Critical values of the Lenth method. A new proposal

Sara Fontdecaba[*], Pere Grima and Xavier Tort-Martorell

Department of Statistics and Operational Research, Universitat Politècnica de Catalunya - BarcelonaTech, Barcelona, Spain

Abstract: Different critical values deduced by simulation have been proposed that greatly improve Lenth's original proposal. However, these simulations assume that all effects are zero – something not realistic-- producing bigger than desired critical values and thus significance levels lower that intended. This article, in accordance with George Box [2] well known idea that Experimental Design should be about learning and not about testing and based on studying how the presence of a realistic number and size of active effects affects critical values, proposes to use t = 2 for any number of runs equal or greater than 8. And it shows that this solution, in addition of being simpler, provides under reasonable realistic situations better results than those obtained by simulation.

Keywords: Industrial experimentation; Lenth method; Pseudo Standard Error; Significant effects; Unreplicated factorial designs.

1. Introduction

Identifying which are the active factors is one of the steps required in analyzing the results of a factorial design. When replicates are available, the experimental error variance can be used to estimate the variance of the effects and judge whether they are active or not. In other cases it can be assumed that three or more factors interactions are zero and then their values can be used to estimate the effect's variance. Unfortunately, this estimate is made with very few degrees of freedom unless we are dealing with full factorial designs with five or more factors, or fractional designs with resolution greater than five, unusual situations in industrial environments where the number of runs is generally limited.

Daniel [5] proposed a procedure based on the representation of the absolute value of the effects on a half-normal plot, since the non-active effects are distributed according to

[*] Corresponding author: sara.fontdecaba@upc.edu

$N(0; \sigma_{ef})$ in these coordinates they are aligned approximately along a straight line that passes through the origin. A variant of this method, also used by Daniel [6], is to represent the effects with their sign on a Normal Probability Plot (NPP). In this case, the non-significant effects are aligned according to a straight line which passes through the point (0, 0.5).

Daniel's method, in the half-normal or NPP version, is the most used for analyzing the "significance" of effects in the absence of replicates; although, being based on a graphic interpretation, the conclusions that are drawn are subjective and may be debatable. What's more, their proper use requires a good deal of judgment and experience. If one simply applies the motto, "the active effects are those that are separated from the line," one can commit glaring errors, as described by De León et al. [8]. Another drawback from basing the observation on a graph is that it is not possible to automate the process of selecting active effects.

To solve these problems, Lenth [15] proposed a method for estimating the standard deviation of the effects on the basis that if $X \sim N(0, \sigma)$, the median of $|X|$ is equal to 0.6475σ and therefore $1.5 \cdot \text{median}|X| = 1.01\sigma \cong \sigma$. Considering that κ_i ($i = 1, \dots, n$) are the effects of interest and that their estimators c_i are distributed according to $N(\kappa_i, \sigma_{ef})$, he defines $s_0 = 1.5 \cdot \text{median}|c_i|$ and uses this value to calculate a new median by excluding the estimates $|c_i| > 2.5s_0$ with the intention of excluding those which correspond to effects with $\kappa > 0$. In this way he obtains the so-called *pseudo standard error*:

$$PSE = 1{,}5 \cdot \underset{|c_i|<2.5s_0}{\text{median}}|c_i|$$

Lenth verifies that the PSE is a σ_{ef} estimator which is reasonably good when the significant effects are rare (sparse) and he proposes calculating a margin of error for c_i by means of $ME = t_{1-\alpha/2,d} \cdot PSE$ for a confidence level of $1 - \alpha$ and where *t* is distributed according to a Student's t distribution with $d = n/3$ degrees of freedom. On the other hand, since n significance tests are performed simultaneously, calculating a simultaneous margin of error (*SME*) is also proposed, which is obviously greater than the *ME*. Finally, he proposes performing a graphical analysis rather than formal significance tests. He suggests graphics that are similar to those used by Ott [19] for the

analysis of means. If $c_i < ME$ should be considered non-significant, if c_i is larger than ME but smaller than SME could be described as possibly active and if $c_i > SME$, then the effect is probably active.

Since its publication, the Lenth method has gained prominence. In fact, it has become the most employed method, being described in the books which are most prevalent in the field of industrial experimental design, such as those by Montgomery [18] and Box, Hunter and Hunter [3]. It is also used by statistical software packages which are very common in the field of industrial statistics, such as MINITAB or SigmaXL that use the critical values originally proposed by Lenth or JMP which calculates critical values ad hoc for each case by simulation.

In our opinion the principal advantage of Lenth method is that it uses a simple and effective procedure (given the scarcity of the information available) to estimate the standard deviation of the effects. However, the critical values originally proposed are not the most appropriate and in many cases produce results that a representation in NPP show that are clearly incorrect. For example, Box, Hunter and Hunter [3] classical book provides a 2^{4-1} design example, page 237. The analysis is accompanied by the comment that it is reasonable to consider that factors A and B influence the response, which is very consistent with what is observed in the NPP representation of the effects. But, for example Minitab that uses Lenth's method with its critical values only presents A as significant (Figure 1).

Although the errors are less frequent when the number of runs increases, these also occur. The well-known example (Box [2]) of optimizing the design of a paper helicopter using experimental design techniques can be useful to illustrate that. The article presents the results of a 2^{8-4}, 16 run design, saying that factors B and probably C are active. Minitab's representation of the effects on NPP and identifications of the significant ones using Lenth's method with its original critical values only identifies B as active (Figure 2).
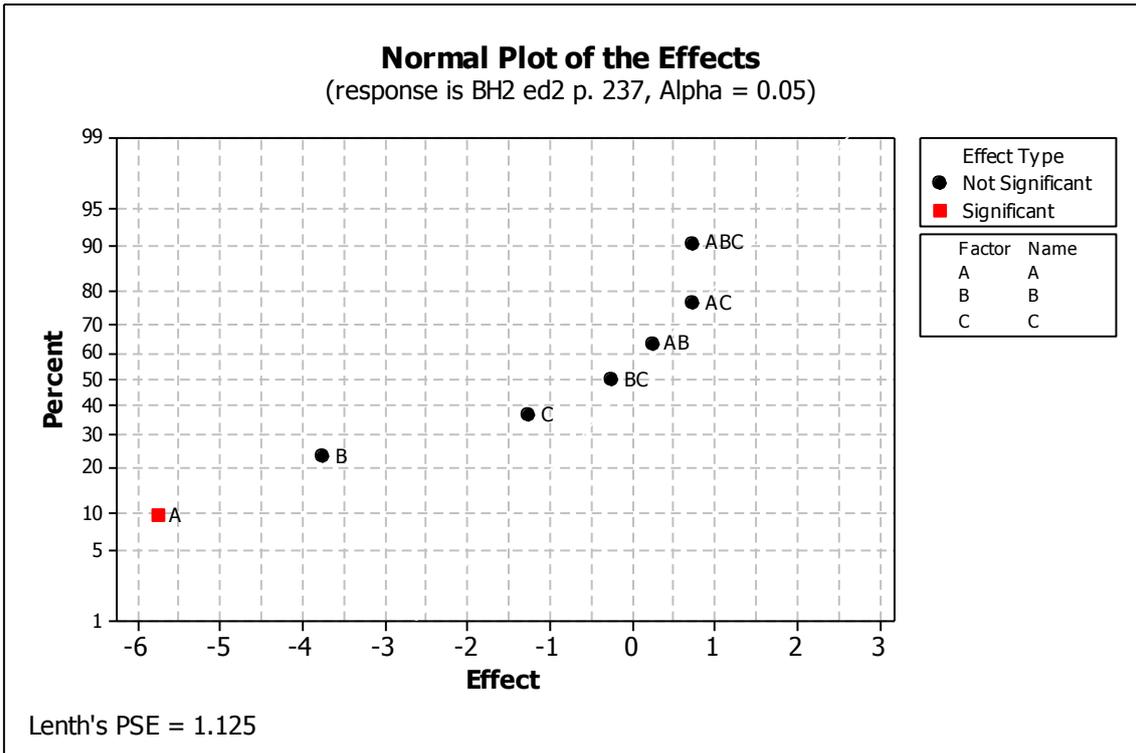
3

Figure 1. Effects of the Box, Hunter and Hunter example represented in NPP by Minitab. Significant ones identified by Lenth's method using its original critical values
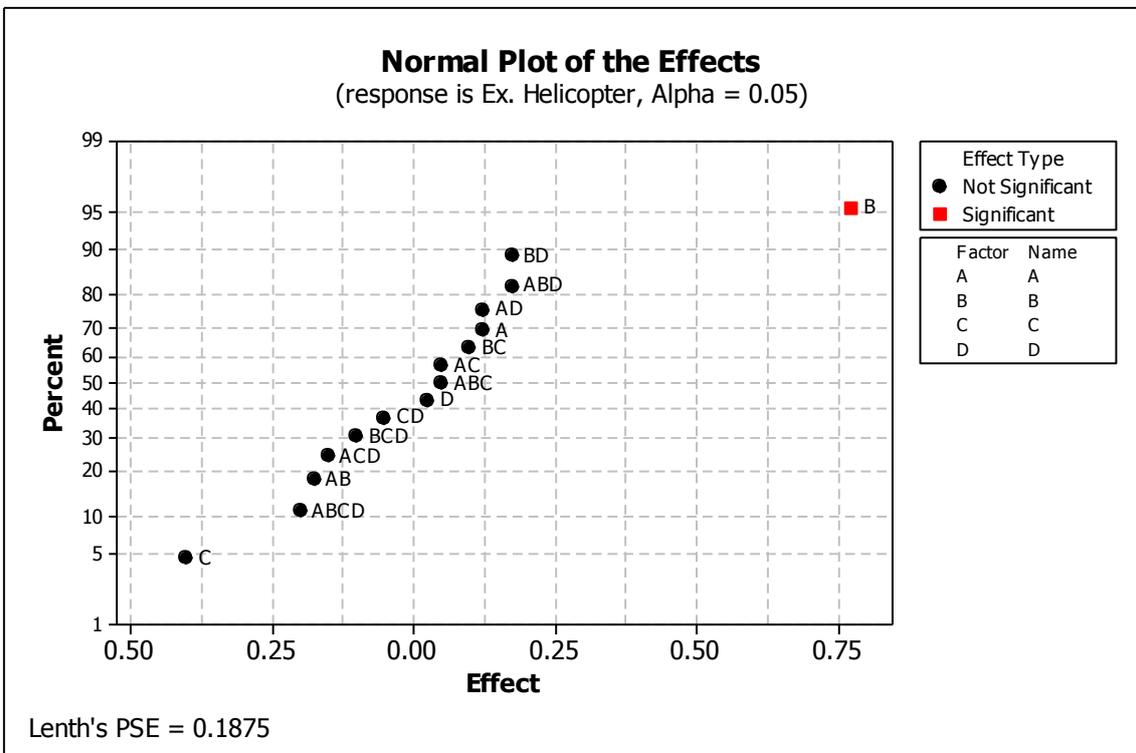

Figure 2. Effects of Box's paper helicopter represented by Minitab in NPP. Significant effects identified by Lenth's method using its original critical values

2. Alternatives to the Lenth method

Since its publication, many alternative methods to Lenth's proposal have appeared. Some, like Loughin [16], show that a Student's t distribution with $d = n/3$ degrees of freedom is not a good reference distribution for Lenth's t-ratio, and he obtained by simulation that for $\alpha = 0.05$, it is necessary to have $t = 2.300$ and $t = 2.152$ for 8 and 16 runs, respectively. Along these lines, and also by simulation, Ye and Hamada [21] determine the critical values for certain values of α by means of a computational method that is more efficient than Loughin. They also address three level and Plackett and Burman designs. In this case, the values obtained for $\alpha = 0.05$ and eight and sixteen-run designs are $t = 2.297$ and $t = 2.156$.

Other authors propose using different methods from Lenth but based on similar procedures, such as Dong [9], who instead of using the effects' median to calculate the σ_{ef} estimator, uses the average of the squares and states that when the number of significant effects is low (less than 20%) and its value is not small (its tests are performed with a $\kappa \geq 5\sigma_{ef}$ value of significant effects), his method delivers better results than that of Lenth. In a similar vein, Juan and Peña [14] deal with the problem by identifying outliers (significant effects) in a sample and, under certain circumstances, the Lenth method is a particular case of the one they propose. They claim that their method works better than Lenth when the number of active effects is large (greater than 20%), but it is more complicated.

Other strategies are those that may be called "step by step", like that of Venter and Steel [20], which, through a simulation study for sixteen-run designs (with the same configurations that we have used), conclude that their σ_{ef} estimator is as good as Lenth's PSE when there are few significant effects, and that it is much better when there are many. Ye, Hamada and Wu [22] also propose a method they call the "step-down version of the Lenth method". They begin by calculating the Lenth statistic (t-Lenth) for the biggest effect and, if it is considered significant, the PSE is recalculated with the remaining effects and a new t-Lenth is obtained for biggest of them. This is then compared with a critical value obtained by simulation in accordance with the number of effects considered. The procedure is repeated until the biggest effect remaining is not considered significant. After performing a simulation study, they conclude that their

results are better than those obtained with the original Lenth method or that of Venter and Stell [20]. Mee, Ford III and Wu [17] also propose an iterative system, but in this case of stepwise regression assuming that at least a certain number of non-significant effects exist.

Edwards and Mee [10] propose to calculate by simulation the p-value that corresponds to the t-ratios. To calculate the critical values they use a method similar to that of Ye and Hamada, but it can be applied to a wide variety of designs, particularly those which are not orthogonal. Bergquist et al. [1] propose a Bayesian procedure. They begin with a study of published cases to establish the *a priori* probabilities based on the principles of sparsity, hierarchy and heredity, and then illustrate their method by analyzing cases in the literature.

A proposal that we find particularly pragmatic is that of Costa and Lopes [4], which takes advantage of current computational capabilities. Since most of the methods are simple to apply, they propose using various methods. If all of them give the same solution, it is clear that this it is a good one. On the contrary if there is a discrepancy, the solution can be decided by a majority, perhaps with careful consideration of the method or of the characteristics of the situation being analyzed. When there is a great disparity between some methods and others, the solution is certainly not clear; not for any failure of the methods but for lack of information, and the only way out is to perform new experiments.

Detailed studies and comparisons of existing methods have also been published, such as Haalang and O'Connell [12]. After a simulation analysis of sixteen-run designs, they conclude that the Lenth method has good properties in a wide range of number and size of significant effects. Other alternatives, such as the previously mentioned Dong, are only recommended if the number of significant effects is known to be low. Another study, which is very complete and detailed, is that of Hamada and Balakrishnan [13], where they summarize and discuss 17 methods and also propose another 5 that are new proposals or modifications of existing ones. They advise against the use of some methods and conclude that among those which perform reasonably well, there is no clear winner.

3. Lenth method with improved critical values

In our opinion, of all the alternatives considered, the ones that best combine simplicity and precision are those, in accordance with the proposals of Ye and Hamada [21] or Edwards & Mee [10], based on comparing the *t*-ratios with reference distributions obtained through simulation.

The problem with these *t*-ratios and the critical values proposed is that PSE values tend to be higher than those obtained by simulation under the assumption that all effects are zero. In practice, some non-null effects ($\kappa \neq 0$) may be incorporated into the PSE calculation producing an overestimation of σ_{ef} and therefore affecting the critical values. This drawback was already highlighted by Lenth and even improved versions as the widely acknowledged proposed by Ye and Hamada [21] suffer from this problem. To further illustrate this phenomenon, Figure 3(a) shows the distribution of 10,000 PSE values obtained by simulation from a design with 8 runs, when all effects are values from a $N(0, 1)$ and Figure 3(b) gives the distribution obtained when there are five effects belonging to a distribution $N(0, 1)$ and two of a $N(2, 1)$. It can be clearly seen that the estimate of PSE tends to be higher in the latter case.
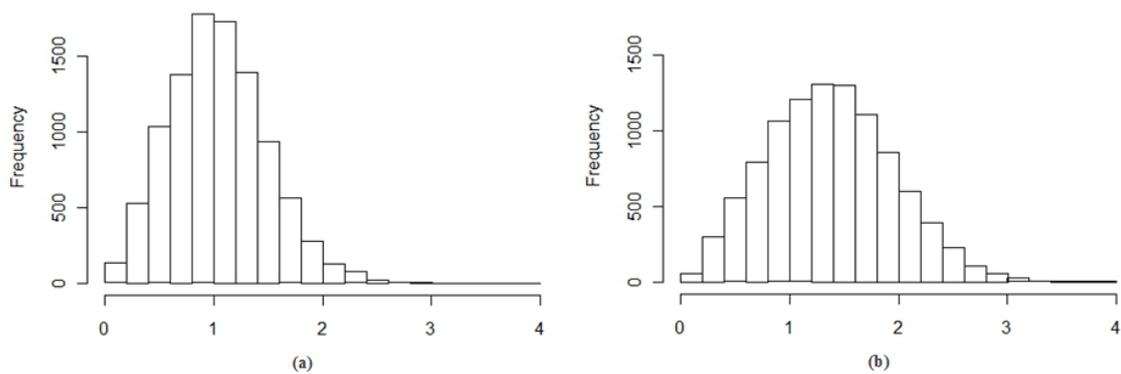
Figure 3. Distribution of values of the PSE in a design with 8 experiments. (a) when all 7 effects belong to a N(0,1) and (b) when five belong to the N (0,1) and two to N (2,1) (b).

Depending on the number and magnitude of non-null effects ($\kappa \neq 0$) the overestimation of σ_{ef} can be higher or lower. To show this fact we have simulated a set of situations that reflect what one would expect in practice and calculated their PSE. For 16-runs designs we have used the same 6 configurations that both Ye, Hamada and Wu [22] and Venter and Steel [20] use in their proposals. They are:

C1: $\kappa_1 = \cdots = \kappa_{14} = 0, \kappa_{15} = \Delta$

C2: $\kappa_1 = \cdots = \kappa_{12} = 0, \kappa_{13} = \kappa_{14} = \kappa_{15} = \Delta$

C3: $\kappa_1 = \cdots = \kappa_{10} = 0, \kappa_{11} = \cdots = \kappa_{15} = \Delta$

C4: $\kappa_1 = \cdots = \kappa_8 = 0, \kappa_9 = \cdots = \kappa_{15} = \Delta$

C5: $\kappa_1 = \cdots = \kappa_{12} = 0, \kappa_{13} = \Delta, \kappa_{14} = 2\Delta, \kappa_{15} = 3\Delta$

C6: $\kappa_1 = \cdots = \kappa_{10} = 0, \kappa_{11} = \Delta, \kappa_{12} = 2\Delta, \kappa_{13} = 3\Delta, \kappa_{14} = 4\Delta, \kappa_{15} = 5\Delta$

Similarly, for eight-run designs, we propose four configurations with up to three significant effects:

C1: $\kappa_1 = \cdots = \kappa_6 = 0, \kappa_7 = \Delta$

C2: $\kappa_1 = \cdots = \kappa_5 = 0, \kappa_6 = \kappa_7 = \Delta$

C3: $\kappa_1 = \cdots = \kappa_4 = 0, \kappa_5 = \kappa_6 = \kappa_7 = \Delta$

C4: $\kappa_1 = \cdots = \kappa_4 = 0, \kappa_5 = \Delta, \kappa_6 = 2\Delta, \kappa_7 = 3\Delta$

In all cases, the effects were generated as independent values of a Normal distribution with the specified mean and $\sigma_{ef} = 1$. As Ye, Hamada and Wu [22], we call the Δ parameter *Spacing* and its value varies from 0.5 to 8 in steps of 0.5. For each of the configuration-spacing combinations, 10,000 situations were simulated and for each one of them we have calculated the PSE. Figures 4 and 5 clearly confirm that, in eight as well as sixteen-run designs, the PSE overestimates the value of σ_{ef}.
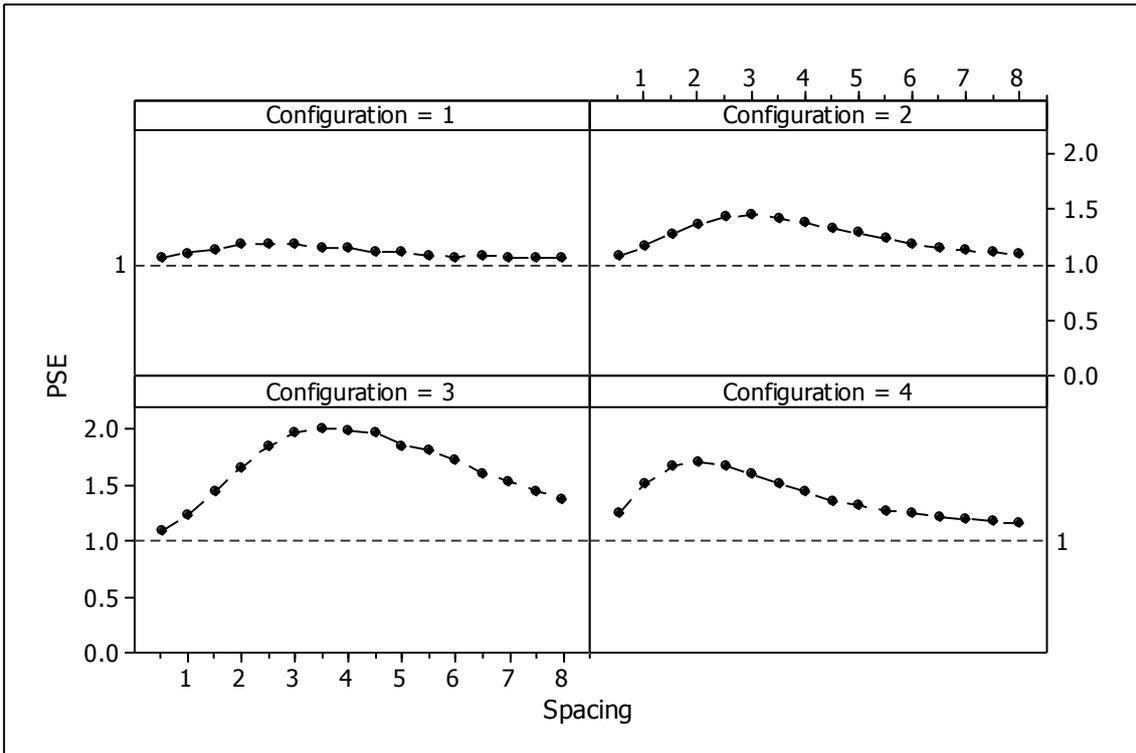
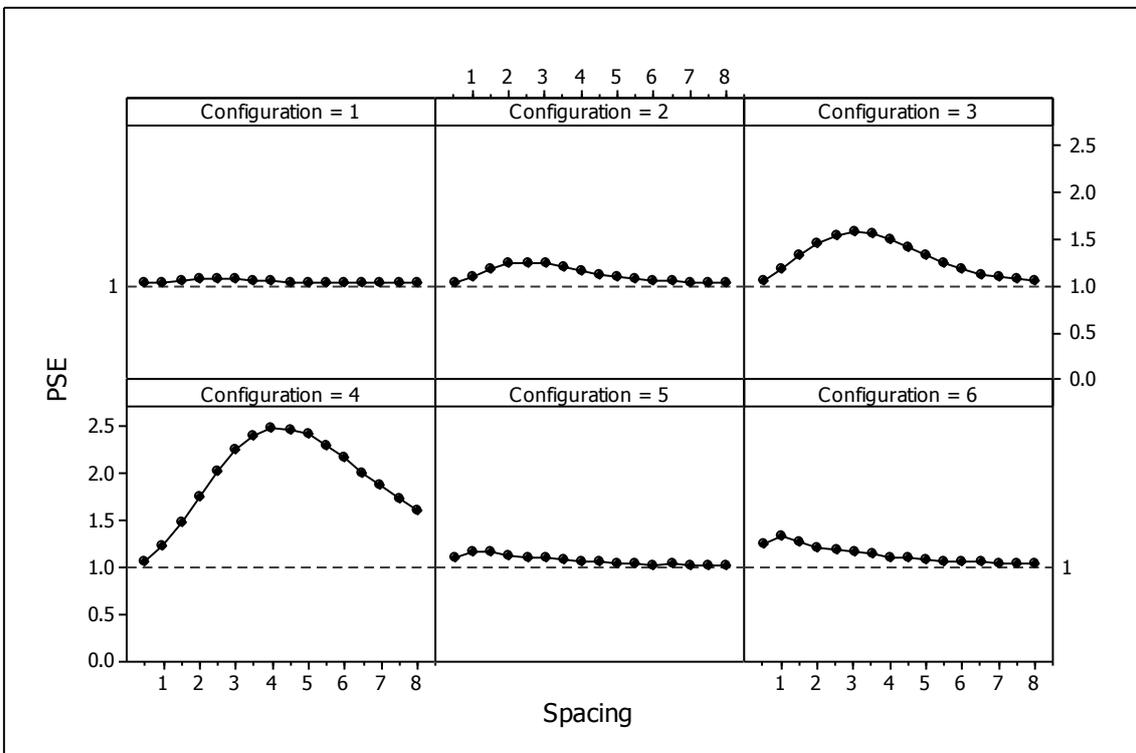Figure 4. Eight run design. Average of 10,000 PSE estimates made for each configuration-spacing combination.

Figure 5. Sixteen run design. Average of 10,000 PSE estimates made for each configuration-spacing combination.

4. New proposal

The fact that in practice PSE values tend to be higher than those obtained by simulation, suggest to use critical values smaller than the ones deduced by this procedure. At the same time it is not possible to propose a critical value that will always produce the desired significance level since it depends on the number and magnitude of significant effects, which is precisely what we want to know.

Therefore, using critical values that seem very accurate -with several decimal places- is misleading; it conveys the idea that the significance test has a meticulousness that it lacks. In accordance with Box's idea mentioned in abstract that aim of Design of Experiments should be to learn about the subject and not to test hypothesis, Box, Hunter and Hunter [3] do not perform significance tests for each effect, nor even when they have replicates. They place the effect values in a table together with their standard deviation in such a way that, in light of this information, the experimenter decides what factors should be considered to have an influence on the response. To go further than that when there are no replicates and, as such, less information is available, it doesn't make sense. Lenth himself doesn't even propose conducting formal significance tests, but rather an indicative graphical representation, as has been previously mentioned. Depending on which area of the graph the effects fall, they should be considered significant or not with higher or lower probability.

Paul Velleman, a student of John Tukey and a statistician at Cornell University, recounts that when he asked the reason for the 1.5 value when determining the area of anomalies in the boxplots, Tukey answered because 1 was too small and 2 was too large (see web reference Exploring Data [11]) and the number halfway between 1 and 2 is 1.5. With this same idea of simplicity, and in the absence of an exact value, our proposal is to always use a critical value of $t = 2$, independently of the number of runs conducted.

To show the usefulness and good properties of this value we compare the results produced by $t = 2$ with the ones with the t values proposed by Ye and Hamada [21]. The comparison is made by computing the percentages of type I and type II errors that occur in each of the scenarios presented in the previous section. The significant effects identification error rates were calculated as the number of errors committed with respect

to the total number that could have been committed. For example, for eight-run experiments in configuration 1 and with $\Delta = 0.5$ and considering the critical value proposed by Ye and Hamada ($t = 2.30$) 2759 type I errors have been obtained (effects which are considered significant when in reality they correspond to a distribution with $\kappa = 0$) and 9730 type II errors (effects with $\kappa \neq 0$ but which are considered not significant). As in configuration 1, there is only one effect with $\kappa \neq 0$. The error rates are: type I: $\frac{2759}{6 \times 10000} \times 100 = 4.60\,\%$ and type II: $\frac{9329}{1 \times 10000} \times 100 = 97.3\,\%$.

Figures 6 and 7 show the type I error rate for each configuration-spacing combination. Except in one case (16 runs experiment, configuration 1, one significant effect) the type I error rate is closer to the intended 5% with $t = 2$ than with the values proposed by Ye and Hamada [21].

Figures 8 and 9 show the type II error rate for each configuration-spacing combination. Although the differences seem small, in 8 run designs exceeds 10% in configurations 1, 2 and 3 and 8% in configuration 4. In designs with 16 runs the difference exceeds 6% in configurations 1, 2, 3 and 4 and 3% in 5 and 6.
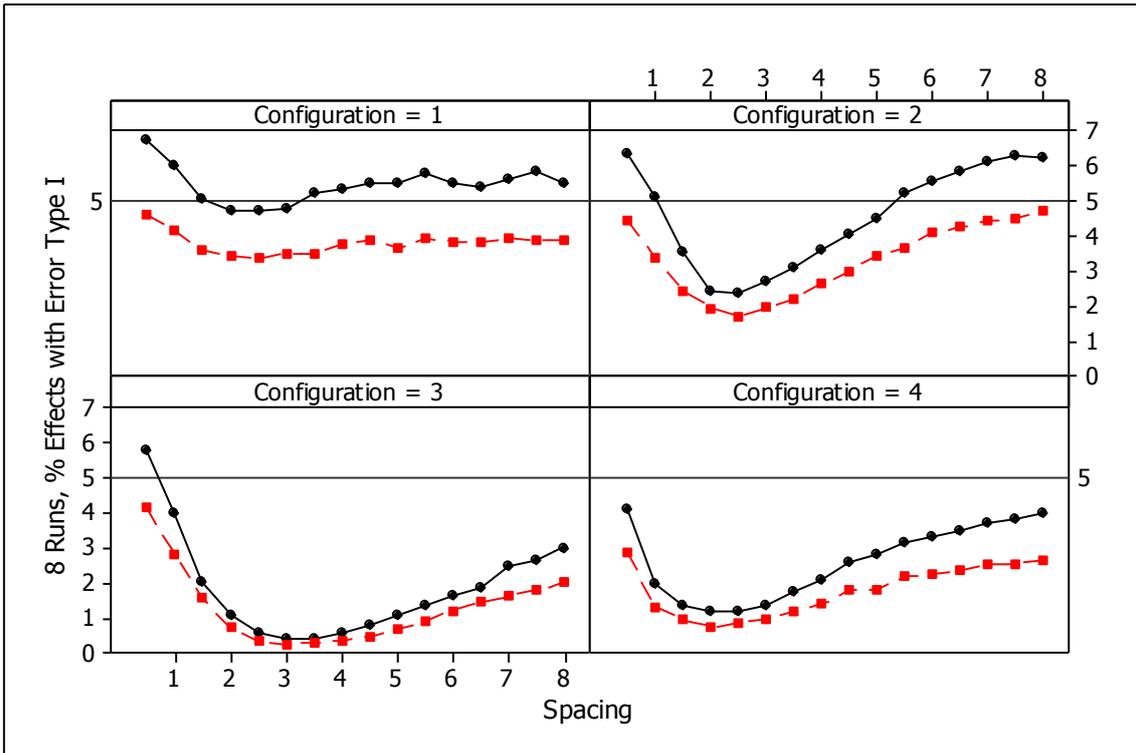
Figure 6. Eight-run designs. Percentage of Type I errors using $t = 2.3$ (square symbols) and $t = 2$ (round symbols).
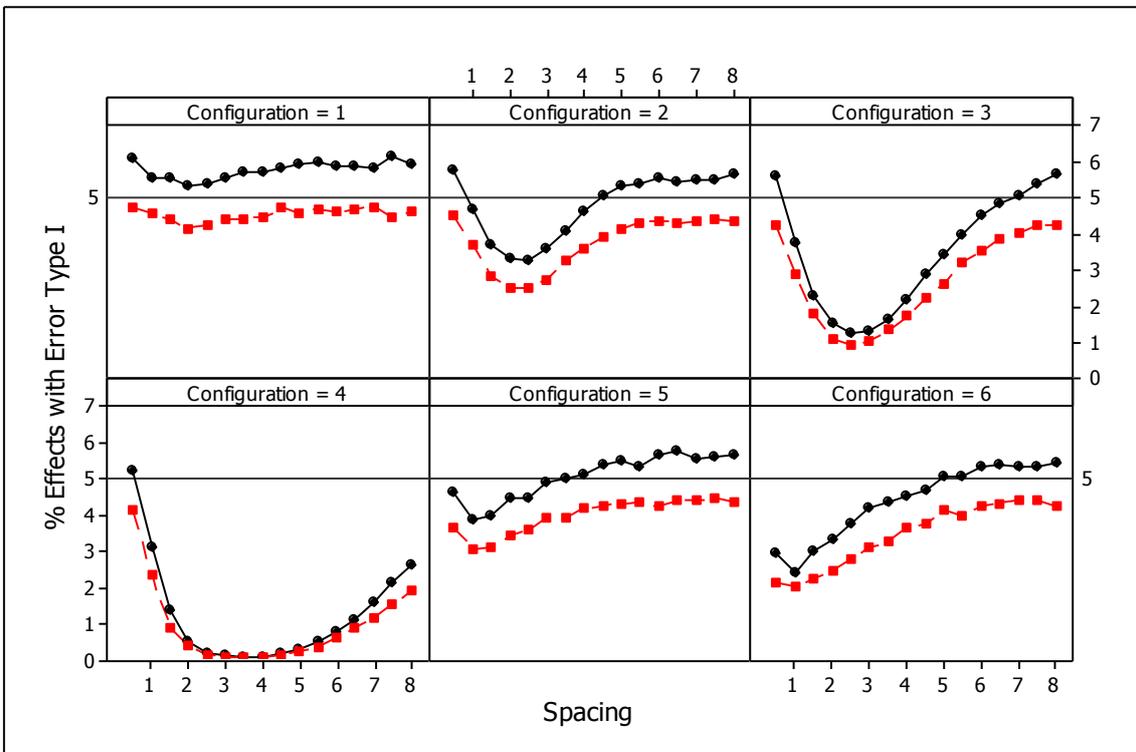

Figure 7. Sixteen-run designs. Percentage of Type I errors using $t = 2.156$ (square symbols) and $t = 2$ (round symbols).
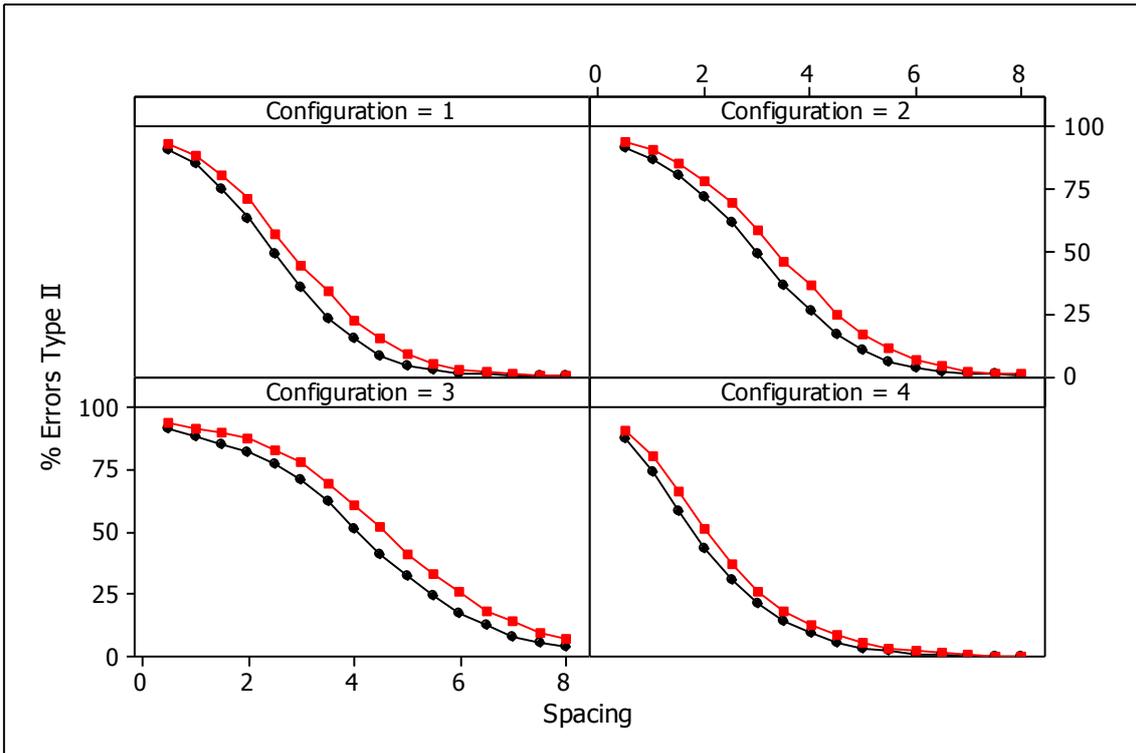
Figure 8. Eight-run designs. Percentage of type II errors using $t = 2.3$ (square symbols) and $t = 2$ (round symbols).
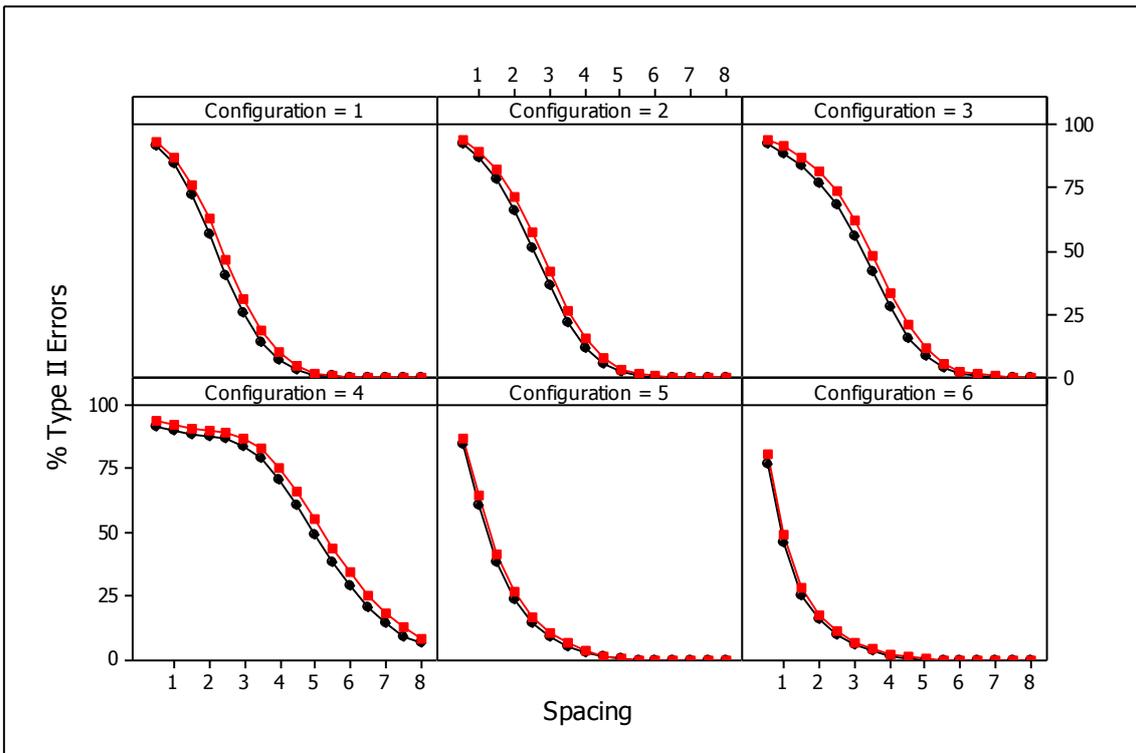

Figure 9. Sixteen-run designs. Percentage of type II errors using $t = 2.156$ (square symbols) and $t = 2$ (round symbols).

13

Even with $t = 2$, the type I error rate is below 5% in many configuration-spacing combinations; naturally this implies a higher type II error risk, which is something unwanted in industrial contexts (De León et al [7]). To mitigate this and in accordance with the already mentioned practice of Box, Hunter and Hunter [3] it seems adequate to define a doubtful zone for t-ratio values; a zone where it is not clear if the effects are significant or not but that the experimenter should be aware of this uncertainty. A reasonable choice for this zone could be the one corresponding to t-ratios between 1.5 and 2.

5. Conclusions

Among the many analytical procedures that have been proposed for identifying significant effects in not replicated two level factorial designs, the Lenth method constitutes a central reference. It is included in well-known books on experimental design (Montgomery [18]; Box, Hunter and Hunter [3]) and it is used by statistical software packages commonly employed in technical environments. And this is so, in spite that it is well known that the critical values employed lead to a type I error probability that is less than what is expected and, as an unwanted consequence, a greater probability of type II error. More accurate critical values derived by simulation produce the desired probability of a type I error when all factors included in the PSE calculation are inert. Unfortunately, the inclusion in the PSE calculation of one or more non-null factors, something difficult to avoid in practice, produces an overestimation of the σ_{ef} with the already mentioned undesired consequences.

The paper proposes a simple solution: to use a critical value of t=2 for any 2^k or 2^{k-p} design. The value is slightly lower than the best ones proposed by simulation, the difference decreases as the number of runs increases. This proposal has several advantages:

- It conveys the idea that the procedure is approximate, it cannot be otherwise, and that therefore it requires a dose of good judgment by the experimenter. Completely in line with G. Box [2] ideas of sequential experimentation and using DOE to learn.

- In a wide range of reasonable scenarios --number and size active effects-- it produces type I errors closer to the desired and announced 5% than the values proposed by simulation methods

- Finally, it is simple and easy and does not need simulations or tables.

References

1. Bergquist. B., Vanhatalo, E., Nordenvaad, M. L. (2011). A Bayesian analysis of unreplicated two-level factorials using effects sparsity, hierarchy, and heredity. *Quality Engineering* 23, 152-166.

2. Box, G. E. P. (1992). Teaching Engineers Experimental Design with a Paper Helicopter. Quality Engineering 4(3), 453-45

3. Box, G. E. P., Hunter, J. S., Hunter, W. G. (2005). *Statistics for experimenters: Design, innovation, and discovery*, 2nd edition. NJ: Wiley-Interscience, Hoboken.

4. Costa, N., Lopes Pereira, Z. (2007). Decision-making in the analysis of unreplicated factorial designs. *Quality Engineering* 19, 215-225.

5. Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments, *Technometrics* 1(4), 311-341.

6. Daniel, C. (1976). Applications of statistics to industrial experimentation, NY: Wiley.

7. De León G., Grima, P., Tort-Martorell, X. (2006). Selecting significant effects in factorial designs taking type II errors into account. *Quality and Reliability Engineering International* 22(7), 803-810.

8. De León, G., Grima, P., Tort-Martorell, X. (2011).Comparison of normal probability plots and dot plots in judging the significance of effects in two level factorial designs. *Journal of Applied Statistics* 38(1), 161-174.

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

10. Edwards, D. J., Mee, R.W. (2008). Empirically determined p-values for Lenth t-statistics. *Journal of Quality Technology* 40(4), 368-380.

11. Exploring data [website on the Internet]. http://exploringdata.net/why_1_5.htm. [15 January 2013].

12. Haaland, P. D., O'Connell, M. A. (1995). Inference for effect-saturated fractional factorials. *Technometrics* 37(1), 82-93.

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

14. Juan, J., Peña, D. (1992). A simple method to identify significant effects in unreplicated two-level factorial designs. Communications in Statistics. *Theory and Methods* 21(5), 1383-1403.

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

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

17. Mee, R. W., Ford III, J. J., Wu, S.S. (2011). Step-up test with a single cutoff for unreplicated orthogonal designs. *Journal of Statistical Planning and Inference* 141(1), 325-334.

18. Montgomery, D. C. (2008). *Design and analysis of experiments* (7th edn). NJ: Wiley, Hoboken.

19. Ott, E. R. (1975). Process quality control: Troubleshooting and interpretation of data. NY: McGraw-Hill.

20. Venter, J. H., Steel, S. J. (1998). Identifying active contrasts by stepwise testing. *Technometrics* 40(4):304-313.

21. Ye, K. Q., Hamada, M. (2000). Critical values of the Lenth method for unreplicated factorial designs. *Journal of Quality Technology* 32(1), 57-66.

22. Ye, K. Q., Hamada, M., Wu, C. F. J. (2001). A step-down Lenth method for analyzing unreplicated factorial designs. *Journal of Quality Technology* 33(2), 140-152.