This is the Submited Manuscript of the article published in Quality and Reliability Engineering International on July 2020. Final Published Version: https://doi.org/10.1002/qre.2702

Selecting relevant effects in factorial designs

Pere Grima, Lourdes Rodero, Xavier Tort-Martorell
Department of Statistics and Operational Research
Universitat Politècnica de Catalunya - BarcelonaTech, Spain

Industrial contexts tend to be as much or more concerned about the probability of ignoring an effect when its influence on the response is relevant (type II error) than about the probability of considering an effect to be active when in fact it is not (type I error). Here, we present a methodology for taking into account both types of error by fixing an effect value that is considered large enough to control the probability of it going unnoticed. In addition, we propose a plot to visualize the results obtained.

Key words: Factorial designs, significant effects, Type I and Type II errors

1. Introduction

When carrying out an experimental plan in an industrial environment, if the person in charge of the process is asked which of the following errors they would rather commit – 1) considering that a factor influences the response when in fact it does not, or 2) ignoring the influence of a factor that actually has an effect – they will generally choose the first. While it is indeed true that arriving at the erroneous conclusion that factor A affects the response – when in fact it does not – can lead to controlling A unnecessarily and even to make an unnecessary investment, the quality of the final product will nevertheless not be affected. It is even likely that with the passage of time the experience reveals the error. On the other hand, failure to detect that B affects a response can cause multiple problems. For example, it may lead to an increase in the variability of the response due to a lack of control over B; or additional costs may be incurred when trying to optimize the process while ignoring the influence of B. What is worse, it will be very difficult to detect that B has any relevant effect without conducting a new study.

However, when analyzing which effects should be considered active, attention is usually given only to the probability of committing the first type of error, which is generally referred to as type I error and is usually set at 5%. Meanwhile the probability of the second – the so-called type II error – is ignored and often turns out to be much greater than what could be considered reasonable.

Undoubtedly, this way of proceeding originates in the usual practices of science, where experimental design began and developed as a methodology. For a scientist, saying that A affects the response when in fact it does not is a clear error, as it makes a false contribution to scientific knowledge. However, overlooking that factor B indeed does have an influence is not considered a serious error and will probably be detected in future investigations. In industrial environments, we continue reasoning in the same way – even though, as we have mentioned, the priorities of industry and science are different.

De León *et al*.[1] make clear the importance of considering the type II error in the analysis of the significance of the effects by showing that this probability of error is usually greater than the experimenter suspects. In this article we propose a simple methodology to take into account this type II error and to graphically show the results of the analysis.

The following section presents two examples taken from well-known experimental design books, in which the usual methods for analyzing the significance of the effects fail to consider factors whose influence is borderline. Then, we present a methodology to take into account the probability that an effect with relevant influence on the response goes unnoticed. Next, the methodology is illustrated by applying it to the examples described at the beginning and finally we explain how to represent the obtained results in a clear way so that the most appropriate decisions can be taken.

2. Situations with factors of borderline influence. Examples

Based on data published by Prat and Tort-Martorell[2], Box *et al*.[3] present a 2^3 design carried out in a pet food factory (p. 194). Several responses are analyzed, but we consider only the amount of product obtained (yield). The factors considered are:

Factors	Levels	
	−	+
A: Conditioning temperature	80% of max.	Max
B: Flow	80% of max.	Max
C: Compression zone	2	2.5

Table 1 shows the design matrix together with the results obtained (left) and the values of the effects (right).

Table 1: Design matrix and effects obtained in the example of Box et al. (2005)

Factors			y
A	B	C	
-1	-1	-1	83
1	-1	-1	85
-1	1	-1	99
1	1	-1	102
-1	-1	1	59
1	-1	1	75
-1	1	1	80
1	1	1	73

Effects	
A:	3.5
$\boldsymbol{B}$:	**13.0**
$\boldsymbol{C}$:	**-20.5**
AB:	-5.5
AC:	1.0
BC:	-3.5
ABC:	-6.0

The most usual way to estimate the standard deviation of the effects is through Lenth's Pseudo Standard Error (PSE)[4]. It is based on the fact that if $X \sim \mathrm{N}(0, \sigma^2)$, the median of $|X|$ equals 0.6745σ and therefore $1.5 \cdot \mathrm{median}|X| = 1.01\sigma \cong \sigma$. Assuming that e_i $(i = 1, \ldots, n)$ are the values of the effects of interest and that their estimators $\hat{e}_i$ are distributed according to a $\mathrm{N}(e_i, \sigma_e^2)$, Lenth defines $s_0 = 1.5 \cdot \mathrm{median}|\hat{e}_i|$ and calculates a new median by excluding the estimated effects with $|\hat{e}_i| > 2.5 s_0$. In doing so, he expects to exclude the effects with $e > 0$ and use the others to calculate:

$$PSE = 1.5 \cdot \underset{|\hat{e}_i|<2.5 s_0}{\mathrm{median}} |\hat{e}_i|$$

An effect is considered significant if its estimator satisfies $|\hat{e}| > t_{1-\alpha/2,,\nu} \cdot PSE$, with α being the level of significance and ν the degrees of freedom in the t distribution. In his original article, Lenth proposes using $\nu = n/3$, with n being the number of effects considered. Authors such as Loughin[5], Ye and Hamada[6], and Fontdecaba *et al*.[7] have shown that the t-values proposed by Lenth lead to probabilities of type I error (α) that are considerably lower than the intended value, with the unwanted result of a higher probability of type II error. To achieve a significance level of $\alpha = 0.05$, we will use the values 2.297 and 2.156 proposed by Ye and Hamada for designs with 8 and 16 runs, respectively, and from now on we will use k to designate the PSE multipliers to avoid the impression that they always come from a t-Student distribution.

In this case, $PSE = 8.25$ and effects satisfying $|e| > 18.95$ should be considered significant. Therefore, only the main effect of factor C $(= 20.5)$ appears to be significant, although taking a look at Figure 1 generates doubt that we should rule out the possible influence of factor B.

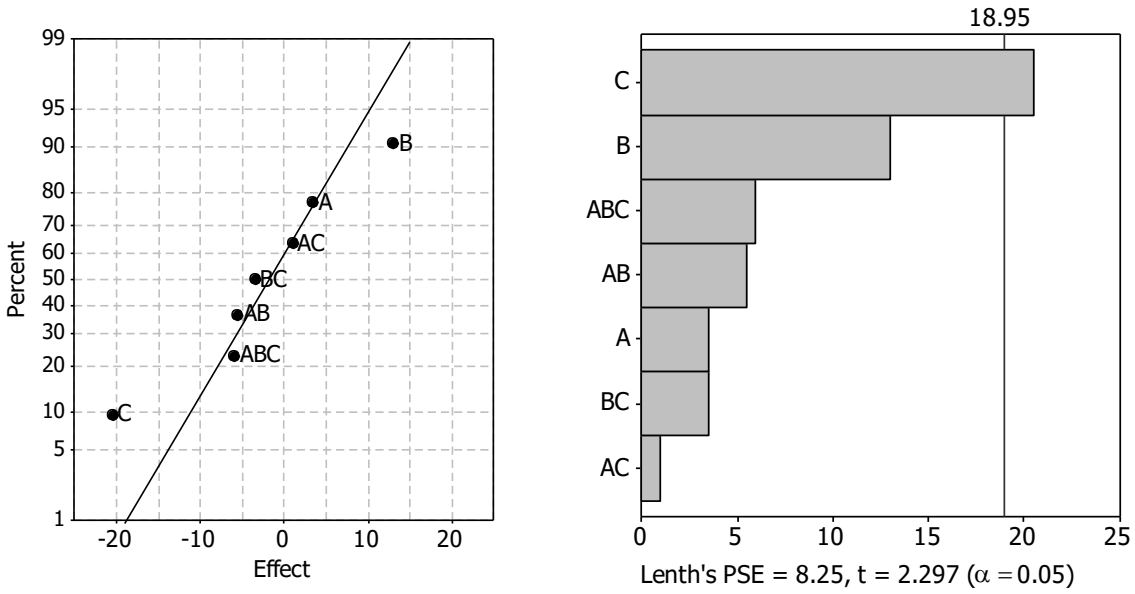

Figure 1: *NPP (left) and Pareto chart (right) representations of the effects in the example by Box et al. (2005)*

Regarding designs with 16 runs, Wu and Hamada[8] present a 2^4 design (p. 155) intended to analyze which variables affect the thickness of the epitaxial layer on polished silicon wafers. The target value is 14.5 ± 0.5 µm. This example studies the influence of factors on the mean and on the variability of the response, although our focus here is only on the influence on the mean.

	Level	
	−	+
A: Deposition time	low	high
B: Deposition temperature (°C)	1210	1220
C: Nozzle position	2	6
D: Susceptor-rotation method	continuous	oscillating

On the left side of Table 2 is the design matrix together with the responses obtained. On the right we have the values of the effects.

Table 2: Design matrix and effects obtained in the example of Wu and Hamada (2009)

Factors				y
A	B	C	D	
-1	-1	-1	-1	14.59
1	-1	-1	-1	13.59
-1	1	-1	-1	14.24
1	1	-1	-1	14.05
-1	-1	1	-1	14.65
1	-1	1	-1	13.94
-1	1	1	-1	14.40
1	1	1	-1	14.14
-1	-1	-1	1	14.67
1	-1	-1	1	13.72
-1	1	-1	1	13.84
1	1	-1	1	13.90
-1	-1	1	1	14.56
1	-1	1	1	13.88
-1	1	1	1	14.30
1	1	1	1	14.11

Effects	
$\boldsymbol{A}$:	**-0.4900**
B:	-0.0775
C:	0.1725
D:	-0.0775
$\boldsymbol{AB}$:	**0.3450**
AC:	0.0300
AD:	0.0500
BC:	0.0575
BD:	-0.0925
CD:	0.0075
ABC:	-0.1100
ABD:	0.0300
ACD:	-0.0250
BCD:	0.0975
$ABCD$:	-0.0200

In this case, the value of PSE is 0.08625. So we should consider significant effects to be those for which $|e| > 0.1860$, that is, A and AB (Figure 2). The main effect of factor C appears to be non-significant but can be considered a borderline case. Indeed, if in a first approximation we calculate a confidence interval of 95% for the true value of C (-0.013, 0.358), we will observe that – even though it includes zero – it reaches 0.35, so it is perfectly reasonable to consider that it can take values like 0.20 or 0.25. Given that, as said, the target value is 14.5 ± 0.5 µm, it does not seem prudent to ignore effects that represent 20 or 25% of the tolerance interval.

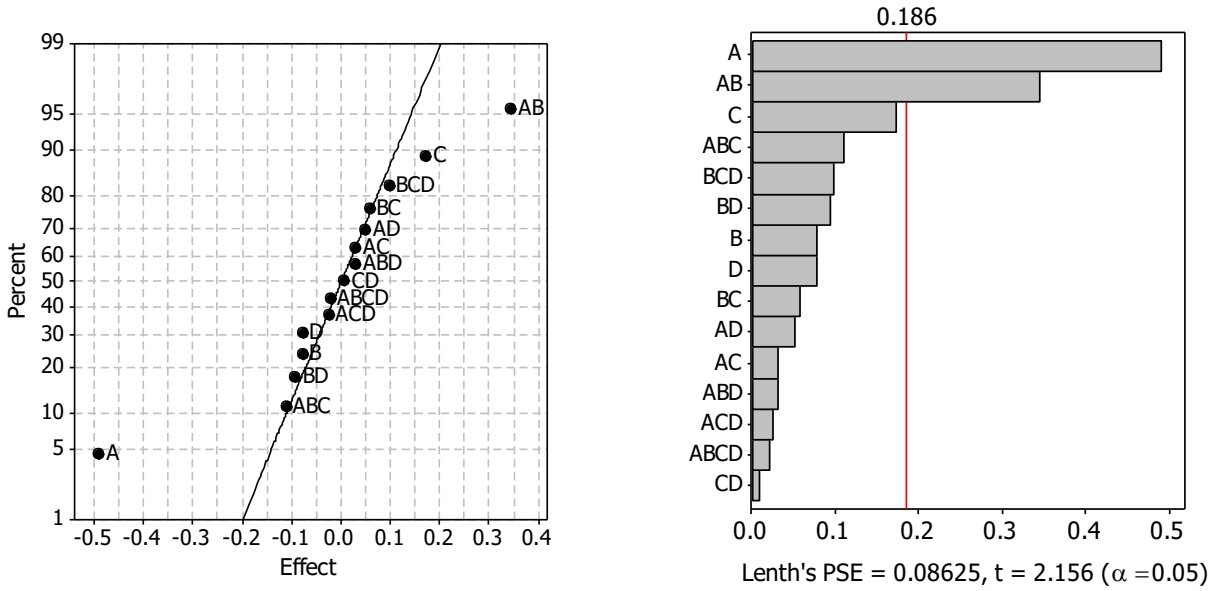

Figure 2: NPP (left) and Pareto chart (right) representations of the effects of the Wu and Hamada example[8]

3. Analyzing the effects while taking type II error into account

Taking into account the type II error is especially important in two situations: when a first analysis shows effects whose significance is in doubt because they are close to the critical value, and when there are effects that, in spite of being non-significant, have a value that is considered too large to be ignored. In these cases, we propose the following procedure:

1. Set a value such that if the effect exceeds it, there will be a low probability of ignoring it.

   We believe that, in industrial contexts, identifying this value is a relatively simple exercise, which is always worth doing.

   Following De León *et al*.[1], we will call this value $MESI$ (Minimum Effect Size of Interest). The probability that it goes unnoticed can be decided in each case; but in the same way that α is generally set at 0.05, we propose setting the value of this probability at $\beta = 0.10$, the most usual one.

2. Estimate the standard deviation of the effects.

   Once the effects that can be considered significant have been identified, their standard deviation can be estimated in different ways:

   a) Using the same value of the PSE already used to analyze the significance of the effects. The main advantage of using this estimator is that it is not necessary to calculate anything new. In addition, it seems more consistent to use the same estimator for σ_e when considering both type I and type II errors. The problem with this approach is that information about which effects are significant are not taken into account.

   b) Recalculating the PSE only based on effects that are not significant. This method takes advantage of all the information available, and is coherent in the sense that estimates σ_e by the same procedure for both types of errors. Unfortunately, this option has the disadvantage that the value of k --that multiplied by the PSE will provide intervals of a given confidence-- is unknown and does not have an analytical expression. In what follows, by means of a simple simulation, we show this fact for a 8 run design.

      We have generated 7 random numbers (for the 7 effects) from a $N(0,1)$ distribution. Then, we calculate the PSE and identify those that appear to be significant, that is, they fall outside the range $0 \pm k \cdot PSE$. This operation is repeated 10,000 times in increments of 0.01 for each value of k between 0 and 3.

Figure 3 represents the proportion of significant effects depending on k. For example, for $k = 2.30$, 10000 samples of 7 observations each have been generated. In these samples, a total of 349 effects would be considered significant, which represents 0.049857 of the total (349/7000), this being the value represented in the graph. The values obtained in our simulation coincide with those reported in the article by Ye and Hamada[6].

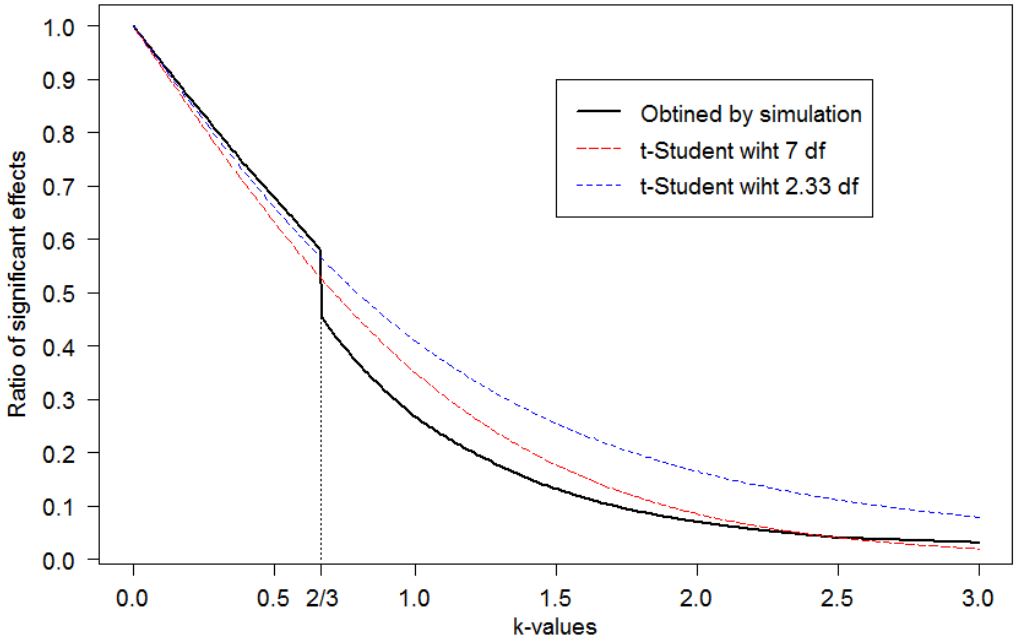

Figure 3: *Ratio of non-significant effects depending on the k-value used*

The sudden drop in the ratio of significant effects when $k = 2/3$ is notable, and it results from the method of calculating the PSE. We have seen that $s_0 = 1.5 \cdot \text{Me}|\hat{e}_i|$ is calculated first, and then the median is recalculated but excluding the values at which $|e_i| > 2.5s_0$. If we do not exclude any value, which happens frequently, we have $PSE = 1.5 \cdot \text{Me}|\hat{e}_i|$. If $k = 2/3$, then significant effects will be those that have an absolute value greater than $2/3 \cdot PSE$, that is greater than $\text{Me}|\hat{e}_i|$. Therefore, the median effect lies just outside the border of the interval and is not considered significant for values of $k \geq 2/3$ while it lies just inside for values of $k < 2/3$. In designs with 16 runs, the problem is analogous. As the number of runs increases, the magnitude of the jump decreases because the effect that is located on the median represents a smaller percentage of the total effects.

Figure 3 also includes the curve obtained using a t-Student value with $\nu = 2.33$ df (= 7/3), as proposed by Lenth[3], and also with $\nu = 7$, which fits better in the tails – although not satisfactorily.

As seen, there is not an analytical expression that allows to easily determine the value of k that should multiply the PSE. This value can be obtained by simulation. In Wu and Hamada[8], one will find values for $m \geq 7$. For our purposes, this would imply generating by simulation a table of k-values for different number of effects. To the best of our knowledge, this has not been done and we believe that a better option is to avoid this complication using method c) explained below.

c) Estimating the variance of the effects from the values of those that are considered non-significant after applying the Lenth method. This method makes vary easy estimating σ_e and has the important additional advantage of allowing a simple analytical way to obtain k because, as will be seen bellow in the 3rd step of the procedure, it follows a non-central t-Student

Estimating σ_e from non-significant effects is a common procedure that, in addition, Xampeny *et al*.[9] show by simulation that –in general– when there are three or more effects that can be considered null the estimate obtained is better than the one provided by Lenth method (no matter which k is used). Notice that according to the Effect Sparsity Principle in 16 runs designs there will very frequently be three or more null effects and this will also quite often be the case in 8 run experiments.

3. Determine the Critical Value for Relevance (CVR). Represent it graphically

In what follows we call, as usual, critical value (CV) the one that separates the effects that are considered significant from those that are not significant, and that is obtained by fixing a maximum value to the probability of committing a type I error. We propose to use, in combination with this value, a new one based on regulating through the $MESI$ the risk of committing a type II error, we call it the Critical Value for Relevance (CRV).

The CVR is calculated as follows: we have $\hat{e}/s_e \sim t_{\nu,,d}$, with $d = e/s_e$ being an estimate of the the noncentrality parameter and ν the degrees of freedom of the t-Student distribution (notice that there will be as many degrees of freedom as values used, since in this case the mean is known). So $\hat{e} \sim t_{\nu,,d} \cdot s_e$ and if $e = MESI$, we have $d = MESI/s_e$ and the CVR will be equal to $t_{\nu,,d,(\beta)} \cdot s_e$, being $t_{\nu,,d,,(\beta)}$ the β-quantile of $t_{\nu,,d}$.

4. Analysis of the examples presented

In the 2^3 design example[3], excluding the significant effect $C = -20.5$ allows us to estimate the standard deviation of the effects using the other 6 that we assume to have an average $\mu = 0$, thus $s_e = 6.58$ with

6 degrees of freedom. Let us suppose that a 20 unit change in yield is a relevant change ($MESI = 20$) and, as we have indicated, we set $\beta = 0.1$ as the probability that a change of this magnitude will go unnoticed, that is, it will be considered non-significant. Then we have $d = \frac{e}{s_e} = 3.04$ and $t_{\nu=6,,d=3.04,,(0.1)} = 1.69$ (Figure 4). The critical value at which the effects should be considered relevant is: $\beta\text{-}CV = t_{0.10,,\nu,,d} \cdot s_e = 11.12$. Therefore, the possible influence of factor B (= 13.0) should also be taken into account.

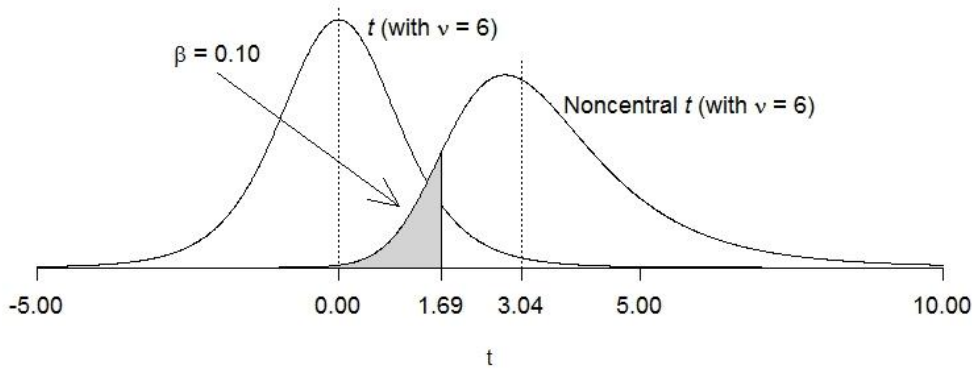

Figure 4: *Noncentral t distribution for determining the critical value with* $MESI = 20$ and $s_e = 6.58$

As for the design with 16 runs[8], excluding the effects that appear significant in the first analysis allows us to obtain $s_e = 0.0787$ with $\nu = 13$ df. If we consider that an influence of 0.25 µm on the thickness (equivalent to 25% of the tolerance interval) should not go unnoticed and, further, set the probability of this happening at $\beta = 0.1$ then we have $d = 3.1766$ and $t_{\nu=13,,d=3.1766,.(0.1)} = 1.85$ (Figure 5). Therefore, the critical value turns out to be $t_{\nu,,d,,(0.1)} \cdot s_e$ = 0.1455. Then, we should also consider as relevant the main effect of factor C (= 0.1725).

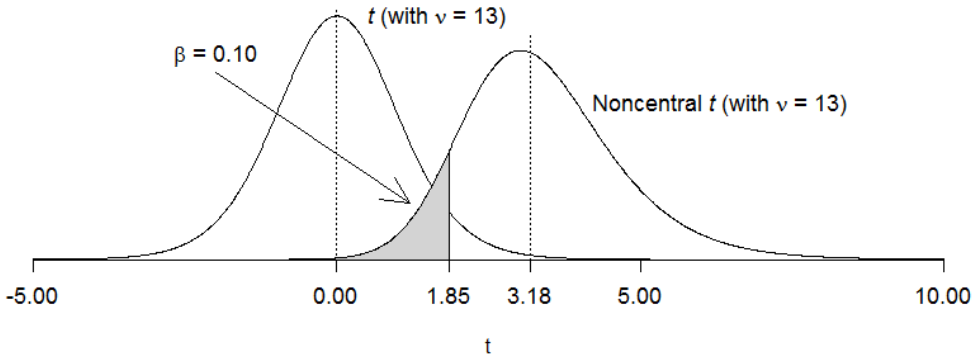

Figure 5: Noncentral t distribution for determining the critical value with $MESI = 0.25$ and $s_e = 0.0787$

Another approach could be to calculate the magnitude of an effect that has a 10% probability of going unnoticed, doing so by using the critical value obtained with $\alpha = 0.05$. In the first example, the critical

value that follows from using $\alpha = 0.05$ is α-CV = 18.95, and this value leads to $MESI$ = 29.12. In the second example, we had α-CV = 0.186, which corresponds to $MESI$ = 0.29. No experimenter would feel comfortable with either of these values. Table 3 (left) shows the probabilities (β) of the value in the $MESI$ column going unnoticed when it is decided based only on a type I error probability of $\alpha = 0.05$, as is usually done. On the right side, we have the probability of considering an inert effect to be significant when we want a probability of $\beta = 0.10$ for the $MESI$ values considered.

Table 3: *Critical values when fixing* $\alpha = 0.05$ or $\beta = 0.10$ *for the example of Wu and Hamada*[7]

$\alpha = 0.05$		
$MESI$	Critical value	Prob. β
0.15	0.17	0.58
0.20	0.17	0.35
0.25	0.17	0.16
0.30	0.17	0.06
0.35	0.17	0.02

$\beta = 0.10$		
$MESI$	Critical value	Prob. α
0.15	0.05	0.54
0.20	0.10	0.23
0.25	0.15	0,08
0.30	0.19	0.03
0.35	0.24	0.01

5. Presentation of the results

The critical values can be calculated easily with the help of a statistical software package that allows using the noncentral t-Student distribution. Here, we have used R. These values can be represented in a clear and visual way by including two vertical lines (one for each critical value) in the Pareto chart of the effects, rather than the usual practice of having only one line. The meaning of each line is clarified by including the conditions for which it was calculated.

Figure 6 and Figure 7 represent the Pareto charts with the lines indicated for the examples that we have been analyzing. Effects beyond the upper critical value must be considered significant, and those that do not reach the lower value do not seem necessary to consider. Those found between the two could be considered borderline, and in each case the experimenter should decide on the most appropriate option.

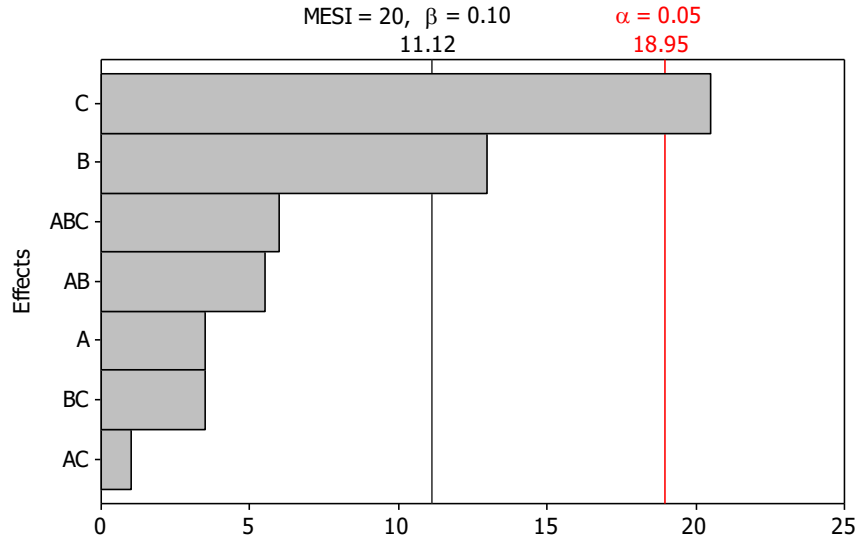

Figure 6: Example of Box et al.[2]. Pareto chart with the lines corresponding to the critical values for each type of error

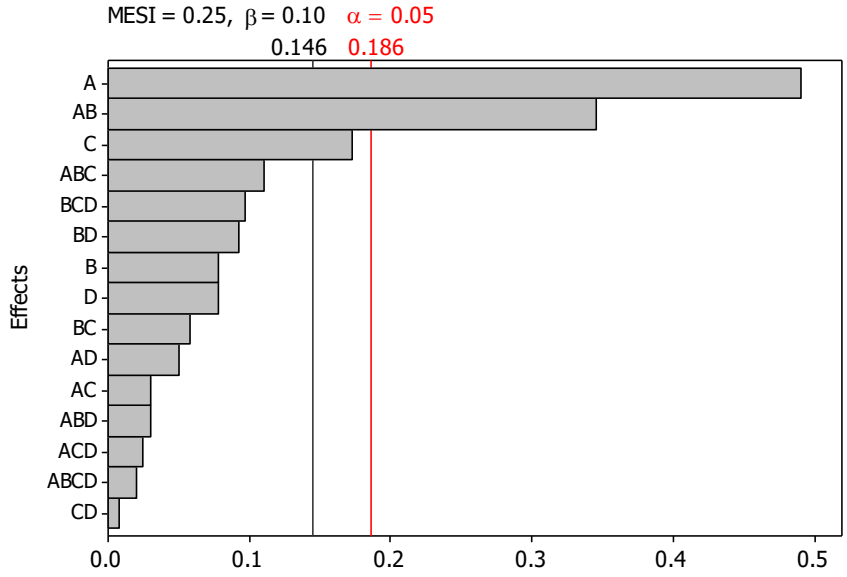

Figure 7: Example of Wu and Hamada[7]. Pareto chart with the lines corresponding to the critical values for each type of error

It can also happen that the effects to be considered are the same regardless of which critical value is considered. For example, in Box et al.[3] (p. 199), an example is presented where the significant effects stand out very clearly over those that are not. It is a 2^4 design, and the results (in the standard order of the design matrix) are:

70, 60, 89, 81, 69, 62, 88, 81, 60, 49, 88, 82, 60, 52, 86, 79

The effects are graphically represented in Figure 8. In this case, it is very clear that the effects to be considered must be A, D, B and BD. The estimator of σ_e calculated from the non-significant effects is $s_e = 0.643$ with 11 degrees of freedom. Then, considering $MESI = 3$ we have $\delta = 4.67$ and $t_{\nu=11,,\delta=4.67,,(0.1)} = 3.188$, for which the critical value is $t_{\nu,,\delta,,(0.1)} \cdot s_e = 2.05$. Unlike the previous cases, the critical value calculated from the probability of committing a type II error is greater than that which corresponds to type I, which is equal to 1.62. In any case, as it could not be otherwise, if the results are very clear it makes no difference whether we use one or the other criterion.

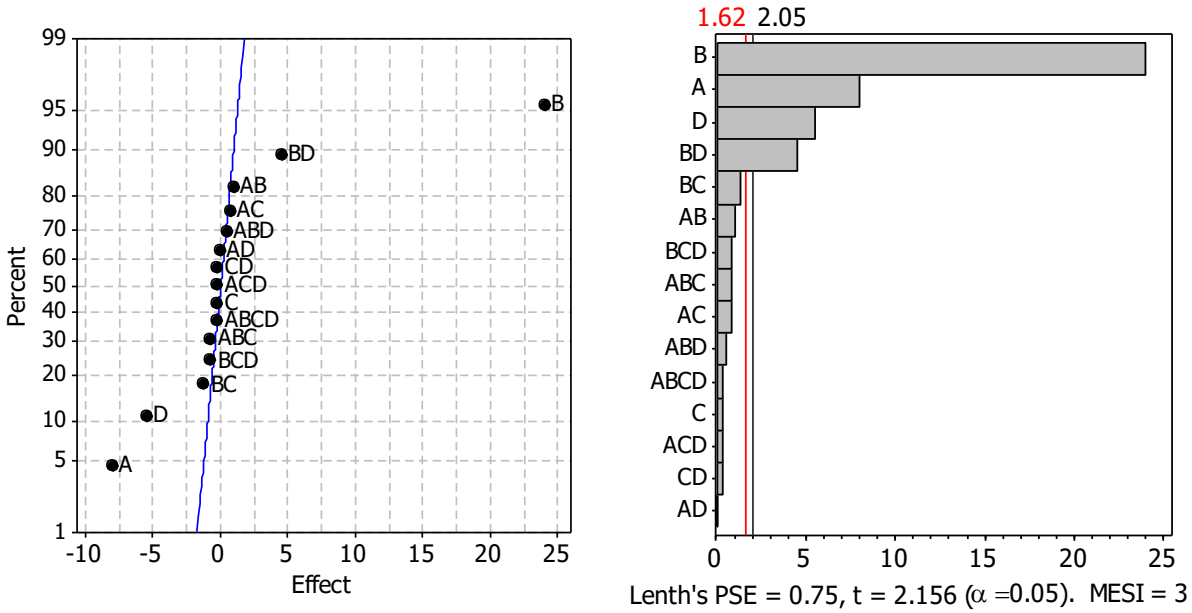

Figure 8: *NPP (left) and Pareto chart (right) representations of the effects in the example of Box et al.[2], p. 199*

Conclusions

In industrial contexts, ignoring the existence of effects with a relevant influence on the response is an error that can have important consequences. Nevertheless, the decision criterion that is commonly used ignores the probability of committing this type of error and, instead, focuses only on the avoidance of considering relevant effects that do not actually influence the response.

Setting a Minimum Effect Size of Interest ($MESI$) makes it possible to determine – in a simple way – a new critical value that also takes into account the probability of committing this type of error. By adding a new vertical line in the Pareto chart of the effects, together with the usual line based on a significance level of $\alpha = 0.05$, we are able to clearly visualize the situation.

In addition to having the "didactic" value of not considering a clear boundary between relevant effects and those that are not, establishing a zone of borderline effects that must be assessed as to whether or not they should be considered relevant.

6. References

1. De León G, Grima P, Tort-Martorell X. Selecting Significant Effects in Factorial Designs Taking Type II Errors into Account. *Quality and Reliability Engineering International* 2006; 22(7):803-810. DOI: 10.1002/qre.729.

2. Prat A, Tort-Martorell X. Case Study: Experimental Design in a Pet Food Manufacturing Company. *Quality Engineering*, 1990; 3(1):59-73. DOI: 10.1080/08982119008918838.

3. Box GEP, Hunter JS, Hunter WG. Statistics for experimenters: design, innovation, and discovery. Haboken: Wiley; 2005.

4. Lenth RV. Quick and easy analysis of unreplicated factorials. *Technometrics*, 1989; 31: 469-473. DOI:10.2307/1269997.

5. Loughin TM. Calibration of the length test for unreplicated factorial designs. *Journal of Quality Technology*, 1998; 30(2): 171-175. DOI: 10.1080/00224065.1998.11979836.

6. Ye KQ, Hamada M. Critical values of the Lenth method for unreplicated factorial designs. *Journal of Quality Technology*, 2000; 32:57-66. DOI: 10.1080/00224065.2000.11979971 .

7. Fontdecaba S, Grima P, Tort-Martorell X. Proposal of a single critical value for the Lenth method. *Quality Technology and Quantitative Management*, 2015; 12:41-51. DOI: 10.1080/16843703.2015.11673365.

8. Wu CFJ and Hamada M. Experiments. Planning, Analysis, and Parameter Design Optimization. New York: Wiley; 2009.

9. Xampeny R, Grima P, Tort-Martorell X. Selecting significant effects in factorial designs: Lenth's method versus using negligible interactions. *Communications in Statistics - Simulation and Computation* 2018, 47(5):1343-1352, DOI: 10.1080/03610918.2017.1311917