Selecting significant effects in factorial designs: Lenth's method versus using negligible interactions

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

ABSTRACT

Among the many analytical techniques that have been published to analyze the significance of the effects in the absence of replications, two have emerged as the most widely used in text books as well as statistical software packages: The Lenth's method and the estimation of the variance of the effects from the values of those considered negligible. This article shows that neither is better than the other in all cases, and by analyzing the results obtained in a wide variety of situations it provides guidelines on when it is preferable to use one or the other technique.

KEY WORDS: Lenth's method, Effects significance, Factorial design, Statistical software, Negligible interactions

1. Introduction

When a 2^k experimental design is carried out, the influence of the factors ($X_1, X_2, \dots, X_k$) over the response (Y) is quantified through the so-called effects. Once the effects are calculated it is necessary to analyze which are significantly different from zero. If there are replicates it is possible to estimate the variance of the response and from it the variance of the effects (σ_{ef}^2) that can be used to assess significance.

In absence of replicates, there is no estimate for the variance, and thus significance has to be assessed by other methods. Several graphical and analytical method have been developed to solve the problem. Among the graphical are: the Pareto chart of the effects based on the sparsity principle and the idea that significant effects will present values that will stand out from those that are not; and representing the effects in a Normal Probability Plot (NPP) or Half Normal Plot based on the idea that non-active effects follow a $N(0, \sigma_{ef}^2)$, and thus will be aligned on a straight line. Graphical methods

have the inconvenient that they require human judgement and cannot be easily automated.

A lot of analytical methods have been proposed, Hamada and Balakrishnan (1998) give a very complete and deep study of many of them. Here we present a very short review.

Lenth (1989) proposed a simple but effective procedure for estimating the standard deviation of the effects that will be explained in section 2. Other authors tried to improve this method adapting it to specific situations, for example, Dong (1993) proposed a procedure useful when the number of significant effects is low (less than 20%) and Juan and Peña (1992) when the number of significant effects is large (greater than 20%). The so-called "step by step" strategy proposed by Venter and Steel (1998), also tries to improve Lenth's method for the case of many significant effects. A more general approach is proposed by Ye *et al.* (2001) with their step-down version of the Lenth's method. Box and Meyer (1986, 1993) proposed a Bayesian approach. Recently, Espinosa *et al*. (2016), proposed a Bayesian sequential method based on posterior predictive checks to screen for active effects.

In spite of this wide variety of analytical methods, Lenth's original method and to estimate the variance of the effects by using interactions that can be considered negligible from the outset are the most widespread. They are explained in commonly used textbooks as Montgomery (2013) and Box *et al.* (2005), and implemented in the most common statistical software packages for industrial applications (Fontdecaba *et al.,* 2014). Accepting to reduce the problem to choose among these two methods, it seems reasonable to suppose that it will be more adequate to apply one or the other according to the characteristics of the situation; for instance, depending on the number of interactions that can be considered negligible. However, several statistical software packages use always by default the same method regardless of the characteristics of the case being analyzed (Fontdecaba *et al.*, 2014).

The aim of the paper is to characterize in which situations it is better to use each method, based on the results of a simulation under some general scenarios representing a wide variety of situations that may occur in real life experimentation. The number of Type I and Type II errors made by each method is used to compare them.

The paper is organized as follows: in the next section we describe the two methods to be compared: Lenth's method and variance estimation using negligible interactions. In Section 3 we describe the simulation scenarios. In Section 4 we show and analyze the obtained results. Finally we provide some conclusions and final remarks in Section 5.

2. Description of the compared methods

Lenth's method consists of estimating the standard deviation of the effects based on the fact that if $X \sim N(0, \sigma^2)$, the median of $|X|$ equals 0.645σ and therefore $1.5 \cdot \text{median}|X| = 1.01\sigma \cong \sigma$. Assuming 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 a $N(\kappa_i, \sigma_{ef}^2)$, Lenth defines $s_0 = 1.5 \cdot \text{median}|c_i|$ and calculates a new median excluding the estimated effects whith $|c_i| > 2.5 s_0$. By doing so he expects to exclude the effects with $\kappa > 0$ and use the others to calculate the so-called Pseudo Standard Error:

$$PSE = 1.5 \cdot \underset{|c_i|<2.5 s_0}{\text{median}}|c_i| \tag{1}$$

From this PSE, a margin of error *(ME)* can be calculated. For a 95% confidence level it will be $ME = t_{0.975,\nu} \cdot PSE$. If $|c_i| > ME$ the effect c_i is considered active.

Lenth proposes that $\nu = n/3$ where n is the number of effects considered. This is the value that has been used in some software packages (e.g. Minitab) although it has been demonstrated to produce type I error probabilities under 5% with the additional and unavoidable inconvenience to produce bigger type II errors. Ye and Hamada (2000) and Fontdecaba *et al.* (2015) proposed t values that give better results (Table 1).

Table 1: Proposed values for $t_{0.975}$ that should be applied with the PSE

The other option is to assume from the outset that some effects are negligible and to use their values c_j $(j = 1, \dots, m)$ to estimate their variance:

$$s_{ef}^2 = \frac{\sum_{j=1}^{m} c_j^2}{m} \tag{2}$$

In general, interactions of three or more factors are considered negligible. In some cases technical expertise of the phenomenon being studied allow to consider negligible some particular two factor interaction and add them to the ones used to estimate s_{ef}^2.

The problem is that the two analytical methods do not always give the same result. (Montgomery, 2013, p. 279), taken from Bell *et al.* (2006) presents a 2^4 design which was conducted to test new ideas to increase direct mail sales of credit cards. The response is the number of orders obtained and the factors are related to the offered conditions (Table 2).

Table 2: 2^4 design with the results obtained from Montgomery (2013) p. 279

Computing Lenth's *PSE* from the effects we get $PSE = 11.4375$ and $ME = 2.57 \cdot 11.4375 = 29.40$ so with a 95% confidence the active effects are $A = 30.37$, $B = -38.88$ and $D = -37.37$. Statistical software packages use always analytical methods to suggest which effects should be considered active, even when they show the results graphically in a normal or other plot. For example, Minitab (Minitab, 2010) uses Lenth's method and marks as significant *A, B* and *D,* both in the NPP chart (Figure 1) and also in the Pareto chart were it draws a line at 29.4 value.

Figure 1: Normal Probability Plot presented by Minitab from the effects obtained in the Example of Montgomery (2013) p. 279

However, if we estimate the variance of the effects considering that the interactions of three or more factors are zero, we have $s_{ef} = 5.24$ and using also a 95% confidence level the effects that should be considered significant are the ones with a value higher than $t_{0.975,5} \cdot 5.24 = 13.46$. Therefore, $C = 18.88$ and $AB = -22.63$ will also be considered significant. This is the result given, for instance, by Statistica (Statistica, 2015) which by default analyse the significance of the effects by this method. (Figure 2).

Figure 2: Pareto chart of effects presented by Statistica for the effects obtained in the Example of Montgomery (2013) p. 279

3. Tested scenarios and simulation

In order to identify the conditions under each method gives better results we have considered a set of combinations of the number and the magnitude of active effects that are a good reflection of the variety of situations that can occur in practice. In each situation we consider that k_i $(i = 1, \dots, n)$ are the effects of interest and their estimators c_i are distributed according to $N(k_i, \sigma_{ef})$. Then, without loss of generality, we can fix $\sigma_{ef} = 1$ so that the subset of *j* inert effects follows a $N(0, 1)$ distribution and the $(n - j)$ subset of active effects is distributed as a $N(a\Delta,\ 1)$ where a can be different

for each effect and Δ are called Spacing and varies in all cases from 0.5 to 8 with increments of 0.5 as it is done in the references cited below.

For eigth-run designs we use the same testing scenarios that were used by Fontdecaba *et al.* (2015):

S8-1: $k_1 = \cdots = k_6 = 0, k_7 = \Delta$
S8-2: $k_1 = \cdots = k_5 = 0, k_6 = k_7 = \Delta$
S8-3: $k_1 = \cdots = k_4 = 0, k_5 = k_6 = k_7 = \Delta$
S8-4: $k_1 = \cdots = k_4 = 0, k_5 = \Delta, k_6 = 2\Delta, k_7 = 3\Delta$

And for 16-run designs the same that were used by Venter and Steel (1998), and later also by Ye *et al.* (2001) and Fontdecaba *et al.* (2014):

S16-1: $k_1 = \cdots = k_{14} = 0, k_{15} = \Delta$
S16-2: $k_1 = \cdots = k_{12} = 0, k_{13} = k_{14} = k_{15} = \Delta$
S16-3: $k_1 = \cdots = k_{10} = 0, k_{11} = \cdots = k_{15} = \Delta$
S16-4: $k_1 = \cdots = k_8 = 0, k_9 = \cdots = k_{15} = \Delta$
S16-5: $k_1 = \cdots = k_{12} = 0, k_{13} = \Delta, k_{14} = 2\Delta, \ k_{15} = 3\Delta$
S16-6: $k_1 = \cdots = k_{10} = 0, k_{11} = \Delta, \ k_{12} = 2\Delta, \ k_{13} = 3\Delta, \ k_{14} = 4\Delta, \ k_{15} = 5\Delta$

We have conducted 10,000 simulations for each scenario and each value of Spacing, using the R statistical software package [15].

These situations have been analysed by Lenth's method as well as by estimating σ_{ef}^2 from the effects considered negligible, which have been selected at random from the ones with $k = 0$. The number of negligible effects go from one to three in eight-run designs, and from one to six in 16-run ones, far beyond the normal situation of having 1 (rarely 2) in 8 runs and up to 5 in 16 runs designs.

As an example, one result from the S8-4 scenario with $\Delta = 3$, is indicated in Table 3. There are four inert effects: $k_{1,2,3,4} = 0$, and three active effects: $k_5 = 1\Delta = 3$, $k_6 = 2\Delta = 6$ and $k_7 = 3\Delta = 9$. With these c_i values we get a Lenth's PSE = 1.84 and using the t-value proposed by Lenth (t = 3.76) the effects that show $|c_i| > 6.92$ should be considered active. In this case only c_7. Since $k_5 = 3$ and $k_6 = 6$ are different from zero, the Lenth method has produced two type II and zero type I errors. If we use c_1 and c_2 – randomly taken from the effects with $k = 0$– to estimate the variance of the effects with two degrees of freedom, we obtain $s_{ef}^2 = 1.12$ which gives a critical value of $t_{0.975,2} \cdot s_{ef} = 4.56$. Therefore, c_6 and c_7 are considered active and c_5 not. Thus, the method has produced just one type II error.

Table 3: Results with the values of the effects obtained by simulation for a design with 8 runs, S8-4 scenario, Δ = 3 estimating the variance with 2 df

After generating 10,000 sets of effects for each scenario and Δ value and identifying active ones by the two methods, we calculated the percentage of type I and type II errors.

The total number of possible type I errors is not the same in both cases, because in the variance estimation method the effects used to estimate the variance are not analysed. For instance, with the S8-4 scenario and Δ = 3 the Lenth method can commit up to 40,000 type I errors (4 in each situation) while the variance estimation method can only commit 20,000 type I if the variance is estimated with 2 df, because those who are supposed zero and are not analysed. With both methods there are 30,000 opportunities of type II error. The results obtained in this specific case are in Table 4.

Table 4: Type I and Type II errors found in the 10,000 simulations with scenario C8-4 with Δ = 3 and estimating the variance with 2 df.

4. Obtained results

In this section we present the results of all simulations. For the Lenth method the results are calculated using Lenth's original t values (still in use in several places like Minitab (Minitab, 2010) and also the t value proposed by Fontdecaba *et al.* (2015).

4.1 Eight-run designs

The percentage of type I errors produced by estimating the variance by the effects considered negligible is, as expected, around 5%. When the Lenth's method is applied the percentage varies depending on the scenario and spacing. Figure 3 shows the percentage of type I errors for each scenario and Spacing value.

Figure 3: Eight-run designs. Percentage of type I errors using Lenth's method (long dashed: t=3.76; short dashed; t= 2) and estimating the effects' variance with 1, 2 and 3 df (solid lines)

The most important differences, and probably also the most relevant in this context of the design of experiments occur in type II errors (Figure 4). The percentage of these errors always decreases while increasing the Spacing value.

Figure 4: Percentage of type II errors using Lenth's method (long dashed: t=3.76; short dashed; t= 2) and estimating the effects' variance with 1, 2 and 3 df (solid lines)

Table 5 shows the number of degrees of freedom used to estimate the variance of the effects that are needed to get a better result than Lenth's method. Results from Lenth method are calculated with $t = 3.76$ and $t = 2$. For instance, for scenario 4, if we use two degrees of freedom or more, type II errors will be smaller than using the method of Lenth with $t = 3.76$, while with $t = 2$ three degrees of freedom or more are needed to get smaller type II errors.

Table 5: Eight runs. Degrees of freedom needed for the negligible interaction method to give lower probability of type II errors than the Lenth's method

Therefore, in an eight-run design, to analyze the significance of the effects by estimating their variance with only one degree of freedom has, in all scenarios considered, bigger probabilities of type II error than with the Lenth method.

4.2 Sixteen-run designs

Figure 5 shows type I errors in 16-run designs' scenarios. As in the eight run designs case, the negligible interaction method produces errors at around 5% whereas Lenth's method varies depending on the scenario and the spacing.

Figure 5: Sixteen-run design. Percentage of type I errors using Lenth's method (long dashed: t=2.57; short dashed; t= 2) and estimating the effects variance with 1 up to 6 df (solid lines)

As before, in the negligible interaction method the proportion of type II errors decreases when the number of degrees of freedom increases and in Lenth's method this proportion is always smaller with $t = 2$ (Figure 6). Table 6 indicates the number of degrees of freedom the negligible interactions method needs to outperform Lenth's method for different scenarios.

Figure 6: Sixteen-run design. Percentage of type II errors using Lenth's method (long dashed: t=3.76; short dashed; t= 2) and estimating the effects variance with 1 up to 6 df (solid lines, from top to bottom)

Table 6: Sixteen runs. Degrees of freedom needed for the negligible interaction method to give better results than the Lenth's method

5. Conclusions and final remarks

We have analyzed the two most widely used analytical methods to judge the significance of effects in un-replicated factorial designs, and we have seen that they do not always produce the same result and that in such cases the best method is not always the same. However, we have seen that in some situations one is clearly better than the other. In the bullet points below we summarize the conclusions from the study and give some recommendations to practitioners, as well as to software makers:

- There are better alternatives for the t values than those proposed by Lenth. This is not new (see, for instance, Ye and Hamada (2000) or Fontdecaba *et al.,* 2015) and the study makes it evident once again. An improvement point for several widely used statistical software packages.
- To estimate the variance of the effects with a single degree of freedom is a bad practice and nearly always worse than to apply the method of Lenth. Some software packages analyse by default the significance of the effects considering negligible the interactions of three or more factors, and they do this also for 2^3 designs in which, obviously, there is only one three factor interaction.
- In eight-run designs, except when we face scenario 3, at least three degrees of freedom are needed for the negligible interaction method to be better than Lenth's method. Practically we will never know a priori that three of the seven contrasts are negligible, and therefore, as a general rule, in eight runs designs it is better to apply the method of Lenth.
- In 16-run designs, the negligible interactions method provides better results (or almost the same in scenario 1) when 5 or more degrees of freedom can be used for variance estimation. Naturally, this happens in complete 2^4 when, interactions of three or more factors are considered negligible.

- A final recommendation to practitioners is to not forget that in many cases normal probability plots, technical knowledge and common sense will solve the discrepancy.

Applying this recommendations to Montgomery's (2013) p. 279 example, the advice is to use the negligible interactions method, and using it the effects of C and AB would be considered active. A very reasonable assessment by looking at the normal probability plot.

References

Bell, G.H., Ledolter, J. & Swersey, A.J. (2006). Experimental design on the front lines of marketing: Testing new ideas to increase direct mail sales, *International Journal of Research in Marketing,* 23:309-319.

Box, G.E.P., Hunter, J.S. & Hunter, W.G. (2005). *Statistics for experimenters: design, innovation, and discovery*. New Jersey: Wiley.

Box, G.E.P. & Meyer, R.D. (1986). An Analysis for Unreplicated Fractional Factorials. *Technometrics,* 28:11-18.

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

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

Espinosa, V., Dasgupta, T. & Rubin, D.B. (2016). A Bayesian Perspective on the Analysis of Unreplicated Factorial Experiments Using Potential Outcomes. *Technometrics,* 58: 62-73.

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

Fontdecaba, S., Grima, P. & Tort-Martorell, X. (2015). Proposal of a single critical value for the lenth method. *Quality Technology and Quantitative Management,* 12:41-51.

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

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: 1383-1403.

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

Minitab 17 Statistical Software (2010). [Computer software]. State College, PA: Minitab, Inc.

Montgomery, D.C. (2013). *Design and analysis of experiments*. Singapore: Willey.

R Development Core Team. (2011). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing.

Statistica 13.0 (2013). [Computer software]. Dell Software.

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

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

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