their significance by means of their representation in a Normal Probability Plot (NPP) (Figure 1, left), the effects that are truly different from zero (A, B and C) appear as significant and, therefore, in this case no error would be made.

From the values of the effects and with an average equal to 100 (arbitrary value), the following responses are obtained (in the standard order of the design matrix):

i:	1	2	3	4	5	6	7	8
y_i:	95.76	102.22	93.60	96.28	102.81	107.85	97.33	104.15

As in a 2^3design, it does not matter which run we do not perform, we randomly choose one of the response values and consider it missing, for example y_4. Next, by equating to zero the expression of the interaction ABC, we estimate its value and in this case we obtain $\hat{y}_4 =$ 101.84. With this estimated response we calculate the effects, and we get:

A	B	C	AB	AC	BC	ABC
6.64	-2.93	4.68	0.89	-0.71	-1.66	0.00

By ignoring the existence of a certain correlation among the effects and excluding the ABC interaction whose equal to zero value has been forced and, therefore, does not represent the variability of the null effects, we have represented these values in NPP (Figure 1, right), and only the effects A and C appear to be significant. Therefore, a type II error is committed, since in reality B is different from zero.

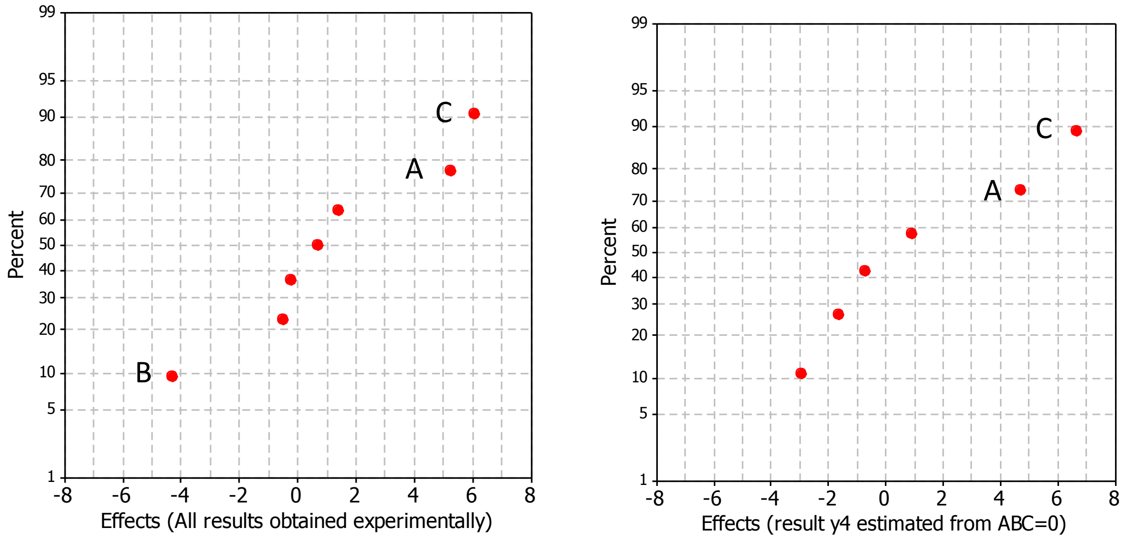

Figure 1: *Analysis of the significance of the effects in a* 2^3 *design with all the responses obtained experimentally (left) and estimating one of them by equating the ABC interaction to zero.*

For each design and for each number of missing runs, we select which ones to skip by following the recommendations in Table 3. In each case, and for each scenario and spacing value, 10,000 situations are simulated; and for each one of those, the percentage of type I and type II errors that have been committed are determined. Table 4 summarizes the conditions under which the simulations are carried out.

Table 4: *Summary of the simulations carried out*

	Design			
	2^3	2^4	2^{6-2}	2^{7-3}
Num. of missing runs	1	From 1 to 5	1 and 2	1
Runs to skip / Interactions used	Those indicated in Table 3			
Scenarios:	From 1 to 4	From 1 to 6	From 1 to 6	From 1 to 6
Simulations:	10,000			

Since in the above example (scenario $S8_3$) there are four null effects, up to 4 type I errors can be made. Therefore, the 10,000 simulations provide opportunities for 40,000 type I errors. On the other hand, having 3 non-null effects there are 30,000 type II error opportunities. After applying the Lenth method with the value of *t* proposed by Ye and

Hamada[9], and once all the experiments have been carried out, the results indicated in Table 5 are obtained. These values will be compared with those obtained when there are estimated response values.

Table 5: Error types produced in the 10,000 simulations of the values of the effects in configuration $S8_3$ using the Lenth's Method with Δ = 3.

Type I error		Type II error	
Absolute value	Percentage	Absolute value	Percentage
641	$\frac{641}{40000}100 = 1{,}60$	20928	$\frac{20928}{30000}100 = 69{,}76$

The problem lies in how to assess the significance of the effects automatically in such a way that it can be implemented in the simulation programs. This issue is dealt with in the following section.

3. Assessing the significance of the effects

Among the disadvantages involved in using estimated responses, first place is given to the great difficulty in assessing the significance of the effects. When all the runs have been carried out, this task can be done, as many software packages do, either using the variability of effects based on the values of those that can be considered null or by using the method of Lenth[7]. It also can be done manually representing the effects in a Normal Probability Diagram (NPP), a task that requires the analyst's judgment. An analysis of how some well-known statistical software packages address the issue of assessing the significance of the effects can be found in Fontdecaba *et al.*[10].

Neither of the three methods is appropriate in our case. The judgement by representing the effects in NPP cannot be automated. Nor can we estimate the variance of the effects using those considered null, since they have been used to deduct the missing values. And with respect to Lenth's method is based on if $X \sim N(0, \sigma)$, then the median of $|X|$ is equal to 0.6475 and thus $1.5 \cdot \text{median}|X| = 1{,}01\sigma \cong \sigma$. This circumstance is exploited in order to define $s_0 = 1.5 \cdot \mathrm{median}|c_i|$, where c_i are the values of the effects. Naturally, s_0 is not a good estimator of σ_{ef}, since the values of the active effects also intervene in its calculation. To eliminate them, a new median is calculated by excluding the values $|c_i| > 2.5 s_0$. In this way we get what is called the Pseudo Standard Error (PSE), from which is defined an interval of $0 \pm t \cdot PSE$ that contains the effects that are considered inert and where t depends on the confidence level and number of effects being considered. The procedure is very attractive both for its simplicity and for being well-known and commonly used. It is based on if $X{\sim}N(0,\sigma)$, then the median of $|X|$ is equal to 0.6475 and thus $1.5 \cdot \text{median}|X| = 1{,}01\sigma \cong \sigma$. This circumstance is exploited in order to define $s_0 = 1.5 \cdot \text{median}|c_i|$, where c_i are the values of the effects. Naturally, s_0 is not a good estimator of σ_{ef}, since the values of the active effects also intervene in its calculation. To eliminate them, a new median is calculated by excluding the values $|c_i| > 2.5s_0$. In this way we get what is called the Pseudo Standard Error (PSE), from which is defined an interval of $0 \pm t \cdot PSE$ that contains the effects that are considered inert and where t depends on the confidence level and number of effects being considered. The above only holds if effects are independent, which never occurs when there are missing values. In addition, if the effects whose values have been forced to zero are excluded, the probabilities of error increase rapidly when considering less than 7 effects. On the other hand, including effects whose values have been forced to zero decreases the PSE, which also leads to major errors.

Hamada and Balakrishnan[11] discuss and compare a great variety of procedures for assessing the significance of the effects in factorial designs without replicas. From among all of them, we have chosen the Bayesian approach of Box-Meyer, since it is a recognized method that is not restricted to a specific number of effects and does not require independence. In addition, there is an R package that allows it to be applied automatically.

The method of Box and Meyer (1986[12], 1993[13]) considers the set of all possible models: $M_0, M_1, \cdots, M_m$ that can be contemplated. The value of m is equal to $2^a - 1$, with a being the number of effects that are going to be analyzed. So for example, in a 2^3design 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 requires using the Bayes theorem to determine the probability of each model M_i, given the response vector $\boldsymbol{y}$. In other words:

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

The calculation of $p(M_i)$ is simple. 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 the model M_i, then we have $p(M_i) = \pi^{f_i}(1-\pi)^{N-f_i}$. The value of π must be previously fixed. Box and Meyer propose the value of 0.25 and that is the one we have used.

For calculating $f(\boldsymbol{y}|M_i)$, it is necessary to assign an a priori distribution for the values of the effects. Box and Meyer propose using $N(0,\gamma^2\sigma^2)$, where the mean is 0 due to the direction of each effect being unknown a priori and the magnitude of the effect relative to the experimental noise is captured through the parameter γ. By also following the suggestion of these authors for each case, we have taken the value of γ that minimizes the probability that all effects are null. The expression of $f(\boldsymbol{y}|M_i)$ and the details of deducing it can be seen in the Appendix of the second article of Box and Meyer[13].

Barrios[14] has developed the BsMD package for R[15] that allows determining the probabilities $p(M_i|\boldsymbol{y})$. By introducing the design matrix, the response vector and the values π and γ, a list of models is obtained in order of their assigned probability. The effects that the model contains are those most likely to be taken as significant.

We have established the reference of what would happen in the case of no missing runs, by using both methods: Box and Meyer and Lenth. There is controversy about which values of t should be used. For a confidence level of 95%, Lenth proposed the values of 3.76 and 2.57 for designs with 8 and 16 experiments, respectively. These values have been discussed by authors such as Loughin[16], Ye and Hamada[9], and Fontdecaba *et al.*[6], all of whom show that a type I error closer to 5% is obtained and that there is a notable decrease in type II errors when using lower values of t. In our study, we used the values proposed by Ye and Hamada: 2.297 and 2.156 for 8 and 16 experiments respectively.

4. Results in 2^3 designs

As an a priori estimate of the proportion of active effects, we have used Box and Meyer's[13] recommended value of $\pi = 0.25$. When we have a missing response value, forcing an effect to be null leads to think that the proportion of active effects will be greater. However, we have also tested with a value of $\pi = 0.30$, and the results do not improve. Therefore, we have maintained the same value regardless of whether we have all the responses or there is a missing value.

Choosing the value of γ is more complicated. In their first article, Box and Meyer use a different metric that they call k, which is related to γ in the form of $k^2 = n\gamma^2 + 1$, where n is the number of experiments that the design requires. After analyzing a set of cases in this first article, they observe that the values of k vary between 2.7 ($\gamma = 0.89$) and 18 ($\gamma = 6.35$); so they propose using the value of $k = 10$ $(\gamma = 3.52)$, because it is a round number that represents approximately the average of the observed values. In their second article they propose choosing the value of γ that minimizes the probability of obtaining a model with all the effects null; and this is the criterion we have used.

To determine those values of γ, we simulated 1000 cases for each Scenario-Spacing combination, which identified for each case the value of γ that minimizes the probability that all effects are null. The value chosen for each Scenario-Spacing combination is the average of the 1000 values obtained. The calculations were made with the help of the `BsSProb` function included in the `BsMD` package of R that calculates the probability associated with each of the models that can be proposed. In each case, probabilities have been evaluated for 20 values of γ that are equidistant within the range of $\gamma = 0.5$ to $\gamma = 10$, which is wider than the one proposed by Box and Meyer in their first article. The values obtained are those we have used in our study, and they are shown in Table 6.

Table 6: Values of γ *used in each Spacing-Scenario combination of values for* 2^3 *designs.*

Spacing	Scenario			
	1	2	3	4
0.5	0.74	0.79	0.76	0.77
1	0.75	0.75	0.77	0.96
1.5	0.81	0.87	0.73	1.33
2	0.89	0.88	0.79	1.67
2.5	1.08	1.09	0.84	2.17
3	1.22	1.24	1.01	2.64
3.5	1.45	1.38	1.15	3.11
4	1.64	1.79	1.28	3.57
4.5	1.79	2.00	1.51	4.24
5	2.02	2.18	1.83	4.56
5.5	2.22	2.44	2.07	5.03
6	2.42	2.72	2.36	5.57
6.5	2.58	2.92	2.82	5.85
7	2.80	3.20	3.02	6.25
7.5	2.96	3.36	3.25	6.64
8	3.21	3.67	3.52	7.04

Instead of previously calculating average values of γ, we could have calculated its value in each case. However, we have verified that the best one obtained is not relevant and doing it in this way greatly extends the computing time, especially when working with 16 experiments in which for each value of γ it is necessary to calculate the probability of the 2^{15} models that can be built.

Figure 2 shows the obtained results and also includes – for reference – those of the Lenth method when there are no missing values. The differences are barely noticeable for type I errors and are not relevant for type II errors, especially when using the results of the Lenth method as a reference.

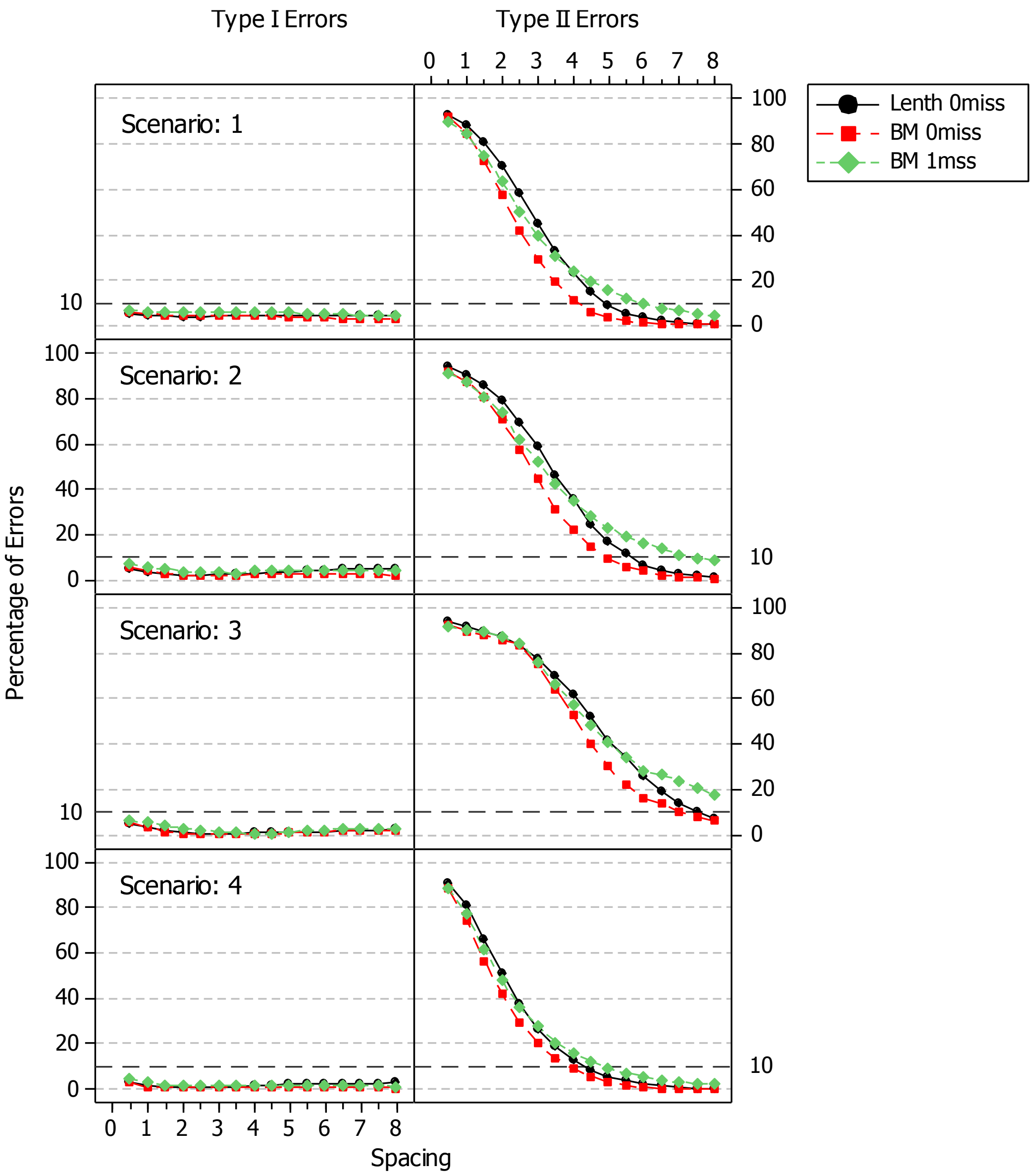

Figure 2: 2^3 *Designs. Percentage of effects for which a type I and type II error is committed in the analysis of their statistical significance. Without missing values (Lenth and Box-Meyer method) and with one missing value (Box-Meyer).*

5. Results in 2^4 designs

The same procedure has been applied as for 2^3 designs. The value of $\pi = 0.25$ has also been taken and the values of γ are the average of those obtained by performing 1000 simulations in each Scenario-Spacing combination. To find the value that minimizes the probability that all effects are null in this case, the range of γ values is 0. 5 to 8 (also slightly wider than the one proposed by Box and Meyer). The values obtained for γ are those listed in Table 7.

Table 7: *Values of γ used in each Spacing-Scenario combination of values for* 2^4 *designs*

Spacing	Scenario					
	1	2	3	4	5	6
0.5	0.53	0.54	0.54	0.53	0.54	0.58
1	0.54	0.54	0.53	0.53	0.62	0.86
1.5	0.55	0.55	0.54	0.52	0.86	1.35
2	0.57	0.58	0.55	0.51	1.12	1.80
2.5	0.61	0.65	0.58	0.52	1.37	2.26
3	0.66	0.71	0.63	0.54	1.69	2.66
3.5	0.72	0.87	0.74	0.57	1.98	3.11
4	0.78	0.97	0.91	0.58	2.23	3.62
4.5	0.87	1.14	1.04	0.66	2.59	4.08
5	0.95	1.34	1.26	0.74	2.84	4.58
5.5	1.06	1.45	1.46	0.90	3.12	4.92
6	1.18	1.56	1.67	1.12	3.44	5.44
6.5	1.26	1.69	1.88	1.39	3.67	5.75
7	1.38	1.82	1.98	1.62	4.00	6.11
7.5	1.48	1.97	2.16	1.74	4.32	6.40
8	1.59	2.12	2.28	1.98	4.57	6.74

The results obtained (Figure 3) show that the percentage of type I errors increases, in general, when the number of missing values increases. However, it remains at values below 10%, except in the worst case of 5 missing values (Scenario 1), where it rises to around 15%. Regarding the proportion of type II errors, the increase is either not relevant or it even drops, except with 4 missing values, in which case it clearly increases. In scenario 4, a singular behavior occurs in which it remains above 80% even for high Spacing values, especially for 4 missing values.

6. Results in other designs

In 2^{6-2} designs with the right generators, for example $E = ABC$ and $F = BCD$, there are two contrasts in which only interactions of 3 or more factors intervene. Therefore, values of 1 or 2 missing values can be estimated. The results obtained with this design are summarized in Figure 4. It can be seen that type I errors are maintained at similar values in all scenarios and the same is true for type II errors in scenarios 5 and 6. In scenarios 1-3, there is an increase in the proportion of type II errors when there are missing values, although only for some spacing values. In scenario 4, with spacing values between 3.5 and 5.5, the Box-Meyer method performs poorly both with and without missing values.

For 2^{7-3} designs with a missing value (Figure 5) the result is similar to the one we just discussed.

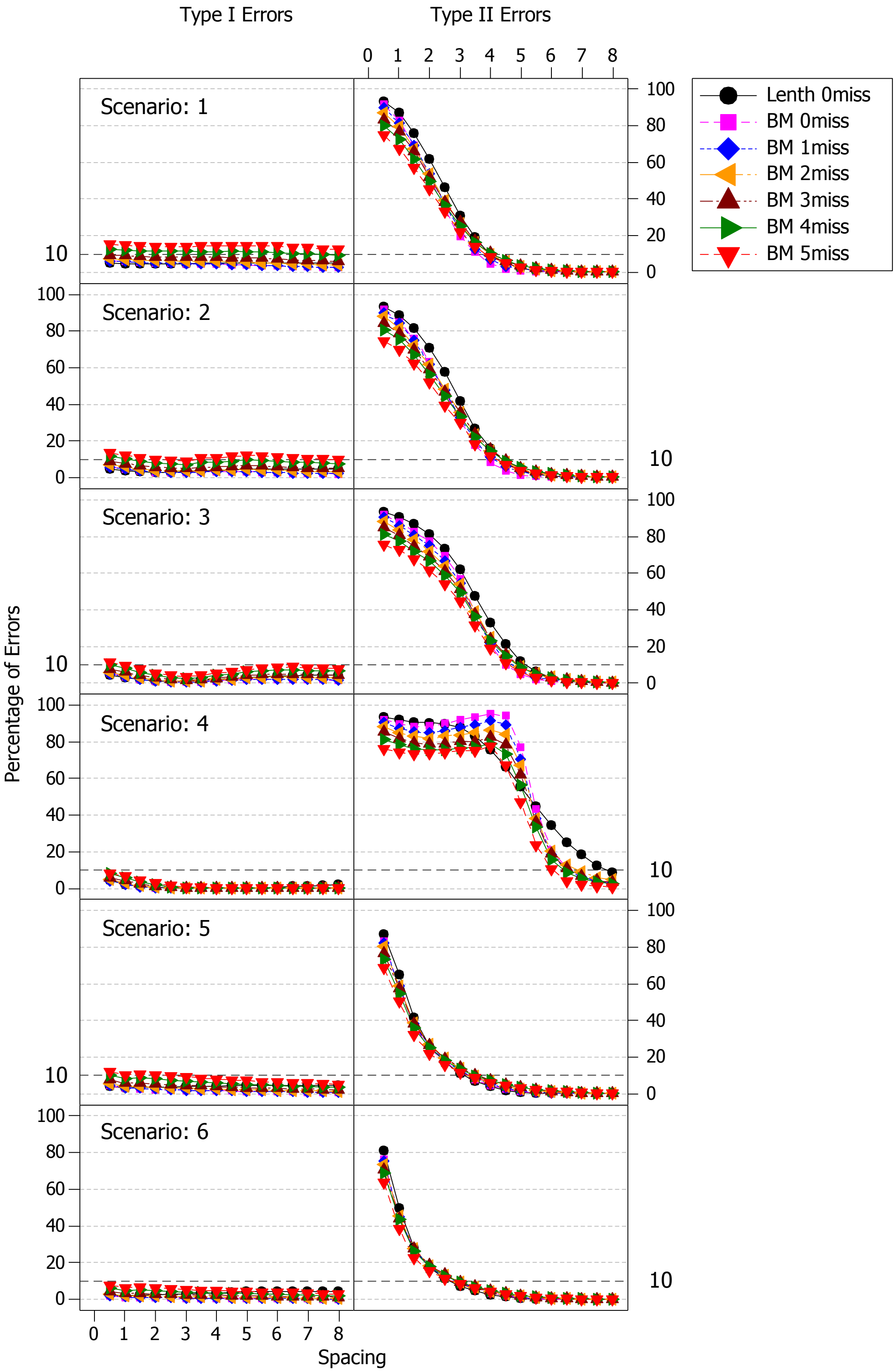

Figure 3: 2^4 *Designs. Percentage of effects for which a type I or type II error is committed in the analysis of its statistical significance with and without missing values. Box-Meyer method.*

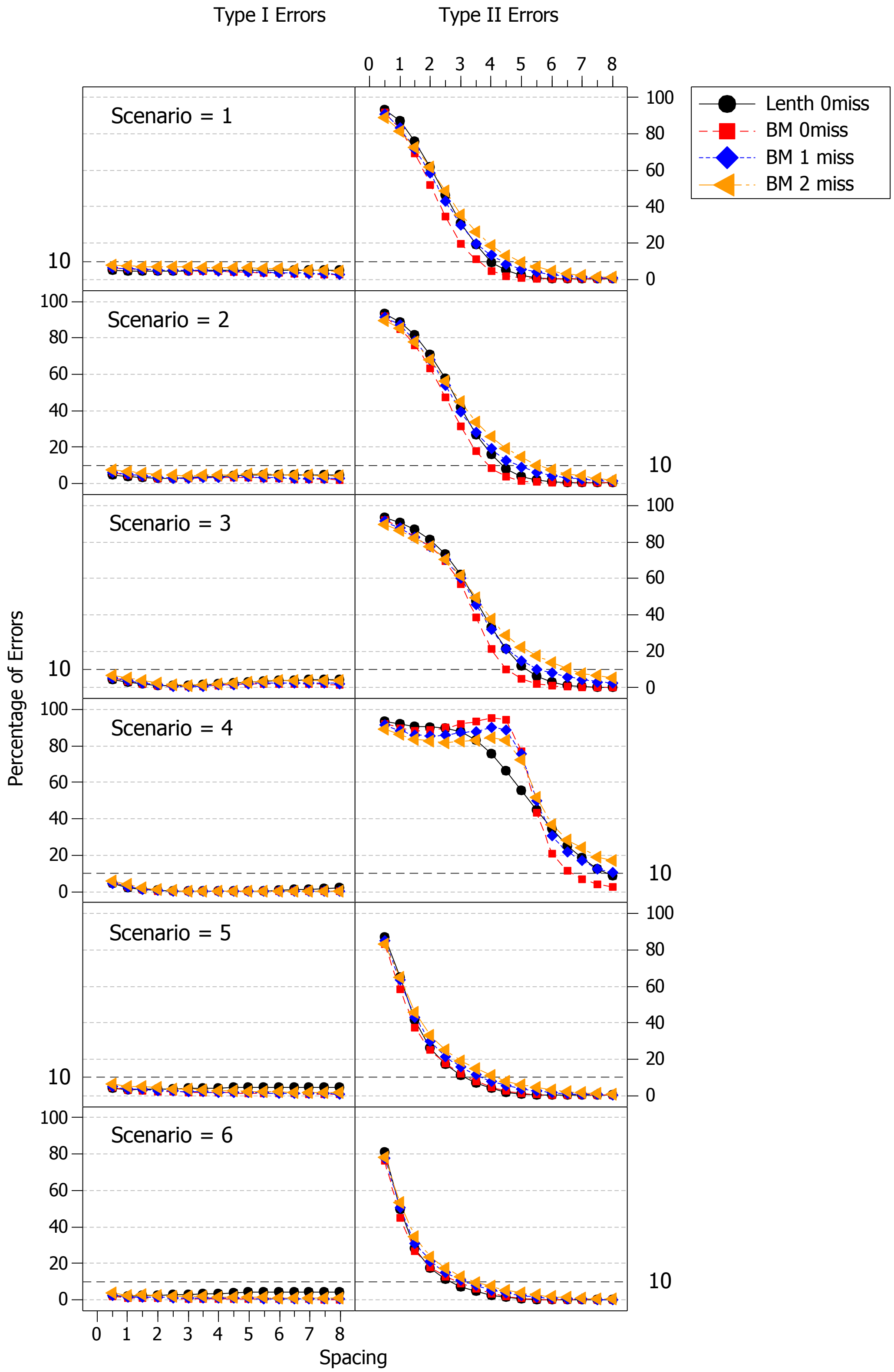

Figure 4: 2^{6-2} *Designs. Percentage of effects for which a type I or type II error is committed in the analysis of its statistical significance with and without missing values. Methods of Lenth and of Box-Meyer.*

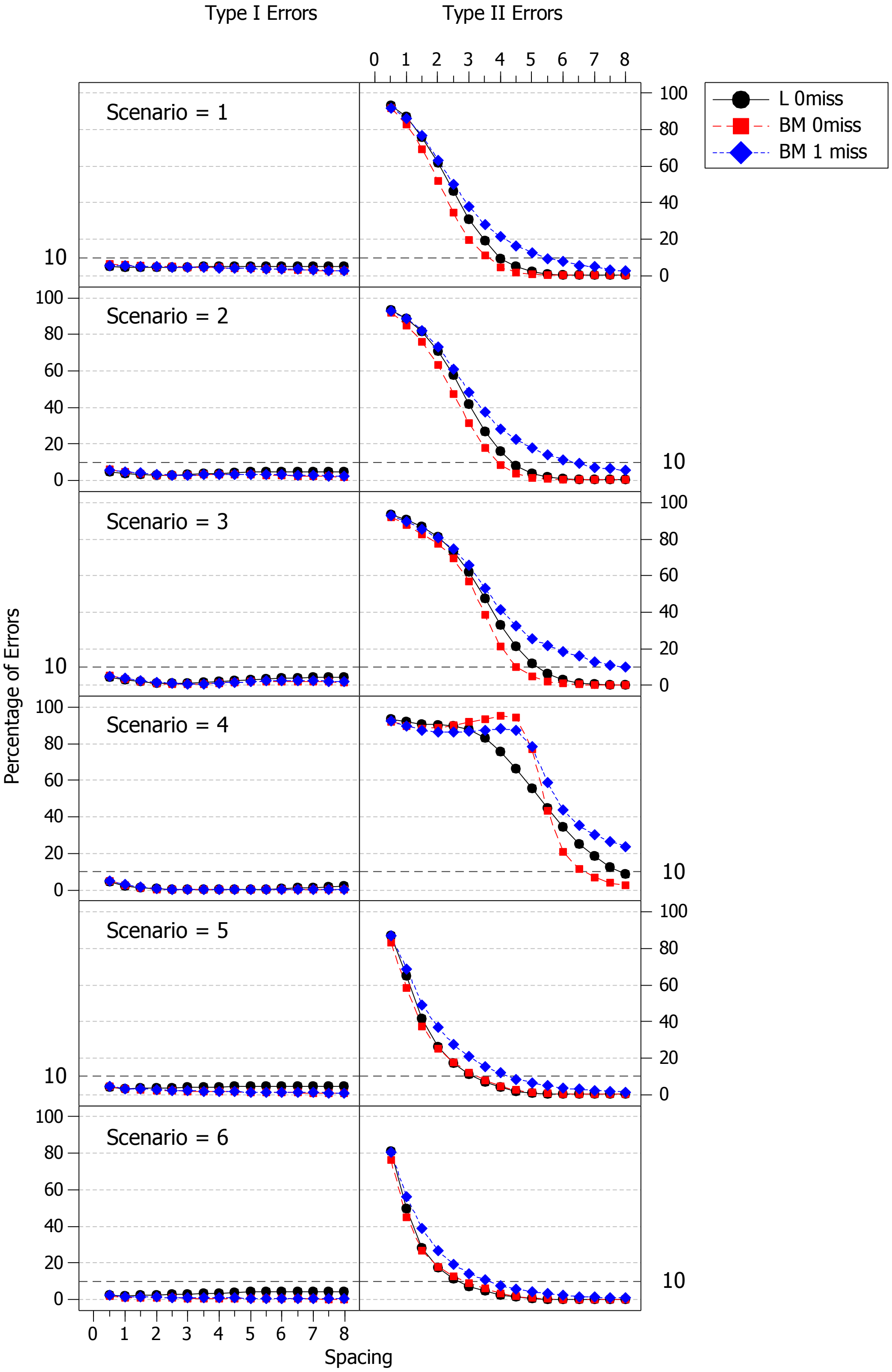

Figure 5: 2^{7-3} *Designs. Percentage of effects for which a type I or type II error is committed in the analysis of its statistical significance with and without missing values. Methods of Lenth and of Box-Meyer.*

7. Summary and conclusions

We have studied the increase in the probability of committing type I and type II errors in assessing the significance of the effects in 8 and 16 run designs when some properly selected runs have not been carried out and their responses have been estimated from the interactions considered null from scratch.

The only 8 run design with a suitable interaction is the 2^3 design. In it a missing response value can be estimated by clearing its value from the expression of the ABC interaction – which is considered null – equated to zero. The problem that arises is that the variance of the estimated value is greater than that of the values obtained directly from the experimentation; and this in turn causes a greater variance of the effects that, moreover, cease to be independent.

In 16 run designs there are more possibilities. The 2^4 allows to estimate up to 5 missing values since there are 5 interactions of 3 or more factors that can be considered null. In addition, the 2^{6-2} design allows to estimate up to two missing values since it has two contrasts that only estimate interactions of 3 or more factors and the 2^{7-3} design has one suitable contrast and thus allows the estimation of one missing value. In these cases, the variance of the missing values depends on which runs have been skyped as well as which interactions are used and how they are used to perform the estimation. Xampeny *et al.* [4] have identified which is the best strategy in each case, and that is the one that has been followed in this work.

One consequence of having estimated values is that it complicates the task of assessing the significance of the effects. The degrees of freedom that could be used to estimate the effect variance are used to estimate the missing values and, therefore, this method cannot be used. The conditions for applying Lenth's method are not met either, and therefore using it would lead to important errors. A good possibility that we have used in this paper is the Box-Meyer method.

Another consequence is the greater probability of error when assessing the significance of the effects. By analyzing simulations – in a wide variety of situations – of the proportion of type I and type II errors that have been discussed, our conclusions are:

- Estimating one response value, no matter which one, in 2^3 designs is barely noticeable in terms of the difference in the proportion of type I errors. For type II errors, the difference is slightly bigger but hardly relevant. The analysis also serves to show the good performance of the Box-Meyer method compared to Lenth's. It is interesting to note that the proportion of errors when applying the Lenth method without missing values is approximately the same as when the Box-Meyer method is applied to a 2^3 design with one estimated value.

- In 2^4 designs, working with up to 3 missing values does not produce relevant changes in the proportion of errors, whether they be type I or type II. With 4 and 5 missing values, there is indeed an increase in the proportion of errors – whether they be type I, type II, or both.
- In 2^{6-2} designs when a single missing value is estimated, the results hardly change. When in this same design two missing values are estimated – or one is estimated in a 2^{7-3} design – the increase in the proportion of errors is indeed noticeable, especially in some scenarios and for certain spacing values.

George Box said[2] "do not rely on your results if you have too many missing observations. Usually, I would start to feel uncomfortable with the analysis when there was more than one missing observation in an 8-run experiment, or more than two observations missing from a 16-run experiment". Box refers to situations in which the number of runs has not been planned or there is a result suspected of being anomalous and which one prefers to disregard. Our results are consistent with this statement, and we can add that if one can choose the missing runs, up to 3 runs can be omitted in 16-run designs.

References

1. Draper NR, Stoneman DM. Estimating missing values in unreplicated two-level factorial and fractional designs. *Biometrics* 1964; 20(3):443-458. DOI: 10.2307/2528487.
2. Box GEP. George's Column: A simple Way to Deal with Missing Observations from Designed Experiments. *Quality Engineering* 1990; 3(2):249-254.
3. Xampeny R, Grima P, Tort-Martorell X. Estimating missing values from negligible interactions in factorial designs. *Quality and Reliability Engineering International* 2017; **33**(6):1235-1247. DOI: 10.1002/qre.2172.
4. Xampeny R, Grima P, Tort-Martorell X. Which runs to skip in two level factorial designs when not all can be performed. Submitted (In advanced process of review) to *Quality Engineering* (to be updated before publication)
5. Ye KQ, Hamada M, Wu CFJ. A step-down Lenth method for analyzing unreplicated factorial designs. *Journal of Quality Technology* 2001; **33**(2), 140-152.
6. Fontdecaba S, Grima P, Tort-Martorell X. Proposal of a single critical value for the Lenth method. *Quality Technology and Quantitative Management* 2015, 12(1):41–51.

7. Lenth RV. Quick and easy analysis of unreplicated factorials. *Technometrics* 1989, 31(4):469-473

8. Venter JH, Steel SJ. Identifying active contrasts by stepwise testing. *Technometrics* 1998, **40**(4):304-313

9. Ye KQ, Hamada M. Critical values of the Lenth method for unreplicated factorial designs*. Journal of Quality Technology* 2000, **32**(1):57–66.

10. Fontdecaba S, Grima P, Tort-Martorell X. Analyzing DOE with satistical software sackages: controversies and proposals. *The American Statistician* 2014, **68**(3):205–211.

11. Hamada M, Balakrishnan N. Analyzing unreplicated factorial experiments: a review with some new proposal, *Statistica Sinica* 1998, 8(1):1–41.

12. Box GEP, Meyer RD. An Analysis for Unreplicated Fractional Factorials. *Technometrics* 1986, **28**(1):11-18.

13. Box GEP, Meyer RD. Finding the Active Factor in Fractionated Screening Experiments. *Journal of Quality Technology* 1993, **25**(2):94-105.

14. Barrios E (based on Daniel Meyer's code). BsMD: Bayes Screening and Model Discrimination. R package version 2013.0718 (2013). https://CRAN.R-project.org/package=BsMD

15. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2016). https://www.R-project.org/

16. Loughin TM, Calibration of the length test for unreplicated factorial designs. *Journal of Quality Technology* 1998, **30**(2):171-175.