The Case Against Normal Plots of Effects

RUSSELL V. LENTH

The University of Iowa, 241 Schaeffer Hall, Iowa City, IA 52242, USA

When analyzing effects in an unreplicated experiment, normal or half-normal plots of the effects (also called Daniel plots) are a popular way to visualize them and to judge which are active. This article discusses the ways in which these plots can be confusing and misleading. There are other methods available that are less subjective, easier to explain, more powerful, and less likely to be misinterpreted. I recommend against using Daniel plots, even as a supplement to one of these better analyses.

Key Words: Daniel Plots; Normal Plots; Unreplicated Experiments.

1. Introduction

In the analysis of unreplicated 2^k or 2^{k-p} experimental designs, normal or half-normal plots of effects (Daniel, 1959), also called Daniel plots, are a popular method for judging which effects may be active. Many statistical packages, such as Minitab, Design Expert, and JMP, provide these plots automatically as part of the analysis of such designs. Daniel plots are by far the most popular method for analyzing unreplicated experiments. They are so popular that it is not unusual to see them used even with a replicated experiment, where there is a valid error estimate and traditional t tests and such are available.

Normal plots are obtained by plotting the estimated effects versus the *normal scores*, these being equally spaced quantiles of the standard normal distribution. Specifically, let $e_1, e_2, \ldots, e_m$ denote the estimated effects (excluding the intercept); let r_j denote the rank of e_j among all the e_i; and let Φ^{-1} denote the inverse cdf of the standard normal distribution. Then the normal scores are $n_j = \Phi^{-1}\{(r_j - 3/8)/(m + 1/4)\}$.

Similarly, letting s_j denote the rank of $|e_j|$ among all the $|e_i|$, the half-normal scores are defined as $h_j = \Gamma^{-1}\{(s_j - 3/8)/(m + 1/4)\}$, where $\Gamma^{-1}(p) = \Phi^{-1}\{(1/2)(1+p)\}$ is the inverse half-normal cdf. The plot of the $|e_j|$ versus the h_j is the half-normal plot of the effects. Henceforth, we use the term "Daniel plot" to refer to either the normal or the half-normal plot of effects.

The idea of the Daniel plot is that the e_j are mutually independent and have the same variance. If the true effects are all zero, then a plot of the e_j versus the n_j (or likewise a plot of the $|e_j|$ versus the h_j) will fall approximately on a straight line. But if just a few true effects are nonzero (this is called the *effect-sparsity model*), then most e_j (or $|e_j|$) will still display near a straight line, but the few active (nonzero) ones will deviate from the line; so one may be able to visualize them as outliers in the plot.

2. Two Examples

Table 1 displays two datasets and the estimates of their effects. Response y_1 (color of a chemical product) is from an experiment reported in Snee (1985) (also presented in Montgomery (2013), problem 8.9). There are five factors and this design is a half fraction of resolution V generated by $E = ABCD$. Response y_2 (yield of a process) comprises data from a 2^{4-1} experiment presented in Montgomery (2013), problem 8.15. This experiment is generated by $D = ABC$; hence, there are data only for the cases where this relationship holds. The effect estimates (in numerical order) for each experiment are also shown in Table 1. Because each is a half fraction, each effect is a combination of effects of two alias pairs, as shown in the tables. We refer to these datasets as Example 1 and Example 2, respectively.

Dr. Lenth is Professor Emeritus of Statistics in the Department of Statistics and Actuarial Science. He is a Fellow of ASA and an Associate Member of ASQ. His email address is russell-lenth@uiowa.edu.

TABLE 1. Two Published Datasets and Their Estimated Effects. The data for y_1 are from Snee (1985), a 2^{5-1} fractional factorial in factors A–E. The data for y_2 are from Montgomery (2013), problem 8.15. It is a 2^{4-1} design in factors A–D

Factor levels					Responses		y_1 Analysis		y_2 Analysis	
A	B	C	D	E	y_1	y_2	Effect	e_j	Effect	e_j
−1	−1	−1	−1	1	−0.63	12	$AD+BCE$	−0.678	$AC+BD$	−2.125
1	−1	−1	−1	−1	2.51		$B+ACDE$	−0.670	$AB+CD$	−0.375
−1	1	−1	−1	−1	−2.68		$E+ABCD$	−0.414	$AB+CD$	−0.375
1	1	−1	−1	1	1.66	16	$AC+BDE$	−0.394	$B+ACD$	0.125
−1	−1	1	−1	−1	2.06		$CD+ABE$	−0.356	$C+ABD$	1.375
1	−1	1	−1	1	1.22	15	$CE+ABD$	−0.120	$A+BCD$	1.875
−1	1	1	−1	1	−2.09	20	$C+ABDE$	−0.074	$BC+AD$	2.125
1	1	1	−1	−1	1.93		$DE+ABC$	0.044		
−1	−1	−1	1	−1	6.79		$BC+ADE$	0.084		
1	−1	−1	1	1	5.47	25	$BD+ACE$	0.123		
−1	1	−1	1	1	3.45	13	$BE+ACD$	0.144		
1	1	−1	1	−1	5.68		$AE+BCD$	0.151		
−1	−1	1	1	1	5.22	19	$AB+CDE$	0.638		
1	−1	1	1	−1	4.38		$A+BCDE$	0.655		
−1	1	1	1	−1	4.30		$D+ABCE$	2.210		
1	1	1	1	1	4.05	23				

For starters, let us look at half-normal plots of the effects for the two examples. They are shown in Figure 1. Other than descriptive axis labels (as all plots should have), these plots are completely unannotated. Thus, we can judge the plots themselves and not be biased or distracted by any other elements.

In teaching how to use Daniel plots, the typical advice is to look for outliers in the plot. Following this rule, there is one point that clearly sticks out in Example 1. It happens to correspond to the D effect, or possibly its alias $ABCE$. The remaining effects seem to follow a fairly linear pattern, though

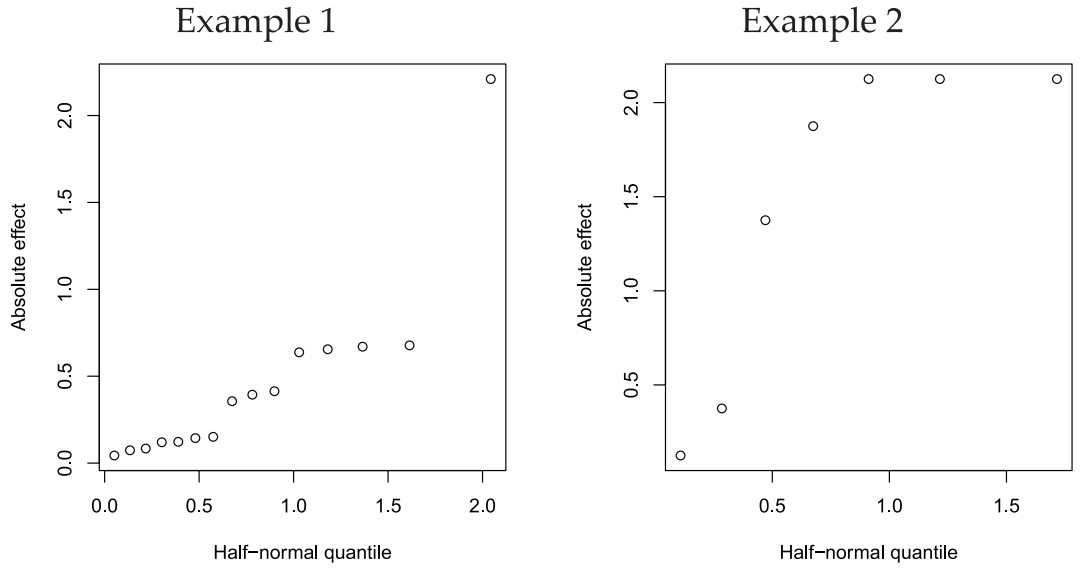


FIGURE 1. Half-Normal Plots for the Two Sets of Estimated Effects.

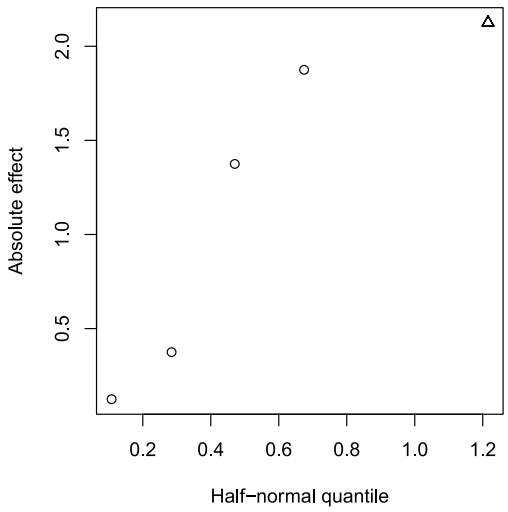


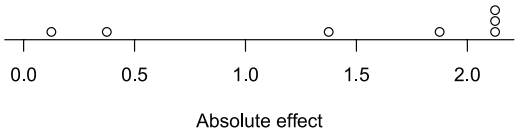
FIGURE 2. Half-Normal Plot for Example 2, Using Tied Ranks.

perhaps if we enhance the plot with a line, it may be possible to see other deviant points—we'll look at that later. For now, based on judgment of the half-normal plot, we will say that there is one active effect, D.

As for Example 2, there are two points that stick out, suggesting two active effects. Looking at Table 1, we note that these points correspond to the two largest effects, $D + ABC$ and $BC + AD$. But wait! Look at the top of the table and note that there is a third effect, $AC + BD$, with the same absolute value of 2.125. The fact is, there is a three-way tie for the largest effects. So are there three active effects (because of the three-way tie) or are there none (because one of the tied effects seems to be in a line with the smaller ones)?

Perhaps we can answer this question by noting that normal scores depend on ranks. What happened in Figure 1 is that small differences in machine precision cause those three absolute effects to have different ranks, whereas we should have averaged those three ranks together and given all three effects the same half-normal score. A revised half-normal plot based on the averaged ranks is shown in Figure 2. The rightmost point (the triangular symbol, representing three tied effects) still seems to deviate from the linear pattern. Based on the idea of calling outlying points active, we have that all three effects, $D + ABC$, $BC + AD$, and $AC + BD$, are active.

However, you may also note that the outlying points in the second example deviate *horizontally* from the linear pattern, whereas the active effect in Example 1 deviates *vertically*. Does this make a difference? Of course it does. What it means in this example is that the three largest absolute effects are outliers because they are *smaller* than what we'd expect relative to the other effects. Even a simple dotplot makes this clear:



Active effects would be noticeably distant from the other effects. But here, the largest three absolute effects are noticeably nearby, and that's why they look like outliers in the Daniel plot.

3. Adding a Reference Line

Often a Daniel plot is constructed with additional elements to aid in interpretation. The most common is a reference line to help assess linearity. However, all sorts of practices are out there for how to define that line. For example, should it be a least-squares line—perhaps after deleting some effects? If so, how many effects to remove? Some of these questions can be answered by appealing to the underlying effect-sparsity principle—that inactive effects come from a normal distribution with mean zero and some unknown standard deviation τ. Because of the mean of zero, a reference line should pass through the origin. Moreover, assuming that the effects or absolute effects are plotted on the vertical axis and the (half-) normal scores are on the horizontal axis, the slope of the line should be an estimate of τ. There are several ways to estimate τ using some kind of robust scale estimate; see Hamada and Balakrishnan (1998) for a survey of the most popular ones. For purposes of our illustrations, we'll simply use the median of the absolute effects, suitably scaled, so that the reference line passes through the origin and the point where the median absolute effect is plotted. We will refer to this as the "median reference line".

Figure 3 displays the half-normal and normal plots for Example 1, with median reference lines added. There is only one effect that deviates substantially from the line in the half-normal plot. This supports our earlier impression that there is only one active effect. In the normal plot, we see a deviant point at the *left* end. However, it is *above* the reference line, indicating it is less deviant than expected, like happens in Example 2.

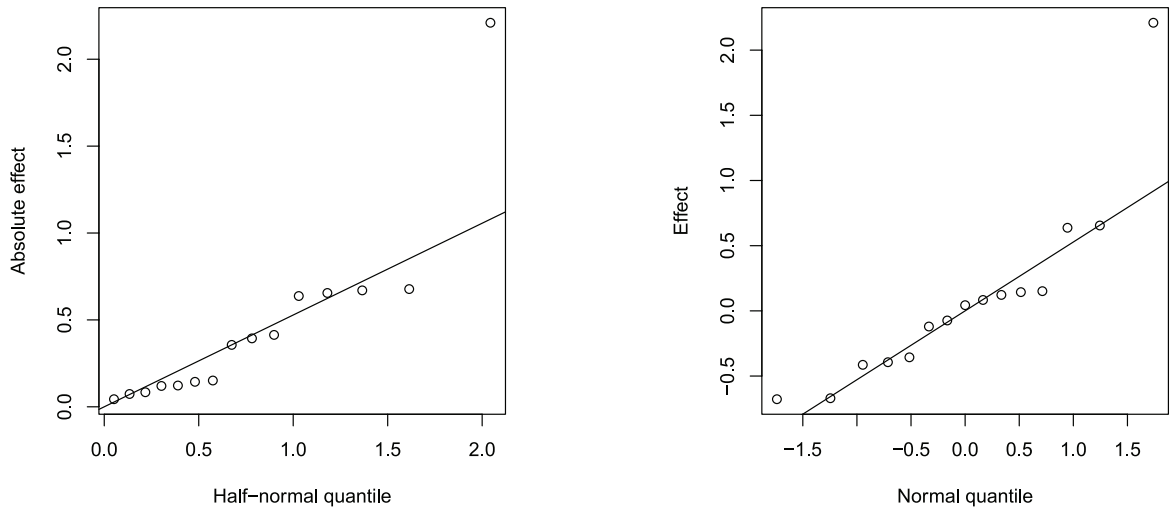


FIGURE 3. Half-Normal and Normal Plots for the First Example, with Median Reference Lines Added.

Figure 4 shows the same two plots for Example 2. The reference line for the half-normal plot seems reasonable and again does not alter our conclusions. However, the very same reference line goes nowhere near most points in the normal plot. The main reason is that most of the effects are positive. If we swapped some of the signs of the effects, this would not affect the median reference line, but the points would fit *it* better. Also shown in the plot, as a dashed line, is the least-squares line based on the middle five points. It obviously fits much better but is not an appropriate reference line for judging whether effects are active, as it departs from the underlying premise that the inactive effects should have mean zero. The chain-dashed line uses regression through the origin, so in that sense it is an appropriate reference line, but with a smaller (more optimistic) estimate of τ, because its slope is lower than that of the solid reference line. It is interesting that it goes nearest the two points that were excluded from the model! All in all, the interpretation of the normal plot is very confusing indeed.

If you want to use such plots, it is worth trying to find out what rules your software uses for adding a reference line, because there is no set rule for this.

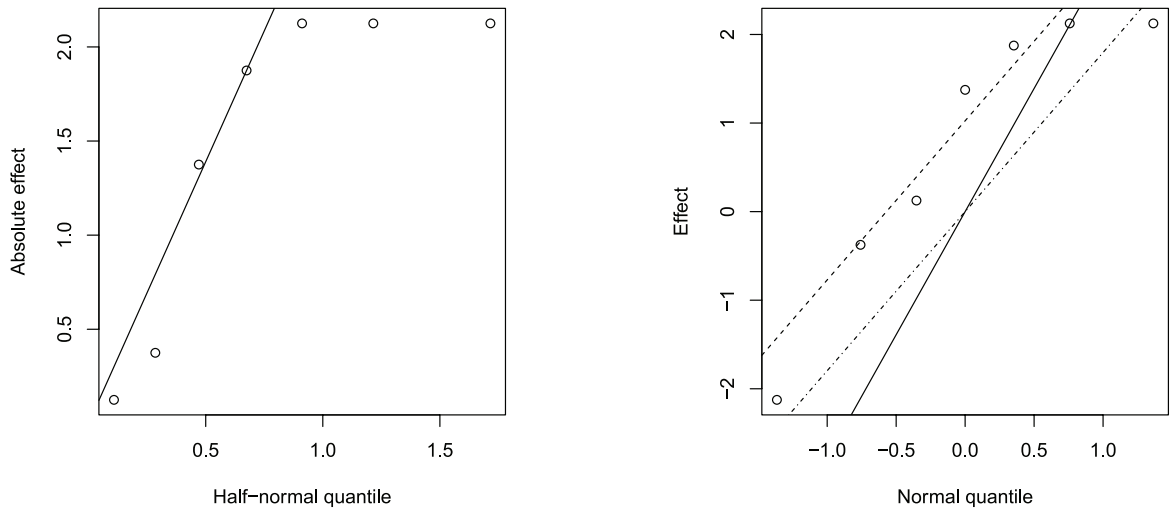


FIGURE 4. Half-Normal and Normal Plots for the Second Example, with Median Reference Lines (Solid Lines) Added. The dashed line and chain-dashed lines are the least-squares lines for the middle five effects, with and without the intercept.

It is especially important to know whether the line passes through the origin. If it doesn't, you are judging linearity of the plot—a separate issue from the goal of judging which effects are active.

Besides reference lines, Daniel plots are often further annotated with the names of one or all of the effects. This can be done either automatically by the software or manually by the user clicking on points. Some software changes the reference line based on which points are identified, assuming that those are surmised as active effects. This seems like a reasonable idea, but there is the hazard of exaggerating the importance of effects that are wrongly identified as active, as could happen in Example 2.

4. Objective Methods

I have already alluded to the fact that objective methods exist for assessing effects. Many of them use a robust estimate of scale to estimate the standard error τ of the effects. The underlying model for most of these is the effect-sparsity model—that the effects are independent, have the same standard error, and the majority of them have expectation zero. Another approach (Box and Meyer, 1986) uses a Bayesian specification of the effect-sparsity model and determines the posterior probability that each effect is active.

When an estimate $\hat{\tau}$ of τ is obtained, then it can be used to construct t-like ratios $t_j = e_j/\hat{\tau}$ for testing the effects. Perhaps a critical value or margin of error of the estimates is produced in addition. These results are most effectively displayed in a bar graph or Pareto chart, preferably with cutoff lines for significance based on $\hat{\tau}$. Figure 5 shows such displays produced by JMP, using its Analyze/Modeling/Screening menu. They are Pareto charts in the sense that the main effects are shown in decreasing order of absolute effects. The plots also include the effect estimates (here called "contrasts"), the t ratios obtained using the method in Lenth (1989), and P values obtained by simulating 1000 cases from the null distribution. An individual P value is the fraction of simulated t statistics that exceed the stated value. The simultaneous one is instead based on the distribution of the maximum of the entire set of t statistics; this adjusts for the error rate of the whole family of t tests.

According to the P values for Example 1, there is one effect with $P \approx .001$ and four more with $P \approx .10$—a kind of middling level of significance. These together comprise the rightmost five points in the half-normal plot in Figure 1 or Figure 3. But four of those points do not stand out in the plots and are unlikely to be identified as active effects, or even marginal ones, by examining the Daniel plots.

As for Example 2, the three largest effects have P values of about .39, so do not even come close to being identified as active.

Most important, the graphs in Figure 5 are completely sufficient for judging the contributions of model effects. They present the effects directly as the lengths of bars, and cutoff lines and P values add directly useful guides for judging them.

5. Explainability and Teachability

Statistical methods are used to explain phenomena that we observe to someone who needs that information to do a better job, improve a process, or advance science. Thus, a statistical analysis is only as effective as our ability to explain it to a client. A related matter is teaching these methods. A lot of in-house short courses in industry, as well as university service courses in statistics, are directed to nonstatisticians who potentially will find use for statistical methods. We do not spend much time with these audiences and they do not learn statistics in great depth. So it is important that what we do teach them will be understood and will likely be used appropriately.

Daniel plots are not easy to explain or teach. In Example 2, we learned that it makes a difference whether points deviate horizontally or vertically. But did you teach that little nuance in your short course for the marketing department? Maybe not. A related side issue is that, originally, normal or half-normal plots were constructed using special graph paper, and often the probablility scale on that paper was on the horizontal axis. Because of this, depending on what book you read or software you use, the axes in the Daniel plot could be switched the other way—the half-normal scores on the vertical axis and the absolute effects on the horizontal axis. This means that it is not enough to tell people to look for deviant effects in a particular direction—you have to teach them to read the axis labels carefully, then follow one of two rules, depending. Add to that the difficulty of explaining how normal plots are constructed (see the introduction: inverse cdfs, ranks, etc.) This is not looking like a very easy-to-teach subject.

Sometimes, when it is the best thing available, a difficult or complex technique is worth spending the

Example 1:

Contrasts

Term	Contrast	Lenth t-Ratio	Individual p-Value	Simultaneous p-Value
D	2.21000	5.81	0.0026 *	0.0159 *
B	-0.67000	-1.76	0.0976	0.6315
A	0.65500	1.72	0.1030	0.6578
E	-0.41375	-1.09	0.2639	0.9812
C	-0.07375	-0.19	0.8608	1.0000
D*B	0.12250	0.32	0.7735	1.0000
D*A	-0.67750	-1.78	0.0954	0.6210
B*A	0.63750	1.67	0.1101	0.6911
D*E	0.04375	0.11	0.9173	1.0000
B*E	0.14375	0.38	0.7358	1.0000
A*E	0.15125	0.40	0.7228	1.0000
D*C	-0.35625	-0.94	0.3262	0.9979
B*C	0.08375	0.22	0.8428	1.0000
A*C	-0.39375	-1.03	0.2837	0.9902
E*C	-0.12000	-0.32	0.7767	1.0000

Example 2:

Contrasts

Term	Contrast	Lenth t-Ratio	Individual p-Value	Simultaneous p-Value	Aliases
D	2.12500	0.76	0.4040	0.9969	
A	1.87500	0.67	0.5002	1.0000	
C	1.37500	0.49	0.6867	1.0000	
B	0.12500	0.04	0.9702	1.0000	
D*A	2.12500	0.76	0.4040	0.9969	C*B
D*C	-0.37500	-0.13	0.9115	1.0000	A*B
A*C	-2.12500	-0.76	0.4040	0.9969	D*B

FIGURE 5. Pareto Charts from JMP's Screening Analysis for Each of the Examples.

time to teach or explain carefully. However, Daniel plots are not in that category because there definitely are better analyses available that are easier to explain (see the preceding section). Granted, the details of calculating a standard-error estimate may be daunting to explain. But explainability has more to do with whether what is shown makes sense to the viewer. Graphs like those in Figure 5 have an inherent sensibility to them, even without an explanation. And the numerical details—t ratios, P values, etc.—are similar to those in many other statistical analyses, so there is good potential for understanding based on transfer of knowledge.

De León et al. (2011) reports a study wherein datasets were simulated, then subjects were asked to judge effects using both normal plots and simple dotplots. Their study includes 8- and 16-run designs (just like the examples in this article). They used 8 simulations of each experiment size, and the effects from each experiment were plotted on normal (not half-normal) plots as well as dotplots, creating 32 graphs total. They then recorded the perceptions of all 32 plots by 211 engineering students. Both types of plots were devoid of annotations or reference lines. The plots were intermixed, and subjects were not told that the dotplots came from the same sets of ef-

fects as the normal plots. Respondents were asked to circle the points that they thought corresponded to active effects. Because the data were simulated from known parameter settings, it was possible to identify when type I and type II errors were made. They found that, for the 16 datasets included, there was no evidence to suggest that normal plots are any better than dotplots in either their type I or type II error rates. They also found that type II errors were much more common with both types of plots.

I disagree with De León et al.'s choices for their simulations. They included no complete null cases (where no effects are active), and several of their scenarios include more active than nonactive effects—violating the effect-sparsity model. However, I arrived at the same conclusions as theirs when looking only at the cases where 1/3 or less of the effects are active. Those authors recommend displaying the dotplot in addition to the normal plot. They do not mention the objective methods in the preceding section.

6. Conclusions

Compare the Pareto plots in Figure 5 with the Daniel plots presented in Figures 1, 3, and 4. The Pareto plots display the effects side by side, while the Daniel plots display them along a crooked path. Judgment of their significance is a direct product of the Pareto plot, whereas a Daniel plot requires a subjective judgment with no clear cutoffs or numerical guidance. And as we have seen, those subjective judgments are prone to error—both in failing to identify potentially important effects, as in Example 1, and in tricking the unwary observer into misidentifying unimportant effects as active ones, as in Example 2. Additionally (also Example 2), near ties of effects look drastically different from exact ties (Figure 1(b) versus Figure 2). De León et al. (2011) put it well: interpretation of normal plots is "enveloped in a cloud of mystery and might easily lead to gross errors". Even a simple dotplot of effects is no worse, while being much easier to explain.

Users and software developers are well advised to abandon Daniel plots as a recommended (or even an available) method of analysis. Simpler, objective methods exist for identifying active effects. These methods are more reliable, direct, explainable, and defensible than subjective judgments. It is not as if the objective methods are new or even computationally difficult. Most, in fact, involve very simple calculations that can even be done by hand. They are easier to compute than normal scores.

I recommend against using a Daniel plot even as a supplement to the Pareto plot. It adds no useful information and can only add confusion and distraction. Just say "no" to Daniel plots. I recognize that they are supremely popular and that change is difficult. But when so much better methods are out there, it is irresponsible to continue using them.

References

Box, G. E. P. and Meyer, R. D. (1986). "An Analysis for Unreplicated Fractional Factorials". *Technometrics* 28, pp. 11–18.

Daniel, C. (1959). "Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments". *Technometrics* 1, pp. 311–341.

de León, G.; Grima, P.; and Tort-Martorelli, 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, pp. 161–174.

Hamada, M. and Balakrishnan, N. (1998). "Analyzing Unreplicated Factorial Experiments: A Review with Some New Proposals". *Statistica Sinica* 8, pp. 1–28.

Lenth, R. V. (1989). "Quick and Easy Analysis of Unreplicated Factorials". *Technometrics* 31, pp. 469–473.

Montgomery, D. C. (2013). *Design and Analysis of Experiments*, 8th edition. Wiley.

Snee, R. D. (1985). "Experimenting with a Large Number of Variables". In *Experiments in Industry: Design, Analysis and Interpretation of Results*, R. D. Snee, L. B. Hare, and J. B. Trout, eds. Milwaukee, WI: Quality Press.

Discussion

BRADLEY JONES

SAS Institute, Cary, NC 27513

I AGREE with Professor Lenth that the use of normal or half-normal plots of effects is now outmoded. There are better analytical tools now available in commercial software for model selection in screening designs. My favorite for the analysis of unreplicated regular fractional factorial designs is Lenth's method as implemented in JMP using Monte Carlo simulation from the null model to obtain p-values for Lenth's pseudo-t statistic.

Here are several reservations about the Daniel plot.

1. If many of the effects are active, it can miss more than one active effect.
2. If there is one huge effect, other effects seem small by comparison—see Figure 1.
3. Plotting order statistics means that the plotted points are correlated, producing patterns that are easily misinterpreted.
4. If there is a severe outlier or a missing observation, then the loss of that point makes the design nonorthogonal and then the Daniel plot is not applicable.
5. The theory behind these plots depends on the orthogonality of all the factorial effects. That is, the design must be a full factorial or a regular fractional factorial. Designs that are orthogonal for the main effects but that have correlated two-factor interactions can produce ambiguous Daniel plots.

I sent an example of the normal plot shown in Figure 1 from a 2^6 full factorial design to most of the discussants.

There were widely varying answers about how many effects were active. In my view, the fact that a group of experts looking at Figure 1 do not come close to each other concerning the set of active effects is damning.

As a developer of software, my personal goal is the democratization of DOE. That is, I would like practitioners to be able to design and analyze industrial experiments without necessarily having a consulting statistician at their elbow.

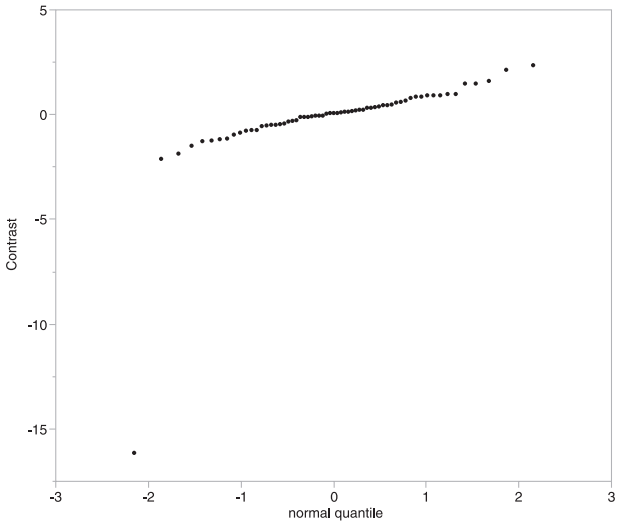


FIGURE 1. Normal Plot of Effects.

What are the main issues in the analysis of screening experiments?

1. Find the active effects—that is the whole point.
2. Check to see if there is any bad data.
3. Check to see if a different response metric works better.

Of course, items 2 and 3 impact item 1, so you need to convince yourself that the data are reasonable and that you don't need to transform the response perhaps iteratively within the process of model selection.

To address item 2 above, it is easier to identify an outlier with a residual plot generated after removing the insignificant terms from the model than to try to see if there are two separate lines in the Daniel plot.

To address item 3 above, providing rules of thumb for power transformations is not too difficult. I base

my own decisions on the range of the observed responses and any knowledge about whether responses must be positive.

It is desirable to provide a graphical display to illustrate the results of the model-selection process. Here a Pareto plot of the effects seems easiest to interpret. Then, a line drawn between the big effects on the left and the small effects on the right separates the "vital few" from the "trivial many" effects.

Reference

LENTH, R. V. (1989). "Quick and Easy Analysis of Unreplicated Factorials". *Technometrics* 31 pp. 469–473.

Discussion

JOSEPH G. VOELKEL

Rochester Institute of Technology, Rochester, NY, 14623

In his 1959 article, Daniel proposes half-normal plots, shows their usefulness through many examples, and finishes by indicating some problems with them. By contrast, Lenth provides a lively proposal, complete with leading examples as well as strong opinions and recommendations. Daniel's article reads like an NPR story, Lenth's like Fox News! I will have to guess that Lenth's writing style was intended to provoke the reader, and he has succeeded with this one.

I would like to make some general comments and then a number of specific replies. My general comments are:

1. The notion of "objective" tests is a myth and one that should be dispelled early and frequently. Such tests, like all statistical tests, are algorithms that use certain tuning parameters and that are based on certain assumptions (some of which may be checkable, but often only to a limited extent). If we change the values of the tuning parameters, we can often arrive at different solutions. Dispelling this myth of objectivity is useful—it enforces the idea that, while decisions need to be made, such decisions are not totally objective.
2. Daniel's paper itself contains an "objective" test. That test, or Lenth's, e.g., is distinct from the argument of half-normal vs. Pareto plots—in fact, just as for tests used in an ANOVA table, no plots are needed at all.
3. Statistical analysis is often not simple. Trying to make it *appear* simple, with a simple set of rules, is not recommended.

   For simplicity, let us crudely group those who will *analyze* an experiment and play a lead role to *write a report* into two extreme but not unreasonable categories: amateur and professional. (Either or neither of these may be involved in the *design* or the *running* of the actual experiment.)

   For the analysis, the amateur is looking for results, has no interest in the theory behind what is being done, and little or no interest in assumptions or data checks. He just wants clear directions on how to proceed; in fact, he wants more: a simple set of rules to follow. We often find this user in short courses and "belt" courses. For this user, a set of rules that works reasonably well a reasonable amount of the time is what we can offer, and about all we can offer. For this user, a Pareto chart of effects with indications of which are active provides such a set of rules, and I agree with Lenth that a Pareto chart appears easier to understand than a half-normal plot. But nothing in life is free—see examples below. (And even in the subset of cases where the Pareto chart of effects leads to a good model, the amateur is still far from performing a good analysis, creating good graphs to communicate the results, and presenting the results adequately in a report.)

   By comparison, the professional is focused on process, understands deeply the theory and the advantages and shortcomings thereof, and, by a mixture of informal and formal methods and an understanding of reality based on readings of past work and his own experience, will perform his analysis. For such professionals, I believe that the half-normal plots provide a nice method for making decisions and for gaining insight—see examples below.

   But does this mean the professional's half-normal plots, part of his *internal analysis*, should be presented to his clients in the main body of report? Usually not. But neither should a Pareto plot of effects—it's hard to see where either would be useful to the typical client. It is rather better to confine the body of the report to present results (not the methods used to reach them)—which factors mattered, graphs that answer the objectives in a way that these particular clients can understand, recommendations, concerns, and so on. (A technical appendix should indicate the methodolog-

ical approach used and its application to the particular data set. This may include plots such as half-normal plots, along with a brief explanation—but that depends on the client.) So, I naturally agree that "[s]tatistical methods are used to explain phenomena that we observe to someone who needs that information to do a better job, improve a process, or advance science. Thus, a statistical analysis is only as effective as *our ability* to explain it to a client." (Italics mine.) However, I don't see how the use of half-normal or Pareto plots help here.

So, for whom is the Pareto plot useful? I would say the amateur—maybe. But if one is conducting a real experiment at a real company for real stakes, I'd prefer the analysis and report be done by a professional. To expect a marketing person (to use Lenth's example) to be able to design, execute, analyze, and report a real experiment correctly after a short course is not reasonable, nor do I believe it is reasonable to pretend this to them—to expect them to understand the value of designed experiments and the idea that it can be analyzed is about as much as we can hope for. (I will offer up a subgroup of engineers and scientists as a partial exception to this rule.) Perhaps the amateur should make the admission in Shakespeare's *Julius Caesar* that "The fault, dear Brutus, is not in our stars, but in ourselves". So, let us not bury the half-normal plot, but praise it.

Next are some specific replies. These are often not intended to contradict Lenth *per se*, but to illustrate how half-normal plots can be useful—at least to the professional. For brevity, I will refer only to references to the experiments, not their descriptions, and will only use one-letter factor abbreviations. I did not go out of my way to find these examples for this discussion—they are ones I have simply come across in readings and practice.

1. Lenth suggests distaste for lack of standardization for the half-normal plot. In addition to his concern about horizontal vs. vertical is a concern about the position of the line on a half-normal plot: "However, all sorts of practices are out there for how to define that line". But lack of standardization also exists for the dizzying array of "objective" rules—the Hamada and Balakrishnan reference provides a list of 24 methods, each of which has a least one tuning parameter. Similarly, even something as simple as a Pareto plot is not standardized—at least one package uses a standard Pareto order, while Lenth's paper uses a Pareto order for the main effects and a second order for two-factor interactions. In fact, neither of these are actually Pareto plots because the cumulative line is (correctly) not part of this plot.
2. "[Half-normal] plots are so popular that it is not unusual to see them used even with a replicated experiment, where there is a valid error estimate and traditional t tests and such are available". A 2^{6-2} experiment with five center points (Box and Draper (2007)) was run to improve the dry strength of plywood. Using the natural t-test from the 4 d.f. available from the center points, and using (IER) $\alpha = 0.05$, a common tuning parameter value, we find from the t-tests that that S, P, SW, and ST are active, with P, T, and SN also active at $\alpha = 0.10$. The half-normal plot tells a different, and I think, better, story: only S and N are active. See (a) in Figure 1, where I display the plots in Daniel's style.

   What happened? Well, either (a) the estimate s from the center points happened to be unusually low for this experiment, leading to an unusual number of declared-significant terms—recall that the points on the half-normal plot are statistically independent, at least under the usual assumptions on the errors, but that these t-tests are not (and especially not so with a small number of error d.f.); or (b) perhaps the center points were run under routine conditions, for which we may expect that the variability might be lower than under experimental conditions. (And, if an amateur *designed* the experiment, the center points are more likely to simply be five samples from what is actually one longer run—but that is another problem!) In fact, a third explanation may be inferred from the description of the problem in Box and Draper (2007). In any event, the amateur who simply follows the rules will likely create an overly complex model. (By the way, in this design, SW is aliased with MT and ST with WM. Many software packages do not make such aliasing clear enough to the user, especially the amateur. In the above, I just displayed what one package reported.)
3. Lenth's Example 1. "In teaching how to use Daniel plots, the typical advice is to look for outliers in the plot. Following this rule, there

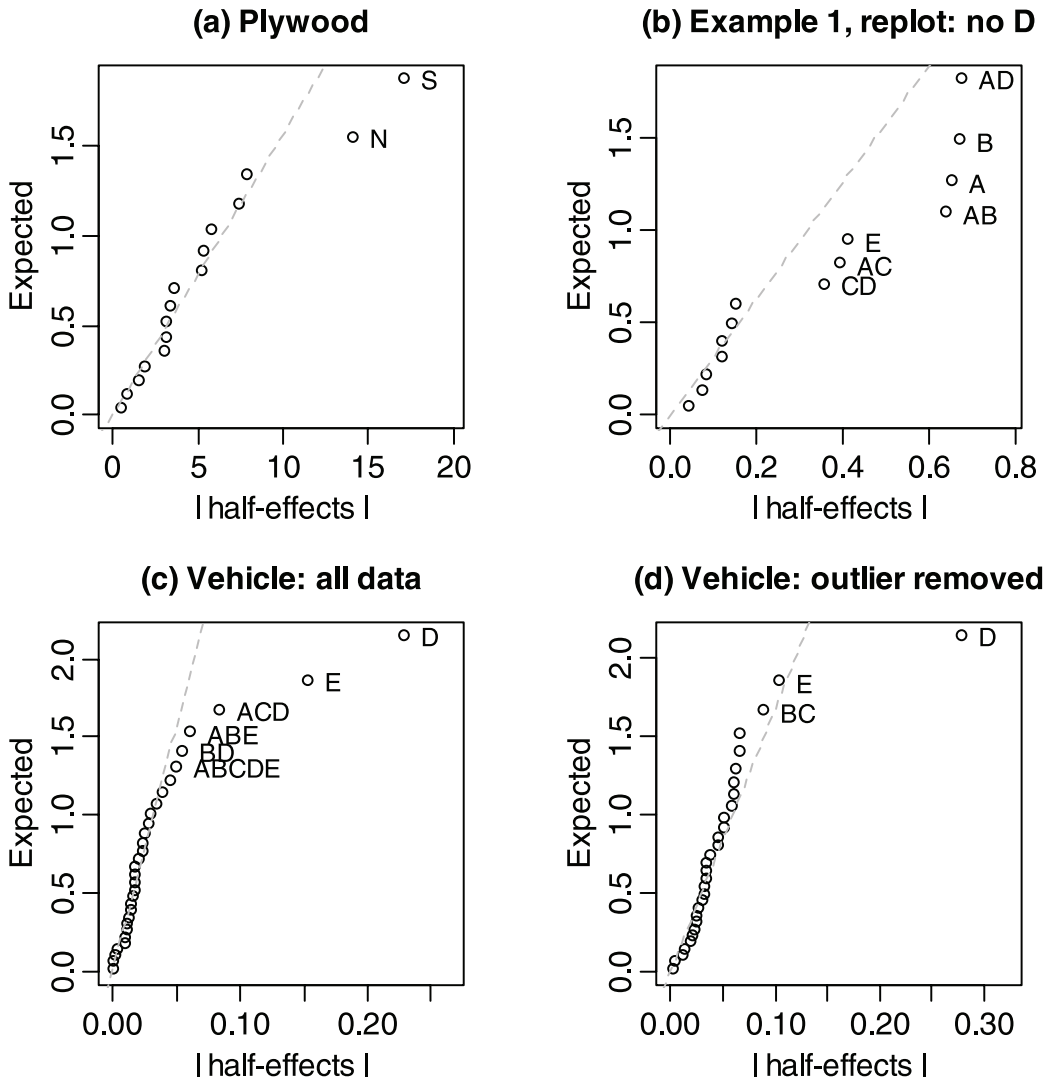


FIGURE 1. Illustrations of Half-Normal Plots for Three Experiments.

is one point that clearly sticks out in example 1. ... For now, based on judgment of the half-normal plot, we will say that there is one active effect, D". And, then, in looking at the Pareto plots, "According to the P values for example 1, there is one effect with $P \approx .001$ and four more with $P \approx .10$—a kind of middling level of significance. These together comprise the right-most five points in the half-normal plot in Figure 1 or Figure 3. But four of those points do not stand out in the plots and are unlikely to be identified as active effects, or even marginal ones, by examining the Daniel plots".

In his first example (p. 315), Daniel notes that, after removal of what are considered to be the largest contrasts, one replots the remaining ones to see what patterns, if any, still remain. (And even if Daniel did not note this, it should be clear to a professional.) Replotting leads to (b) in Figure 1, where the professional can make some decisions. It seems pretty clear that the set of four terms AD, B, A, and AB should either all be included or all be excluded from the model. Because D has already been selected and the remaining terms tell a simple story, a professional would almost certainly include these.

Also, note that these absolute values are similar, but not equal. (But even if they were equal, I would prefer to see them plotted

based on different ranks, for readability and consistency—compare with Lenth's Figure 2 in which tied ranks were used.) In addition, the professional would realize that this near equality corresponds to an often-witnessed phenomenon: $|A| \approx |AD|$ indicates that the effect of A at one level of D is essentially 0; and $|A| \approx |B| \approx |AB|$ indicates that the effect of A and B jointly only exists for one of the four (A, B) combinations. I find this information much clearer to see on a half-normal plot than on the plot provided by Lenth.

Next, for at least two reasons, it's hard to get excited about the triple E, AC, CD. However, if this had been the double C, E the professional would have probably added these to the model. Useful subjectivity! I don't believe these would have been noticed in the Pareto plot. Based on these ideas, I must disagree with the idea that the Pareto graphs "are completely sufficient for judging the contributions of model effects". I will also note that, if the user were following a set of rules and the rule said to use IER $\alpha = 0.10$, then AB would *not* be on Lenth's list ($P = 0.1101$).

4. Example 2. The professional would first have most likely strongly recommended against an eight-run experiment if he were involved in the design stage. But, in the analysis stage at the first glance of the half-normal plot, he would have seen that nothing was active. I will agree that the amateur may have missed this—but it's pretty simple for software folks to lightly gray out the "wrong" side of the plot, and much software incorporates rules to highlight some tentatively active terms. But if the amateur did miss this, and decided to put D and AC (or BD) in his model while respecting hierarchy, he would have found only large P-values in the regression output. But the amateur may have missed this warning too? Do we really want this person to be analyzing experiments for our company?
5. When I teach a short course in DOE, I don't find it hard to explain a normal probability plot. No formulas are presented, only a normal curve; participants are asked to guess where the max of each of $n = 1$, 2, 3, and 7 numbers from a standard normal might lie. On their own—or more often with a slight bit of prodding—many see that splitting up the normal curve with n lines into $n + 1$ equal areas is reasonable, and they then look these up in a normal table. A simulation in software illustrates this idea and sampling variation, including the relatively high amount of "off-the-line" variability for 8-run vs. 16-run designs. I do wave my arms a bit for the transition to the half-normal, but the participants see that the underlying idea is the same. I also provide a rationale, with an example, for preferring the half-normal. I much prefer this, and the acceptance of some subjectivity, to the myth of objectivety. A software package we sometimes use may employ Lenth's method, but I emphasize that (a) this is only a guide and that (b) it is much better to think about the problem and look at the half-normal plot. I do not try to explain Lenth's algorithm in a short course. Nor would I want to explain IER vs. EER in a short course.
6. For another example of the myth of objectivity, consider the results of an experiment to see how acceleration times of a vehicle varied with five factors A–E. The design was a 2^5. (Randomization was not done, but that does not affect the point here. Data available on request, for internal use only.) Using $\alpha = 0.10$ with Lenth's method leads to six active effects: in decreasing size these are D, E, ACD, ABE, BD, and ABCDE. The last two effects are dropped if the short-course instructor arbitrarily tells participants to retune to $\alpha = 0.05$, and one more effect is dropped if $\alpha = 0.01$. Which α is objective? None. But at this point, the amateur seeking objectivity will wonder why it seemed OK to use $\alpha = 0.10$ and even a bit larger (per the instructor) in the previous example, but now some other rule is used. By comparison, the half-normal plot in (c) of Figure 1 shows D and E as standing out far from the line, while ACD is a bit off and the remaining three effects noted above now seem totally uninteresting. The nonobjective short-course instructor might say, "Well, if that ACD were a main effect, say A, I would have included it in the model. But, as we discussed before, 3fi's are rare, and it's especially hard to imagine this is active, when almost all the terms 'underneath' it (instructor asks class what these terms would be) are not active. Do you think we should include these extra terms in our model?"

   Aside from sampling variability, what kind of data would lead to so many effects declared significant? Well, it appears that there was one

outlier in the data and, in this sample, the four higher order interactions declared active at $\alpha = 0.10$ were affected by it to the extent they were declared active. Even using the model based on $\alpha = 0.05$ tends to hide this outlier in a subsequent analysis. However, using only two main effects in the model makes this outlier stand out more clearly. A follow-up analysis with that one point set aside, removing the nonestimable 5fi, and recreating the (now approximately correct) half-normal plot shows that only D is active, with E possibly suggested—see (d) in Figure 1.

7. Some other examples of the usefulness of the half-normal plot appear in Daniel's section "Use of Half-Normal Plots in Criticizing Data".
8. Finally, it is worth repeating Daniel's quote (p. 338, section on "Failures of Half-Normal Plotting"): "It is unnecessary to warn experienced statisticians that the use of half-normal plots suggested here is still full of subjective biases, that it is not offered as a general substitute for the analysis of variance, and that its use in a routine way may be catastrophic. More optimistic and less experienced statisticians may get the impression from the successful examples given that a panacea is being offered that can hardly fail. This is not the case".

Reference

Box, G. E. P. and Draper, N. R. (2007). *Response Surfaces, Mixtures, and Ridge Analyse*, 2nd edition. New York, NY: Wiley.

Discussion

DOUGLAS C. MONTGOMERY

School of Computing, Informatics and Decision Systems Engineering,
Arizona State University, Tempe, AZ 85287

My thanks to Professor Lenth for a clear exposition of the problems associated with the use of normal (and half-normal) probability plots of factor effects in interpreting the results from 2^k factorial and 2^{k-p} fractional factorial designs. I am in agreement with the points that he makes, although I'm not quite ready to throw these plots under the bus. I'll explain why in what follow.

First, let's look at the half-normal plots that Professor Lenth presents in his Figure 1. These plots have the absolute effect value on the vertical axis and the probability (or half-normal quantile scale) on the horizontal axis. This has always seemed to me to be an unnatural way to orient the axes. I remember literally dozens of times in engineering school where we had to construct graphs of some measured quantity and a computed or derived value. We always drew the graph with the measured variable on the horizontal axis. This scale was usually linear, and the computed variable scale often was not. For example, if temperature was a controlled or measured variable and viscosity was the computed or derived variable, then we were instructed to plot viscosity versus temperature. Often the viscosity was actually log viscosity to obtain a more nearly linear relationship. Now I realize that the axes definitions can be arbitrary, but I think that the practice that I describe was drilled into me in a freshmen course on engineering drawing. When I first started using normal probability plots, I had to construct them by hand (I went to engineering school in the BC era, meaning Before Computers). To construct the plots, we typically used preprinted normal probability graph paper (I still remember the K&E variety, white margins, and green grid lines). This paper had the probability scale on the vertical axis. So this just reinforces my natural inclination to want to see the graph with that orientation.

I also prefer the full normal probability plot to the half-normal plot. Because this was the type of graph paper available to me back in the BC era, it's what I learned to do, and I just find it easier to interpret. Another nice thing about the full normal plot is that

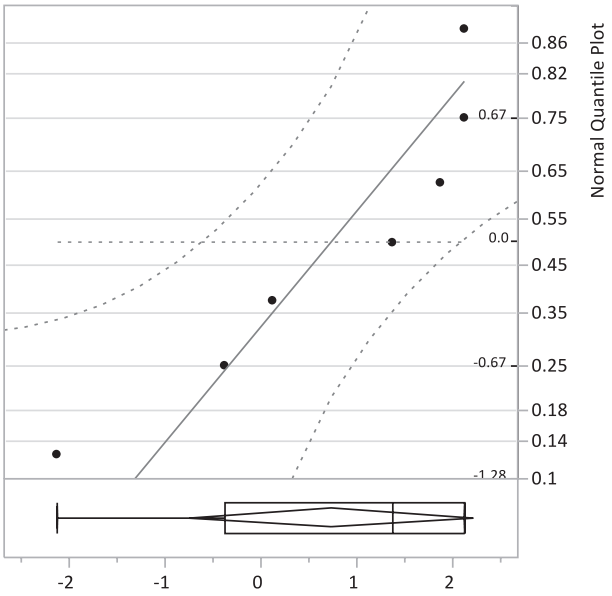


FIGURE 1. Normal Probability Plot and Box Plot from JMP, Lenth's Example 2.

it facilitates adding a dot diagram at the bottom so that you can see the actual relative size of the effects. Professor Lenth observes that simple dot plots may be just as effective as normal probability plots in identifying active effects. I agree. I also always found that combining the two plots was very helpful in interpretation. Sadly, that can't be done directly in modern software. Figure 1 is a normal probability plot from JMP Pro V11 of the effect estimates from Professor Lenth's example 2. JMP's default is to display the normal plot in the vertical orientation with the probability scale on the horizontal axis, but it's easy to change the display to the horizontal orientation that I show here. You can also select either a histogram or a box plot to be displayed along the effect axis. I selected the box plot because the histogram doesn't make sense for seven observations. A dot diagram would be perfect, but that's not an option at present in JMP. I would like to see software display the normal probability plot of effects in the

horizontal orientation accompanied by the dot diagram as the default.

Professor Lenth gives an insightful discussion on drawing the reference line on the plot. I draw the reference line on the half-normal plot emanating from the origin and passing through the 50th percentile of the graph. On the full normal plot, I draw the line by eye between about the 25th percentile and the 75th percentile. It is very helpful when computer software allows you to move the line around by clicking and dragging on the plot. Not all software packages allow you to do that, but they should—it helps in interpretation. Some packages adjust the line as effects are entered and/or removed by clicking on their plotting symbol. I don't care for that, and I think some analysts are potentially confused by this feature. Once I have drawn the line, I tell my students to imagine laying a "fat pencil" along the line—effects covered by the pencil can be assumed to be error and effects not covered are probably active. I know this is pretty subjective, and some of my students laugh at the "fat pencil test", particularly when they remember that I've told them that one of the reasons that we employ statistical methods in experimentation is to achieve scientific objectivity in our conclusions. A few years ago some of my graduate students gave me a fat pencil about 30 inches long and sometimes I take it to class when the normal plots are introduced just to make the point about the degree of subjectivity in what I'm telling them. However, after they see some examples, most students get the concept of using the line to judge outliers and become reasonably adept at using the normal plot.

I am a big fan of Lenth's analytical method for assessing the significance of effects. I always teach this method and, before it was widely available in software,I computed it manually and used it in conjunction with the normal plot. I particularly like the implementation in JMP's Analyze/Modeling/Screening menu, where the t-like ratios are accompanied by individual and simultaneous pseudo-P-values obtained by Monte Carlo simulation. The effects are plotted on a Pareto chart as professor Lenth illustrates in his Figure 5. I could learn to live with the Pareto plot instead of the normal plot, I suppose, but I like them both and don't have any problem showing both displays.

I agree that teaching the technical mechanics of how the plots are constructed can be a challenging task, particularly to nonstatisticians, who are unlikely to be fascinated by the details. My engineering and science students fall into that category and I admit to doing a bit of arm-waving about those details. I focus as much time as I can on how to interpret the plots and provide lots of examples. That's a better use of their time.

I have always found normal plots to be pretty reliable in identifying large and moderately large effects, particularly in 16-run designs. I think they are less reliable in eight-run designs. I always err on the side of selecting an effect to be active in a doubtful case. Type II errors are much more problematic in screening experiments than type I errors. However, I always use Lenth's method and look carefully at the Pareto chart of effects that accompanies it. I feel that having more information doesn't hurt.

Discussion: Better, Not Fewer, Plots

ROBERT MEE

Haslam College of Business, University of Tennessee, Knoxville, TN 37996-0525

In "The Case Against Normal Plots of Effects," Russell Lenth has highlighted the danger of potential misinterpretation. The unaided plots in his Figure 1 are indeed not easily interpreted. But rather than discarding normal plots of effects, I would argue for their enhancement. While I agree that objective test procedures such as Lenth (1987) should be the primary tool, normal plots of effects sometimes help uncover violations of the standard sparsity-of-effects assumption and help convey the risk of making type I errors when basing many tests on individual p-values. Consider the following two examples, which will be familiar to many readers.

Example 1: Engel's (1992) Robust Design Shrinkage Data

Steinberg and Bursztyn (SB) (1994) produced a half-normal plot of the 31 estimates from Engel's (1992) $2^{3-1} \times 2^{7-4}$ product array experiment, where the 31 estimates correspond to the 10 main effects and the 21 clear control-by-noise interactions. SB's Figure 1, re-created here, shows three large estimates, with Lenth's p-values (obtained by simulation) of 0.040, 0.049, and 0.052 for C∗N, A, and E∗N, respectively.

SB note the unexpected shortage of regression coefficients very near zero; while only four estimates are less than 0.05 in absolute value and four more are less than 0.10, 13 are from 0.10 to less than 0.15! As noted by Daniel (1959) and reiterated by SB, when the highest concentration (in the absolute value) of estimates is distanced from zero, the analyst should suspect that one or more outliers are present. SB identify two suspected outliers with the same control factor combination and suggest that these two adjacent values were likely swapped when being recorded. The reversal of these two shrinkage values reduces Lenth's PSE from 0.206 in Figure 1 to 0.033 in Figure 2. In the second analysis, 10 effects have Lenth p-values $<$ 0.05, including three interactions involving the noise factor M (% regrind). If the analyst examined residuals from a reduced model with a few initial effects, the identification of one or two outliers might have been noticed without Figure 1. However, it is also possible that the dearth of significant estimates in the initial (objective) analysis might have discouraged all further investigation and the swap of values gone undiscovered with the half-normal plot.

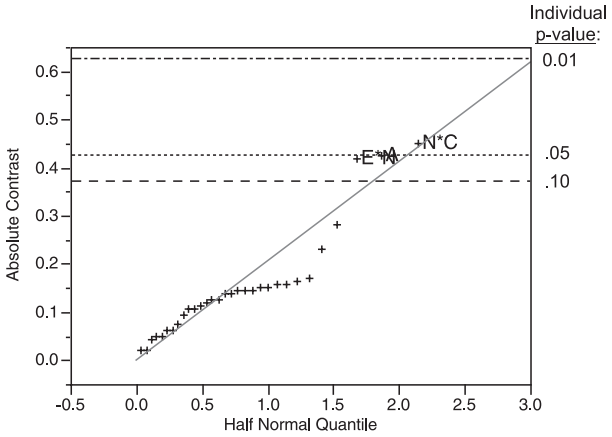


FIGURE 1. Original Half-Normal Effects Plot for Engel (1992).

This author is not aware of an effective test for a mode shifted away from zero in the half-normal plot. A likelihood-ratio test here fails to identify the problem. Thus, while an objective test would be useful,

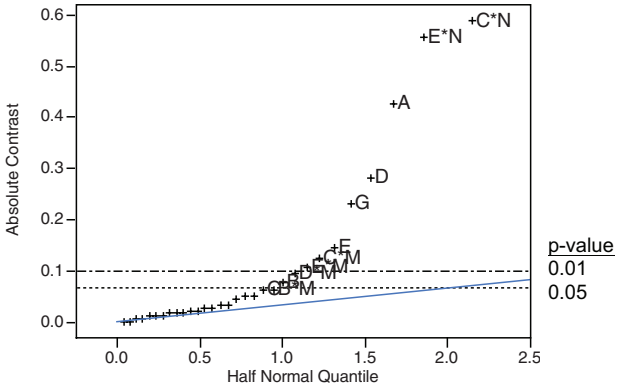


FIGURE 2. Half-Normal Effects Plot for Engel (1992) After Value Swap.

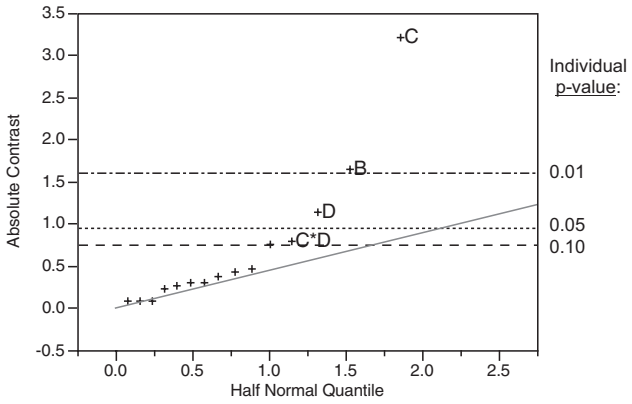


FIGURE 3. Half-Normal Effects Plot for Daniel's (1976) 2^4.

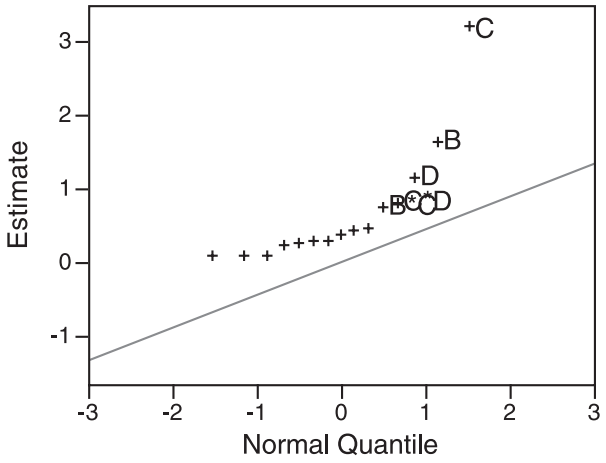


FIGURE 4. Normal Effects Plot for Daniel's (1976) 2^4.

in its absence, the half-normal plot of effects serves as a useful diagnostic. [Note: While a single outlier moves all the estimates by an amount $\pm c$ for a 2^{k-p} design, swapping two values moves exactly half of the regression coefficients by $\pm c$. Thus, having a concentration of values away from zero might be indicative of either an outlier or of the swapping of two entries.]

Example 2: Daniel's (1976) Drill Experiment

Daniel (1976, p. 72) used a 2^4 drill experiment to motivate the search for simple models by transformation of the response. Using Daniel's labels, the half-normal plot (Figure 3) seems typical for a successful 16-run factorial. Lenth's method highlights the presence of three main effects: B, C, and D. However, a normal plot of effects (Figure 4) shows that something is amiss; the small estimates are not scattered about zero but are rather all above 0. Once again, while a residual plot for a suitable reduced model would show unequal variance, the Figure 4 normal plot of effects is a more direct diagnostic. Having all or a preponderance of estimates of the same sign indicates that something systematic is at work, which would not happen under the sparsity-of-effects assumption that underlies the estimation of the standard error of effects here. The Pareto plots of effects shown in Lenth's (2015) Figure 5 would also reveal any preponderance of effects of the same sign, but the line in our Figure 4 draws attention to this departure from expectations. This normal plot is a direct diagnostic because it contrasts what we expect to see under sparsity-of-effects (a mass of estimates about the line) with all the estimates distanced from that line. When this happens, the random error cannot be as large as that reflected by the PSE. Instead, there must be many effects of different sizes. Here, this means the effects for advance rate are not additive; numerous interaction effects are present.

Summary

The purpose of these two examples has been to highlight the advantage of plotting the estimates. Violation of the assumed effect sparsity may show itself in the normal or half-normal plot of effects. Something useful is lost if we do away with these plots. So how can we guard against misinterpretation of these plots by those less acquainted with the theory behind them? I concur that the plots themselves are not sufficient. Lenth's method or some other objective test should be primary for the determination of which effects to include in one's model. And the half-normal plots should reveal more explicitly the results of that test. To that end, Figures 1–3 were enhanced by plotting dashed lines corresponding to individual Lenth p-values = 1%, 5%, and 10%. Even better, the right-hand vertical axis could display the p-value scale. The default might be for individual p-values, but one might also display simultaneous p-values. This simple enhancement would aid all users, both those who are knowledgeable about the construction of normal and half-normal quantile plots and those who are uninformed.

Would such annotated normal plots have utility when the assumption of effect sparsity does hold? Yes, in that they visually communicate the logic behind controlling the experiment-wise risk of type 1 errors. In Figure 1, the normal quantile for the largest estimate is to the right of where the line crosses the

5% significance line. This reflects the notion that, with 31 estimates, even if no effects are active, the expected number of estimates with p-values less than 5% is near to $31(0.05) = 1.55$. When the number of estimates is large, using an individual error rate of 5% means that some type I errors are a likely occurrence.

While type II errors may in fact be more prevalent for small experiments, that is a consequence of the experiment size and/or the magnitude of these real effects. If this concern makes one inclined to allow declaring effects significant using a more lenient type I error rate, then one must surely recognize the need for confirmation experiments. But that is a different matter than the focus of this reply.

As a final technical note, using the normal quantile scores from Mee (2009, p. 45),

$$h_j = \Gamma^{-1}\{(s_j - 0.055)/(m + 0.6)\}$$

produces a closer fit (than Daniel's choice or the scores used by Lenth) for the observed estimates when the true model has no effects.

References

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*. New York, NY:John Wiley & Sons.

Engel, J. (1992). "Modeling Variation in Industrial Experiments". *Applied Statistics* 41, pp. 579–593.

Lenth, R. V. (1989). "Quick and Easy Analysis of Unreplicated Factorials". *Technometrics* 31, pp. 469–473.

Mee, R. W. (2009). *A Comprehensive Guide to Factorial Two-Level Experimentation*. New York, NY: Springer.

Steinberg, D. M. and Bursztyn, D. (1994). "Dispersion Effects in Robust-Design Experiments with Noise Factors". *Journal of Quality Technology* 26, pp. 12–20.

Discussion: On Daniel Plots

DAVID M. STEINBERG

Tel Aviv University, Ramat Aviv 69978, Israel

Russ Lenth advocates that we abandon normal and half-normal plots of effects as a tool for analyzing 2^k and 2^{k-p} experiments. Whether or not we decide to bury these plots, let me begin by praising them.

The half-normal plot for two-level factorial designs was introduced by Daniel (1959) and henceforth, like Lenth, I will use the term *Daniel plots* to refer to both the normal and half-normal versions. Considering the tools available for data analysis in the 1950s, Daniel's ingenuity produced a remarkably successful method with a number of noteworthy properties:

- It is easy to produce.
- It gives a quick visual summary of which effects are the most important.
- Effects can be compared to a reference—the contrasts on a line through the origin—to assess whether or not they stand out from background noise.
- Outliers leave a distinguishing fingerprint and can often be detected.
- The plot may suggest the need for a transformation of the response data.
- The analysis adapts to split-plot experiments and can be used as a diagnostic tool to detect inadvertent split-plotting.
- All the orthogonal contrasts from the design are plotted, so the analysis is not dependent on specifying a particular model.

All the points above (with the exception of the last) were noted and explored by Daniel in his original paper and further elaborated in his book (Daniel (1976)).

More than 50 years have passed since Daniel's paper was published. Modern software offers many analysis and visualization alternatives for factorial experiments. Has the Daniel plot outlived its usefulness? On this basic question, I disagree with Lenth and will explain in what follows why I think that Daniel plots remain useful.

The greatest single advantage of the Daniel plot is its ability to stimulate discussion of the results of an experiment by encapsulating, in a single display, such a variety of information. In a Daniel plot, the experimenter can see how large are the factor effects, if they stand out from noise (including an automatic adjustment for multiple testing) and indicators of problems like outliers or need for transformation.

Lenth recommends the screening summary from JMP©, which includes an effect plot, estimates and p-values using Lenth's method (1989) (see his Figure 5). I agree that this is also a useful tool. Does it provide all the information that is present in the Daniel plot? I find it much more difficult in these regression summaries to make a quick visual comparison of the strength of the effects, to quickly pick out the strongest terms, and to see the relative strength of the effects. Arranging the contrasts in order of magnitude (and not just the main effects) is a helpful option. Even so, the comparisons with noise are conveyed in a separate manner (say by adding p-values); I think the Daniel plot, by placing this information in the same graphic, is more effective.

Much of Lenth's discussion and criticism centers on the ability to judge whether observed effects stand out against noise. The assessment from a Daniel plot, which relates to whether an effect falls "off the line" determined by the small contrasts, is rough and subjective. Although significance tests are important, I think that Lenth gives them excessive weight in his recommendation to drop the Daniel plot. My own experience is that other criteria (such as the importance associated with a given predicted change in mean) are often more crucial than achieving a p-value below a common cut-off, like 0.05. Moreover, the Daniel plot is not meant to be a "stand alone" analysis tool; it can, and should, be complemented by other summaries. My own analyses of choice are Lenth's (1989) test for unreplicated experiments, which is a simple and robust method to produce significance tests and error estimates for factorial designs, or regression analysis for replicated experiments. (Although designed for experiments without replication,

the Daniel plot is still useful when there is replication and can be augmented to include contrasts that reflect pure error.)

Used wisely, Daniel plots should provoke discussion about effects that are at, or just off, the line formed by the nonactive contrasts. I find the Daniel plot especially helpful for identifying such "borderline" effects. When the Daniel plot is the basis for analysis, Lenth argues that these effects may be ignored, but I think this is more likely to happen when there is overemphasis on p-values as summaries. For most experiments, these effects should stimulate discussion, which can be of great value in understanding and exploiting the results of a factorial experiment. This may lead to follow-up experiments to improve precision or, at least, to confirm that predicted gains based on these effects are realized. As noted earlier, the Daniel plot may also point to the presence of outliers, to inadvertent split-plotting, or to the need for a transformation. All of these are valuable by-products of the visual summary provided by the plot.

One of the questions that must be considered in testing for effect significance in factorial designs is whether to adjust for multiple testing. The Daniel plot, by construction, automatically adjusts (though it may need to be redrawn, dropping clearly active effects, to better assess less clear-cut effects). Whether this feature of the Daniel plot is a strength or a weakness is closely related to the analyst's opinion about the importance of making a multiplicity adjustment. Significance tests, like that of Lenth, usually have two versions—one adjusted and the other without adjustment. Lenth presents both test results for the two examples in his article. Multiplicity adjustment is crucial in deciding whether the "edge of the plot" effects are statistically significant in the 16-run experiment. The Daniel plot, with its built-in adjustment, shows these effects as typical of the noise contrasts, as Lenth notes. The multiplicity adjusted p-values for these terms are all greater than 0.6, reaching the same conclusion but more emphatically. This example suggests that the multiplicity adjustment in the Lenth test is much more extreme than the adjustment made by the Daniel plot. For large two-level factorials, Tripolski Kimel et al. (2008) argued that family-wise adjustments (like that in Lenth's test) are overly conservative and showed the benefits of using the false discovery rate (Benjamini and Hochberg (1995)) for this purpose.

Relative effects can be very important in factorial experiments. In the 2^{5-1} experiment shown by Lenth, factor D has a slope of 2.21 and is more than three times the size of the next largest effects, one of them an interaction involving D. The practical implication is that D continues to have a strong positive effect throughout the experimental region. By contrast, the effects of factors A and B may disappear or even reverse direction if the interactions that have borderline statistical significance are real effects.

The advantages of the Daniel plot for these sorts of comparisons are already evident for the 16-run experiment in Lenth's article (compare the visual ease of Figure 1 or 3 there with Figure 5, even with the Pareto principle used to order the main effects). For larger experiments (with 32 or more runs), the advantages are more pronounced.

The second example presented by Lenth is a 2^{4-1} experiment. Experiments with eight runs can be very informative (see Box (1992)), but any formal analysis is a risky endeavour. Note that Daniel (1959) analyzed only experiments with 16 or more runs and Daniel plots appear in many standard texts only for these larger experiments. (See, for example, Box et al. (2005) or Wu and Hamada (2009).)

Perhaps the main reason that Lenth includes the eight-run example is to point out that deviations on a Daniel plot may be in the "wrong" direction, with the observed effect smaller than would be predicted based on the plot. This can happen and, yes, practitioners using the plot should be cautioned about what it means. However, the plots are best used in larger experiments where it becomes less likely to see this feature.

How will the data from a two-level factorial be analyzed if the Daniel plot is not used? For practitioners with dedicated software like the JMP© platform, many good alternatives are readily available. I am a strong proponent of buying and using the right software. But many companies don't (yet) have it and analysis may rely on what can be done easily with a standard, all-purpose statistical package or even with EXCEL©. Most likely, the main statistical tool will then be regression analysis and the experimenter will need to declare a regression model. If there is no replication, the analysis will depend heavily on the choice of terms in the model. An important effect that is left out may never be detected and will inflate the error estimates. Including too many effects may leave insufficient degrees of freedom for evaluating error. The Daniel plot includes by default all the estimable contrasts in the data and so has the major

benefit of robustness to model choice. (By the way, it is easy to produce Daniel plots with EXCEL©, making it available even to those who lack statistical software.)

Lenth's final argument against the Daniel plot is that it is difficult to teach and the time needed to explain it could be devoted to other topics. Lenth is correct that some effort is needed to explain the plots and still more to achieve confidence in interpreting them. A simple effect plot (like Figure 5 in the paper) is easier to grasp. For practitioners who will not analyze many factorial experiments, the effect plot might be a better alternative. However, as noted above, the choice of techniques we should teach to practitioners must also take into account the software that is available to them for analysis.

For all that it offers, the Daniel plot does not answer every question that we would like to answer from a two-level factorial design and the answers are not always clear cut. Daniel was well aware of the limitations and it is worth recalling his words of caution from the original paper (page 338): "It is unnecessary to warn experienced statisticians that the use of half-normal plots suggested here is still full of subjective biases, that it is not offered as a general substitute for the analysis of variance, and that its use in a routine way may be catastrophic. More optimistic and less experienced statisticians may get the impression from the successful examples given that a panacea is being offered that can hardly fail. This is not the case." More recent analysis techniques are a welcome complement.

I will close with an example. Figure 1 presents a Daniel plot from a 2^{6-1} experiment. The data are confidential and the effects have been scaled so that the largest equals one. Most of the 31 contrasts fall on a well-defined line rooted at the origin. The two largest contrasts, corresponding to factors D and E, are off the line. The Lenth test confirms that they are statistically significant, with p-values (not adjusted for multiplicity) of < 0.001 and 0.003, respectively. The third largest contrast is slightly off the line and has a Lenth p-value of 0.03. Moreover, it corresponds to the DE interaction. Given this connection to the two dominant main effects, I believe that most experimental teams would agree that it is important, even if the p-value had been slightly above 0.05. What about the next group of five points, which are slightly removed from the line through the origin? The Daniel plot should encourage experimenters to ask precisely that question. Which effects are involved? Are they

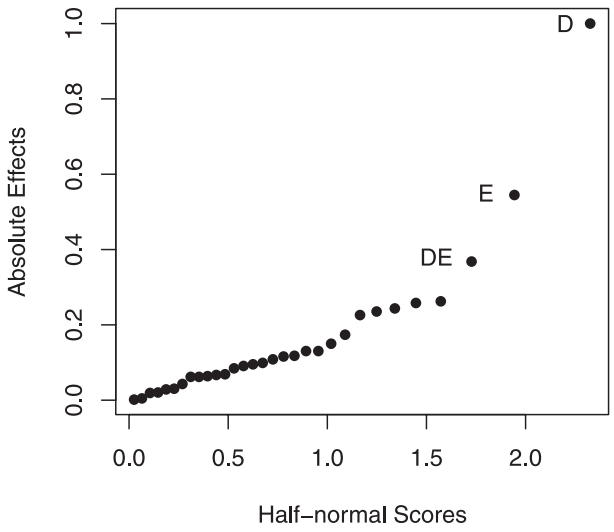


FIGURE 1. Daniel Plot for a 2^{6-1} Experiment.

"real" effects or noise? Could they be important in reaching conclusions or decisions from the experiment? Should they be highlighted in follow-up experiments? Lenth's test is also helpful; the five effects have p-values ranging from 0.08 to 0.15. Qualitative analysis sheds further light; the two largest contrasts both correspond to a pair of three-factor interactions and, in both pairs, factor D is in one of the interactions but factor E is in the other interaction. Thus, neither contrast is an interaction of a third factor with D and E. Discussion with the experimenters led to the conclusion that these terms, and the subsequent contrasts, should be treated as error.

In conclusion, what I like about the Daniel plot is that it presents so much information, related to different aspects of the data analysis. No other single summary touches on such a variety of important features. The conversation surrounding the plot above was a valuable part of the analysis. I think the Daniel plot remains a very useful summary of the experimental data and a great stimulus for discussing the results.

References

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Journal of Royal Statistical Society, Series B* 57, pp. 289–300.

Box, G. E. P. (1992). "What Can You Find Out from Eight Experimental Runs?" *Quality Engineering* 4, pp. 619–627.

Box, G. E .P.; Hunter, J. S.; and Hunter, W. G. (2005).

Statistics for Experimenters: Design, Innovation and Discovery, 2nd edition. Hoboken, NJ: John Wiley & Sons.

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*. New York, NY: John Wiley & Sons.

Tripolski Kimel, M.; Benjamini, Y.; and Steinberg, D. M. (2008). "The False Discovery Rate for Multiple Testing in Factorial Experiments". *Technometrics* 50, pp. 32–39.

Wu, C. F. J. and Hamada, M. S. (2009). *Experiments: Planning, Analysis, and Optimization*, 2nd edition. New York, NY: John Wiley & Sons.

Discussion

ERIC D. SCHOEN

University of Antwerp, Belgium, and TNO, Zeist, Netherlands

PETER GOOS

University of Antwerp and University of Leuven, Belgium

1. Introduction

Lenth pleads against the use of normal or half-normal plots of effects from two-level experiments to judge which effects may be active. Instead, he recommends utilizing Pareto charts of the effects supplemented with cut-off lines based on a robust estimator of the standard error.

Our discussion of Lenth's paper consists of two parts. In the first part, we consider judging effects from full factorial or regular fractional factorial designs by normal or half-normal probability plots or robust estimators of the standard error. We recognize that the plots shown in Lenth's paper are appealing, but we do not think that half-normal plotting should be entirely abandoned. In the second part of our discussion, we consider the case of nonregular designs or optimal designs, where robust standard errors and half-normal plotting of effects are of limited use and other methods of analysis need to be utilized.

2. Regular Designs

2.1. Some History

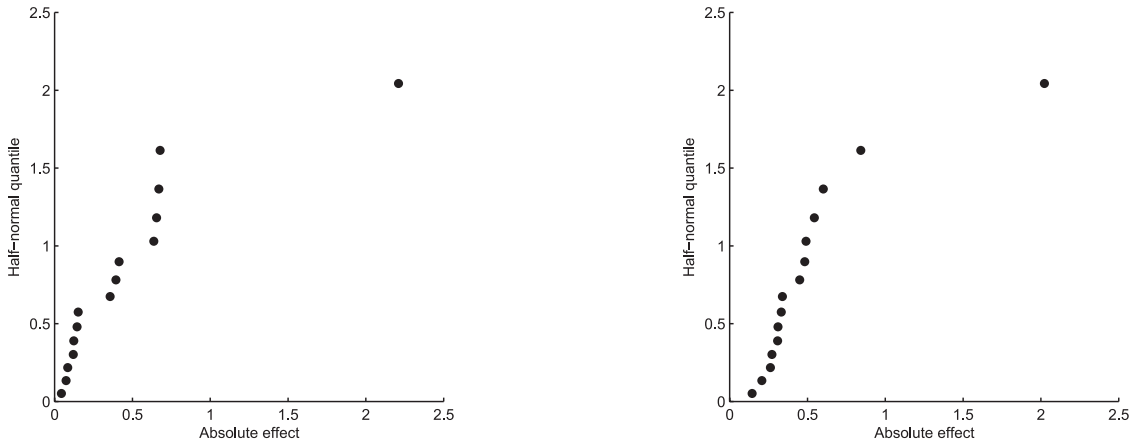
Half-normal plots were introduced by Daniel (1959), who observed that absolute-valued null effects ordered from small to large tend to be on a straight line through the origin when plotted on what he called 'half-normal probability paper'. At that time, blocks of specially prepared sheets of *normal* probability paper could be purchased, where the vertical axis was used for the quantiles. By analogy, Daniel's half-normal plots had the quantiles on the vertical axis and the effect sizes on the horizontal axis. As an example, Figure 1(a) shows the Daniel plot of effects based on the data of the regular 2^{5-1} fractional factorial design from Snee (1985), as reproduced in Lenth's paper.

Daniel (1959) also observed that half-normal plots offer a way to detect one or two defective values, inadvertent plot-splitting, or an anti-lognormal distribution of error. In this discussion, we restrict attention to the detection of a single defective value. For this purpose, we changed the 13th response value of Snee's data from 5.22 to 2.22. Figure 1(b) shows the half-normal plot of effects based on the modified data. Because one response value is three units too small, the effects now are biased by either $-3/16$ or $3/16$, depending on the sign of the contrast corresponding to the defective value. Therefore, in absolute value, the null effects now no longer lie on a straight line through the origin, but on a straight line through the point $(3/16, 0)$. This illustrates that half-normal plots can draw our attention to possible outliers in the data. In any case, Daniel (1959) showed that there may be more than one reason to study half-normal plots. Lenth's present article only deals with the best-known one, which is the detection of active effects.

Daniel (1976, Section 7.6) observed that some peculiarities of the data were not reflected in the half-normal plots. For this reason, he proposed to study signed effects without even using half-normal or normal plots. He also studies normal plots in the book, but only for signed residuals. In the book, no normal or half-normal plots of factor effects were made. Indeed, normal plotting only became fashionable due to the influential book by Box et al. (1978). Thus, the appropriate reference for normal instead of half-normal plotting of effects may well be Box et al. (1978), rather than the work of Daniel.

2.2. Some Technical Issues

Loh (1992) showed that the appearance of the points in normal plots can be entirely changed if the arbitrary $-$ and $+$ labels of the factor levels are swapped. For the example from Snee (1985), which involves five factors, 2^5 different normal effect plots



(a) Data from Snee (1985)

(b) Data with 13th response changed from 5.22 to 2.22

FIGURE 1. Half-Normal Plotting As a Way of Detecting an Outlier.

can be created by swapping the labels of the levels of one or more factors. Some of the possible plots are much easier to judge by eye than others. This problem is absent in half-normal plots. Wu and Hamada (2009) use the same kind of reasoning to prefer half-normal plots. Mee (2009) points out that one reason to prefer these plots might be the fact that the statistical significance of an effect is generally based on the size of its absolute value. Thus, we prefer half-normal plots over normal plots to judge effects from unreplicated experiments.

Of course, when using half-normal plots, there is the problem of subjectivity in drawing the line that marks the inactive effects. Lenth (1989) has written an immensely popular and useful paper on making a more objective evaluation of effects from unreplicated two-level experiments. The effects are evaluated using a pseudo-standard error based on the set of absolute values of the effects. The pseudo-standard error is calculated in three stages. First, the median of the full set of absolute-valued effects is multiplied with a consistency constant, so that an initial standard error estimate becomes available. Effects that are large with respect to this estimate are considered too big to provide information about the error variance. Therefore, in the second stage, these effects are removed from the set. Finally, the median of the remaining effects' absolute values is multiplied by another consistency constant to result in the pseudo-standard error, which we denote by PSE_{50}. The subscript 50 refers to the 50th percentile, which is the same as the median.

In a follow-up article, Haaland and O'Connell (1995) provide an overview of various alternatives to Lenth's approach. One alternative to the PSE_{50} is the ASE (Dong (1993)), where the third stage for the pseudo-standard error calculation is based on the root-mean square of the remaining effects. Another alternative is to use the PSE_{45}, meaning that the first stage is based on the 45th percentile of the absolute values of the effects. Haaland and O'Connell (1995) ended up recommending the ASE, PSE_{50}, and PSE_{45} for practical use, where the choice depends on *a priori* expectations concerning the number of active effects. Schoen and Kaul (2009) provided extensive tables for consistency constants and critical values based on these pseudo-standard error approaches. In recent years, many other authors have published alternative approaches to analyze data from regular fractional factorial designs. We would welcome a review article discussing the pros and cons of the existing methods, possibly including a simulation study comparing the most attractive approaches.

The evaluation of effects both with a half-normal plot and with a robust standard error require that effect sparsity holds. For a half-normal plot, effect sparsity is required to identify a set of null effects pointing to the origin of the plot. For an evaluation with a robust standard error estimate, the effect sparsity is even more important, because the initial stage of constructing this estimate involves the median or the 45th percentile of the absolute values of the effects. If too many effects are active, the initial error estimate is inflated. This inflation, in turn, might lead to an inflated final estimate. If the significance of effects is evaluated using an inflated standard er-

ror estimate, the effects may erroneously be classified as inactive.

The Pareto plots that Lenth advocates instead of the probability plots do not provide checks on effect sparsity. Much as we like these Pareto plots, we think that the half-normal plots should be provided as well, both to check the sparsity assumption and to identify possible outliers in the data and inadvertent split-plotting. We believe that half-normal plots may still work well when less than 50% of the effects are inactive, provided there is a sufficiently large total number of effects in the plots. The Pareto plot approach involving PSEs, however, should be adapted to continue to work if the effect-sparsity condition is violated. In that case, the PSE should be based on a smaller quantile than the one used by default.

Additional conditions that need to be fulfilled for a valid evaluation with a half-normal plot as well as with a robust standard error estimate are that the estimates of the factor effects should be independent and normally distributed and that they should possess the same standard errors. Many scenarios exist in which at least one of these conditions does not hold. One such scenario is when a split-plot design, a strip-plot design, or another design with restricted randomization is used, either inadvertently or intentionally.

2.3. Restricted Randomization

Many industrial experiments are not completely randomized. This may be due to the presence of hard-to-change factors or due to the fact that the experiments span multiple steps of a process; see Goos and Jones (2011, Section 10.3.4) for a discussion of the different disguises of a split-plot design. Depending on the exact two-level design utilized, half-normal plotting or evaluation with a robust standard error may or may not work well.

For a regular split-plot design, half-normal plotting or evaluation with a PSE may work if all effects are either estimated in the whole-plot stratum or in the subplot stratum. In that case, there are two sets of effect estimates: one with a larger standard error and one with a smaller standard error. For each set of effects, a different half-normal plot (Bisgaard (2000)) or a different PSE is required. The main problem is that, oftentimes, one of these sets contains only a limited number of effects. For that set of effects, a PSE-based test will not be powerful and half-normal plots will not permit separation of active and inactive effects. Schoen (1999) states that at least seven effect estimates should appear in a half-normal plot for it to be useful.

For a regular strip-plot design, half-normal plotting or evaluation with a PSE may work if all effects are either estimated in the row stratum, the column stratum, or the run stratum. In that case, there are three sets of effect estimates, one for each of these strata. Again, for each set of effects, a different half-normal plot or PSE is required. Here too, the main problem is that, generally, at least one of these sets contains only a limited number of effects, which renders half-normal plotting or PSE-based evaluation of the corresponding effects problematic.

By revisiting a regular strip-plot example of Vivacqua and Bisgaard (2004), Arnouts et al. (2010) demonstrate that a generalized least-squares analysis may result in the detection of a larger number of active effects than an analysis relying on normal plots. This example is one where there are very few effects in the column stratum.

In case the split-plot, strip-plot, or related design is based on a nonregular fractional factorial design or an optimal design, half-normal plots or PSE-based evaluations become virtually useless. To a large extent, this is also true for completely randomized nonregular fractional factorial designs.

3. Nonregular and Optimal Designs

The best-known examples of nonregular fractional factorial designs are Plackett–Burman (PB) designs. Plackett–Burman designs, as well as other nonregular designs, are increasingly utilized for experimentation. There are two main reasons for this. First, nonregular designs offer more flexibility in terms of run size, as they exist whenever the number of runs is a multiple of four, whereas regular fractional factorial designs only exist when the number of runs is a power of two. Second, it is now recognized that nonregular designs allow a more diverse set of regression models to be estimated than regular fractional factorial designs. A disadvantage is the fact that the analysis of these designs is more complicated (i.e., a careful linear regression analysis is required), so that, in our consulting experience, certain experimenters feel more comfortable using regular fractional factorial designs. Also, the estimable effects in a regular design have the minimum possible standard error. We therefore think that both regular and nonregular designs have merit, but we would like to point

out that the advantages of using nonregular designs, in our view, outweigh the disadvantage of a more complicated analysis. We would rather be able to estimate more interaction effects (with a nonregular design) than fewer interaction effects (with a regular design). The complete aliasing of interactions in regular fractional factorial designs will often necessitate follow-up experiments to disentangle aliased effects, while this will often not be the case for nonregular designs.

3.1. Complete Randomization

When using nonregular fractional factorial designs, the estimates of the main effects and the estimable two-factor interactions often do not have the same standard errors, even when the design is completely randomized. Also, the estimates are generally not all independent. This is due to the fact that the two-factor interaction effects are often partially aliased with the main effects and/or with other two-factor interaction effects. Consequently, the significance of the main effect estimates and the two-factor interaction effect estimates cannot be evaluated in a half-normal probability plot. For the same reason, the PSE approach of Lenth cannot be used.

One special case where half-normal plots and PSE-based evaluation may remain useful is in the analysis of data from resolution-IV nonregular designs where the main effects are independent of the two-factor interactions. The main effects can then be evaluated with half-normal plots or using PSEs, while interactions can be explored using regression techniques (Miller and Sitter (2001)).

Increasingly so, optimal experimental designs are considered as alternatives to traditional regular fractional factorial designs, due to the presence of constraints on the factor levels (i.e., due to the fact that certain combinations of factor levels are disallowed) or due to budget constraints. Both types of constraints often cause the experimental design to be nonorthogonal, as a result of which the estimates of the main effects and the two-factor interactions are dependent and do not have the same standard errors. In those cases, a half-normal plot and Lenth's approach are again not appropriate.

3.2. Restricted Randomization

The combination of nonregular or optimal experimental designs, on the one hand, and restricted randomization, on the other hand, does not have a positive impact on the usefulness of half-normal plots and Lenth's PSE. Frequently, factor effects are estimated using information from more than one stratum in a split-plot design, a strip-plot design, or a related design. This causes the factor effects to have different variances and results in a violation of an important condition for half-normal plots and pseudo-standard errors to be applicable.

4. Conclusion

Half-normal plots have proven to be very useful in an era where no fast computers were available to set up experiments and to analyze the resulting data. The lack of computing power has for a long time stimulated researchers to pay attention to ease of computation when proposing design-of-experiments methodology and methods of analysis. Unavoidably, this has led experimenters to simplify their problem so that it matched the available experimental designs and the available methods of analysis. This approach led to many successful applications of full factorial and regular fractional factorial two-level designs. The approach was successful in the sense that it led to good solutions for a simplified problem.

In the past two decades, things have changed dramatically because it has become possible to analyze more complicated data sets properly by means of powerful statistical software and computers. It has also become possible to create tailor-made experimental designs for virtually any practical problem. In other words, there is no need any more to simplify the problem at hand to fit a given design and analysis method. We view half-normal probability plot and PSE-based significance testing as useful for a specific type of experimental designs, i.e., full factorial and regular fractional factorial designs. In many practical scenarios, however, these designs cannot be used or can be improved upon.

In this contribution, we mentioned that the data analysis for nonregular designs is more complicated than that for regular designs. In doing so, we do not intend to say that one needs a degree in statistics to analyze the data from nonregular designs. All that is required to get most of the salient insights out of a data set is a sound knowledge of linear regression analysis, including concepts such as stepwise regression, some goodness-of-fit diagnostics, and multicollinearity. Nowadays, plenty of user-friendly software is available to perform linear regression analyses for any experimental design available.

Of course, it will always remain difficult to iden-

tify the best model from small experiments with numerous potential effects, independent of the design used. Therefore, we welcome more research on the analysis of data from nonregular two-level designs, in the spirit of Wolters and Bingham (2011) and Mee (2013), and on the implementations of these methods in user-friendly software. This will enable experimenters to combine the best designs available with the best methods of analysis available, so that their original problems rather than simplified ones can be solved in an optimal way.

References

Arnouts, H.; Goos, P.; and Jones, B. (2010). "Design and Analysis of Industrial Strip-Plot Experiments". *Quality and Reliability Engineering International* 26, pp. 127–136.

Bisgaard, S. (2000). "The Design and Analysis of $2^{k-p} \times 2^{q-r}$ Split Plot Experiments". *Journal of Quality Technology* 32, pp. 39–56.

Box, G. E. P.; Hunter, W. G.; and Hunter, J. S. (1978). *Statistics for Experimenters*. New York, NY: Wiley.

Daniel, C. (1959). "Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments". *Technometrics* 1, pp. 311–341.

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*. New York, NY: Wiley.

Dong, F. (1993). "On the Identification of Active Contrasts in Unreplicated Fractional Factorials". *Statistica Sinica* 3, pp. 209–217.

Goos, P. and Jones, B. (2011). *Optimal Design of Experiments: A Case Study Approach*. New York, NY: Wiley.

Haaland, P. D. and O'Connell, M. A. (1995). "Inference for Contrast-Saturated Fractional Factorials. *Technometrics* 37, pp. 82–93.

Lenth, R. V. (1989). "Quick and Easy Analysis of Unreplicated Factorials". *Technometrics* 31, pp. 469–473.

Loh, W. Y. (1992). "Identification of Contrasts in Unreplicated Factorial Experiments". *Computational Statistics and Data Analysis* 14, pp. 135–148.

Mee, R. W. (2009). *A Comprehensive Guide to Factorial Two-Level Experimentation*. New York, NY: Springer-Verlag.

Mee, R. W. (2013). "Tips for Analyzing Nonregular Fractional Factorial Experiments". *Journal of Quality Technology* 45, pp. 330–349.

Miller, A. and Sitter, R. R. (2001). "Using the Folded-Over 12-Run Plackett–Burman Design to Consider Interactions". *Technometrics* 43, pp. 44–55.

Schoen, E. D. (1999). "Designing Fractional Two-Level Experiments with Nested Error Structures. *Journal of Applied Statistics* 26, pp. 495–508.

Schoen, E. D. and Kaul, E. A. A. (2000). "Three Robust Scale Estimators to Judge Unreplicated Experiments". *Journal of Quality Technology* 32, pp. 276–283.

Snee, R. D. (1985). "Experimenting with a Large Number of Variables". In: Snee, R. D.; Hare, L. B.; and Trout, J. B., eds., *Experiments in Industry: Design, Analysis and Interpretation of Results*. Milwaukee, WI: Quality Press.

Vivacqua, C. and Bisgaard, S. (2004). "Strip-Block Experiments for Process Improvement and Robustness". *Quality Engineering* 16, pp. 495–500.

Wolters, M. A. and Bingham, D. R. (2011). "Simulated Annealing Model Search for Subset Selection in Screening Experiments". *Technometrics* 53, pp. 225–237.

Wu, C. F. J. and Hamada, M. S. (2009). *Experiments: Planning, Analysis, and Parameter Design Optimization*, 2nd edition. New York, NY: Wiley.

Discussion

R. DENNIS COOK

School of Statistics, University of Minnesota, Minneapolis, MN 55455

CHRISTOPHER J. NACHTSHEIM

Carlson School of Management, University of Minnesota, Minneapolis, MN 55455

Lenth (2014) has it right: It's time to retire Daniel plots in the analysis of unreplicated 2^k or 2^{k-p} designs. Interpreting these plots is fraught with difficulty for both the statistician and nonstatistician, and there are better alternatives.

Among Lenth's many compelling points were the following:

1. Normal plots can be completely misleading. Example 2, in which three effects are identically 2.125, demonstrates this convincingly.
2. The placement of the reference lines is arbitrary and potentially misleading.
3. It matters whether potentially active effects deviate horizontally or vertically, a point that is easily missed.
4. Correct interpretation of half-normal plots depends on the choice of axes.
5. De Leòn et al. (2011) demonstrate that normal plots are no more effective than dot plots.
6. Magnitudes of effects are difficult to judge because they are displayed along a "crooked" path. Pareto plots, in contrast, display magnitudes appropriately in a side-by-side fashion.

To these points we would add the following. A plethora of applicable model-selection techniques has been developed since Daniel first introduced the normal and half-normal plots of effects in the 1950s. Some popular alternatives include stepwise regression, lars-lasso, and the Dantzig selector. See, for example, Draguljić et al. (2014) for a recent assessment of these techniques for the analysis of interactions. Can these methods contribute effectively to the analysis of saturated experiments? Providing a general purpose (OK, cookbook) approach to model selection would make the analysis a bit less daunting for the nonstatistician.

Given that we've chosen to abandon the Daniel plot, what alternatives are suggested? Simple dot plots and/or Pareto plots of estimated effects offer advantages, but judgment as to statistical significance is still subjective. Lenth's (1989) method, with effects displayed in a Pareto plot with a superimposed reference line and simulated P-values (as in the JMP implementation) is a plausible alternative. But Lenth's method has limitations as well. The breakdown point occurs whenever more than half of the effects are active and the method relies on normality, effect-sparsity, and standard errors. Randomization analysis is an alternative that seems to get little attention, and which offers the potential for robustness to the level of sparsity and normality. Loughlin and Noble (1997) provided methodology for implementing a randomization analysis for unreplicated two-level factorial and fractional factorial designs.

For illustration, we implemented five alternatives to the Daniel plot using five previously published data sets and two simulated data sets. The five methods are (i) Lenth's (1989) method, (ii) randomization analysis as described by Loughlin and Noble (1987), (iii) forward stepwise selection based on the AICc criterion (Hurvich and Tsai (1989)) as implemented in JMP, (iv) the Dantzig selector with selection based on AICc, and (v) the Lars/Lasso method using AICc for selection. Data sets analyzed were:

1. Example 1 from Lenth (2014), taken from Montgomery (2013). $n = 16$.
2. Example 2 from Lenth (2014), taken from Montgomery (2013). $n = 8$.
3. Example II from Lenth (1989), taken from Box and Meyer (1985). $n = 16$.
4. Example IV from Lenth (1989), taken from Box and Meyer (1985). $n = 16$.
5. Example of Table 4 from Loughlin and Noble (1997). This was adapted from problem 9.11 of

TABLE 1. Results of Five Methods for Five Previously Published Data Sets. The table displays the number of active effects identified by each method at the $\alpha = 0.05$ level of significance

Method	Lenth	Randomization	Stepwise AICc	Dantzig AICc	LarsLasso AICc
Lenth example 1	1	1	5	5	5
Lenth example 2	0	0	0	0	0
Lenth example II	2	2	2	2	2
Lenth example IV	2	0	5	5	5
Loughlin and Noble example	1	9	3	3	3
Null log-normal example	2	0	2	2	2
Null exponential example	2	0	5	5	5

Montgomery (1991), which gave two replicates of a 2^4 factorial design. The response used is the sum of the two responses for each design point after removing a replicate (block) effect. $n=16$.

6. Null log-normal case: A 2^4 factorial, where the response vector consists of 16 random draws from the log-normal distribution of $X = \exp(Z)$ where Z is normal with mean $\mu = 1$ and standard deviation $\sigma = 1$. Responses (in standard order) are: 0.762, 0.13, 0.805, 2.059, 2.233, 0.2, 1.812, 3.726, 0.882, 1.631, 0.61, 4.601, 3.074, 0.387, 2.347, 6.439.
7. Null exponential case: A 2^4 factorial, where the response vector consists of 16 random draws from an exponential distribution with mean $\mu = 1$. Responses (in standard order) are: 0.543, 2.51, 5.447, 0.874, 6.043, 0.206, 4.171, 1.17, 3.41, 0.667, 3.032, 3.182, 0.456, 4.425, 0.096, 2.211.

Results are summarized in Table 1. Differences of note arise among the methods for example 1, Lenth example IV, the Loughlin and Noble example, and the two null examples. Any of the methods using AICc are liberal in these cases and it is interesting to compare the more similar results for Lenth's method and the randomization analysis. For the Loughlin and Noble example, the randomization tests identify nine active effects, whereas the Lenth method identifies two. As noted by Loughlin and Noble, when the data are analyzed as two separate replicates, these nine effects are identified as active. This is a case where the Lenth method is at a clear disadvantage because the number of active effects is greater than $n/2$. For both null cases examined, the Lenth method identifies two active effects, while the randomization analysis correctly concludes that no effects are active.

These results are illustrative of the kinds of differences that can arise among objective methods and are not meant to demonstrate the superiority or inferiority of any particular method. But they do demonstrate that applying a battery of methods may not lead to clarity, establish a *prima facie* case for randomization tests, and urge additional study. In particular, we think that the randomization approach deserves further consideration. The results also raise a number of questions.

Is effect sparsity a proper foundation for a general purpose method? While there are contexts where sparsity is a driving concept, some seem to view sparsity as akin to a natural law: if you are faced with many factors, then naturally the effects must be sparse. Others have seen sparsity as the only recourse. In the logic of Friedman et al. (2004), the bet-on-sparsity principle arose because, to continue the metaphor, there is otherwise little chance of a reasonable pay-off. Is it important to allow for an abundance of effects?

How important is robustness? A single outlier can have a dramatic impact on the analysis of a saturated or nearly saturated experiment, and it is generally recognized that analyses based on penalization are sensitive to anomalies in the data. Similar remarks apply to regressions with heavy tails.

How important is fidelity to error rates in a general-purpose method for saturated or nearly saturated experiments? If saturated designs are used primarily for screening, followed by confirmatory experimentation, then perhaps control of the false-negative

rate is paramount. Methods are not equally amenable to error-rate control. If careful control is important, then some method will necessarily be favored on this basis alone.

To sum up, we agree with Lenth's call for the retirement of the Daniel plot, and we see his paper as a welcome and perhaps overdue contribution to the literature.

Additional References

Dragulјić, D.; Woods, D. C.; Dean, A. M.; Lewis, S. M.; and Vine, A. E. (2014). "Screening Strategies in the Presence of Interactions (with Discussion)". *Technometrics* 56, pp. 1–28.

Friedman, J.; Hastie, T.; Rosset, R.; Tibshirani, R.; and Zhu, J. (2004). "Discussion of Boosting Papers." *Annals of Statistics* 32, pp. 85–134.

Hurvich, C. M. and Tsai, C.-L. (1989). "Regression and Time Series Model Selection in Small Samples." *Biometrika* 76, pp. 297–307.

Loughin, T. M. and Noble, W. (1997). "A Permutation Test for Effects in an Unreplicated Factorial Design." *Technometrics* 39, pp. 180–190.

Rejoinder

RUSSELL V. LENTH

University of Iowa, 241 Schaeffer Hall, Iowa City, IA 52242

It is a real gift to receive the comments from so many distinguished researchers and practitioners. So first, a sincere thank you to all discussants. I will respond to the main points, as I see them, on a per-topic basis, and follow-up with a couple of closing recommendations for different audiences.

Responses

I value all of these discussants' opinions, and indeed I have learned a lot from them. I even find some measure of agreement with many of their comments, including those with opposing views. I do not attempt to respond to all the points made, or even to acknowledge all of the thought-provoking ideas that are raised. Here are some of the main areas of discussion.

Normal Paper, Axis Orientation

The historical use of graph paper with a normal scale is the reason to plot the effects on the horizontal scale. But normal plots are also commonly used in residual diagnostics, where we plot residuals (vertically) against predicted values, time order, adjusted or unadjusted predictor values, and normal scores. So there are good reasons behind both orientations.

Daniel's Own Warning

That is something I realize I should have quoted myself. I particularly emphasize his point that the routine use of Daniel plots may be catastrophic. Yet the most routinely used tool in analyzing screening experiments is surely the Daniel plot. (Speaking of quotes, thanks to the same discussants for adapting (in the same way) the same passage from *Julius Caesar*.)

Eight-Run Experiments Are Problematic

This is a good point, but I have seen a lot of articles and texts with Daniel plots of seven effects. The fact that my example 2 is only eight runs does not alter the important points related to tied effects and misidentified-as-active effects.

Experts, Amateurs

I could have been clearer that my main gripe with Daniel plots is with teaching them to nonexpert audiences, usually as the primary tool in assessing whether effects are active. It's interesting that Schoen and Goos say that this is primarily due to the influence of Box et al. (2005) (since its first edition)—I'm not sure of that. Voelkel mentions that pseudo standard errors (PSEs) and Pareto plots are suited for amateurs while Daniel plots are suited for experts; and I agree: obviously, expert data analysts should do whatever they want, and if it includes Daniel plots, that's fine. Easy for nonexperts to do a Daniel plot in Excel? Maybe with a macro or add-in.

Nonregular Designs

These are important; and a key issue is that neither Daniel plots nor PSEs work in such cases, as originally conceived. (However, both procedures may be adapted by orthonormalizing the predictors so as to obtain independent contrasts having equal variance. The results depend on the order in which predictors are entered, but if those with the largest absolute effects are entered first, that helps.)

Split Plots

Daniel plots have limited use for a split-plot design, especially for the whole-plot effects. The absolutely best diagnostic for inadvertent split plotting is asking a lot of questions about how the experiment was conducted. Same with Voelkel's example of center-point replicates not representing the true variability due to improper randomization. A dataset consists of numbers and a story. If you get the story wrong, no methodology can do much to save you.

Versatility of Daniel Plots, Unique Diagnostic Capabilities

To me, this is highly related to the different-audience issue. In general, where discussants express favor for Daniel plots, there is a lot of emphasis on their diagnostic features—and, I'd say, some of the

examples are pretty artful and nuanced. These points are not likely to capture the interest or understanding of a short-course student, so this is all in experts' territory. I do not see as many pro–Daniel-plot arguments that emphasize their ability to identify active effects.

For most statisticians, the main use for normal plots is for residual diagnostics. In contrast, a Daniel plot is primarily for identifying active effects, and diagnostic use of them is not of the same character as in residual normal plots. Experts can understand all this, but even students with a few courses under their belts will struggle understanding these two very different uses of the same plot.

Normal vs. Half-Normal

For identifying active effects, I think the consensus is mostly for half-normal plots; but, like Montgomery and Mee, I also think something is lost, and half-normal plots are less likely to be available in software.

Enhancements, Annotations

In some ways, I like Mee's idea of adding cut-off lines to a Daniel plot. On the other hand, it creates some clutter as well. Is not a half-normal plot with cut-off lines in essence a Pareto chart with unequally spaced bars? A well-constructed graph does make a huge difference. Also, in Montgomery's example, I strongly concur that a dot plot would be a more desirable "rug" than the box plot.

Objectivity, Tuning

All procedures are indeed tunable. This is primarily for experts to do. Maybe objectivity is a myth, but there is a distinction between giving a cut-off for making a decision and leaving it entirely to judgment of what is an outlier. Montgomery's "fat pencil" idea provides some guidance, I suppose. One could argue that it is a good idea to give nonexperts only a vague method rather than a cut-and-dried one. It may discourage them from making overly bold statements. Well, actually, I doubt that.

How My Story Changes

The discussions have gotten me to think about some of the issues differently. And these thoughts are organized around the user/consumer. On the one hand, we often teach two-level experiments in short courses directed to nonexperts; what do we say to those people? On the other, what is most useful to advanced data analysts?

The Short-Course Crowd

We are (I hope) primarily teaching very fundamental principles of design and analysis, so that these people will be able to communicate better with statisticians and understand what's important to us. They won't be dealing with artful analyses, diagnostic plots, or nonregular designs. We show them nicely balanced, clean designs so that they get an idea of how these work.

I still don't think Daniel plots are very well suited for this audience. And I still think a Pareto plot of effects with a cut-off is useful and interpretable. But we also have to remember that an element of "magic" is present when we introduce powerful software tools, and sometimes it's better to slow down and show more about the underlying principles.

For communicating with nonexperts, I continue to like dot plots a lot, and I am not alone. As mentioned before, de Léon shows that a dot plot is as effective as a Daniel plot for identifying active effects. Moreover, Box et al. (2005) is full of dot plots, often with accompanying reference distributions. With that in mind, I offer the reference plot shown in Figure 1, for Voelkel's vehicle example with all data included. The normal curve is an enhancement to the dot plot to serve to guide in identifying active

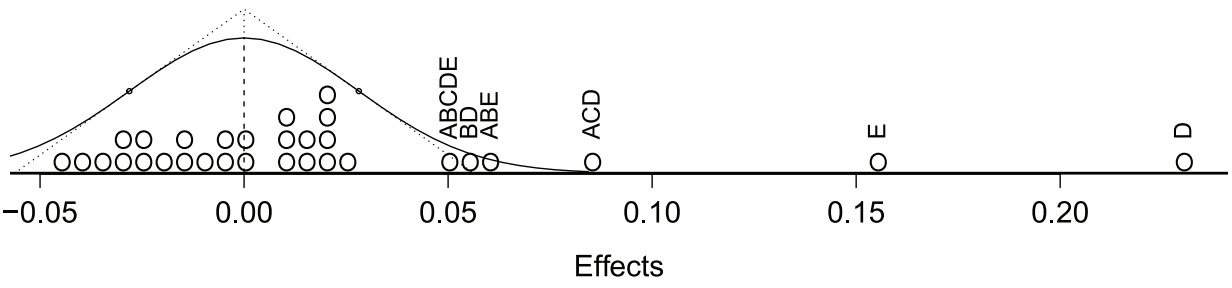


FIGURE 1. Reference Plot for the Vehicle Example.

effects. It has mean 0 and standard deviation equal to $1.5\text{med}|e_j|$—very easy to compute, and it is the first step in calculating the PSE. Superimposing this reference distribution is equivalent to adding a guide line on a Daniel plot determined by the origin and the median absolute effect.

Active effects are those that are obviously out beyond the tails. In this illustration, those would be D, E, and ACD. After identifying these, you may then, if you wish, follow the replotting idea—which would entail tossing out the obviously active effects and recomputing the reference SD with those effects excluded. Oh, by the way, that calculation will yield the PSE, or something similar. In this way, the idea of the PSE arises naturally and meaningfully. (In fact, this PSE-like calculation adapts itself to the number of really large effects, which may be appealing in light of Schoen and Goos's comments about tuning.) The underlying effect-sparsity principle is quite apparent in this reference plot—that we are judging active effects relative to the variations in the smaller ones. And the familiar bell-shaped curve makes this idea a lot clearer than do deviations from a line.

For those who like the informal nature of Daniel plots, I think that Figure 1 is just as effective in visualizing active effects, and it is much less prone to misinterpretation. It also has a self-contained multiplicity adjustment, as mentioned by Steinberg. By the way, this is a borrowed idea—very similar to Figure 5.10 in Box et al. (2005). Admittedly, the reference plot does not have the same diagnostic effectiveness that several discussants ascribe to Daniel plots. But we're not teaching diagnostics to this audience.

A display like this is very easy to do with pencil and paper—a big advantage for short courses where available technology is likely limited, and hands-on work is advantageous. A $N(\mu, \sigma^2)$ curve is easy to draw accurately by hand: start with an isosceles triangle with vertices at $\mu \pm 2\sigma$ and above μ (this is shown faintly in the figure). The inflection points are at the midpoints of the sides of this triangle. Fill-in four J-shaped curves originating at the inflection points, tangent to the sides of the triangle—two ending at a point somewhat below the peak (about 82%), and two leveling off at the baseline at $\mu \pm 3\sigma$. It is easier and more fun to teach this than how to get normal scores.

The Experts

For this group, I think I need to take seriously the point that nonregular designs are becoming quite common, and neither a Daniel plot nor a Pareto chart will work without some adaptations. If you are an expert, you, of course, should choose the methods that you find most valuable. Cook and Nachtsheim mention a few good ones, and show how much they differ (mixed results for the same data are not unique to Daniel plots). Accordingly, the expert group will use more varied and more sophisticated methods, will have more and better technology available, and will value having good diagnostic tools.

In this context, I'd like to highlight, as an additional useful tool, the Bayesian method in Box and Meyer (1986), mentioned in passing in my main article. While it is a relatively early contribution, it survives into the present [thanks to Markov chain Monte Carlo (MCMC) methods] and gracefully handles nonregular designs and even supersaturated ones. It is based on an effect-sparsity model whereby effects are thought of as draws from a mixture of two normals. One of them has a variance of, say, $k = 10$ times the other, and effects drawn from this one are the active ones. The mixing parameter α is set based on one's prior belief about how many effects are active; typically, we choose $\alpha = .2$, or 20% active. We then obtain the posterior probabilities for each effect being active. In the case of nonregular designs, the computations may be done using Gibbs sampling, as detailed in Chipman et al. (1997). The Gibbs results also provide information on which combinations of predictors are visited the most often.

Figure 2 summarizes the effects having the highest posterior probabilities in six Bayesian analyses of the vehicle data. The top three analyses use the same $\alpha = .2$ for all effects as do Box and Meyer, based on (1) the complete dataset, (2) excluding the outlier (observation 26), and (3) accommodating the possible removal of observation 26 by including an indicator for that observation among the candidate predictors. The bottom three analyses are respectively the same, only a hierarchical "weak inheritance" prior Chipman et al. (1997) is used whereby, for example, the conditional prior probability for ACD being active depends increasingly on how many of its contained effects AC, AD, and CD are also in the model. The results displayed are the observed relative frequencies for the last $10,000$ of $10,500$ rounds of Gibbs sampling (which still only took a fraction of a second on a desktop PC).

Looking first at the bottom row of plots—those for the weak inheritance prior—we find that all three have high posterior probabilities only for D and E,

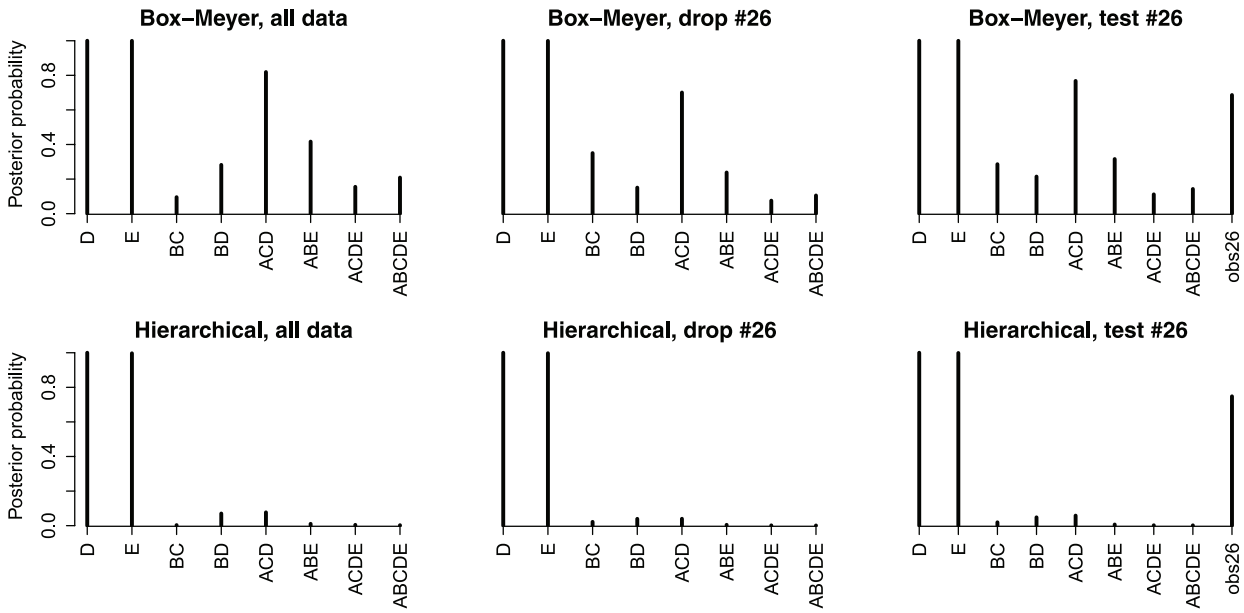


FIGURE 2. Bayesian Model-Selection Results for Effects Having Posterior Probabilities Exceeding 0.1 Somewhere Among Six Analyses of the Vehicle Data. The top three use equal prior probabilities of $\alpha = 0.20$ for all effects. The bottom three use decreasing prior probabilities for higher order effects. The analyses are performed using the complete data, one outlier removed, and with an indicator for the outlier.

with all the interactions pretty much suppressed. The prominence of the `obs26` indicator in the last analysis makes a fairly strong case for deleting observation 26. If we have a prior belief in effect hierarchy, these results give us permission, in a way, to ignore the higher order effects and to delete the 26th observation.

But these hierarchical results are the wrong ones to compare with Voelkel's analysis using Daniel plots, which treat all effects equally regardless of order. The fair comparison is with the top three analyses in Figure 2, which all give a high posterior to ACD. Note that dropping observation 26 does not make this go away, but instead boosts the importance of the two-way interaction BC. The third analysis shows that there is a slightly stronger case for including ACD in the model than for excluding observation 26.

How do these results compare with other analyses? With the outlier excluded, Voelkel dropped $ABCDE$—one of the larger effects with the full data—from the model in order to obtain estimates for his approximate Daniel plot, and obtains only D as an active effect with E "possibly suggested" (though it looks like it's on the line). A JMP analysis of these data (see Figure 3, with an indicator for observation 26, which shows effects D, E, and the `obs26` indicator unequivocally active; and ACD and BC also have individual P values less than .03. Like Voelkel, JMP also discards $ABCDE$ in order to fit the model, but it orthogonalizes the effects used in obtaining the PSE. The JMP analysis with observation 26 excluded has very similar results. So both the JMP and Box–Meyer analyses disagree pretty strongly with Voelkel's final plot. This illustrates how flaky a Daniel plot can become when effects are dependent, unless care is taken to orthogonalize.

I like the Box–Meyer strategy because there is no need to exclude any predictors and we also obtain results on which *models* were selected most often. It could be that there is more than one combination of predictors that can explain the response; and if so, we as experts need to know about it.

This vehicle example is not very straightforward. Maybe the fact that the run order was not randomized is more important than Voelkel states. Otherwise, I am not willing to strongly favor a hierarchical model until I have some belief that I won't lose something: sometimes, high-order effects are important. Years ago, a student of mine did an internship with a local manufacturer, and helped them with an ex-

Contrasts

Term	Contrast		Lenth t-Ratio	Individual p-Value	Simultaneous p-Value
D	0.230000		9.83	<.0001 *	<.0001 *
E	0.154375		6.60	<.0001 *	0.0017 *
D*A*C	0.059115	*	2.53	0.0222 *	0.3442
D*A*B	0.044707	*	1.91	0.0651	0.7481
A*B*C	0.041565	*	1.78	0.0806	0.8359
D*E*A	0.039416	*	1.68	0.0942	0.8838
D*B	0.037176	*	1.59	0.1140	0.9310
D*E*B	0.034374	*	1.47	0.1453	0.9676
D*E*C	-0.028614	*	-1.22	0.2166	0.9973
A	-0.029763		-1.27	0.2008	0.9945
B*C	-0.058246	*	-2.49	0.0233 *	0.3513
obs26	-0.086813		-3.71	0.0033 *	0.0751

FIGURE 3. Strongest Effects in the JMP Analysis of the Vehicle Data, with an Indicator for Observation 26. Effects marked with a "*" are ones that are forced to be orthogonal.

periment on a problem related to product damage incurred during shipping. Among a whole pile of analyses, the student found one three-way interaction. I thought it was probably just a spurious finding, but the company ran with it anyway, and it turned out to save them millions of dollars in scrap costs.

By the way, I applied the Box–Meyer method ($\alpha = .2$, $k = 10$) to the same examples in Cook and Nachtsheim's discussion, and obtained the following numbers of effects having posterior probability greater than .5: $(1, 0, 2, 0, 7, 1, 0)$. These results come out a lot like those for the randomization test. I am not sure I have the same data as in their example 5.

Another aspect of nonregular designs is that, because we can't use Daniel plots, we lose their highly touted diagnostic capabilities. What can we do instead, supposing we have used Box–Meyer or some other model-selection strategy? I suggest just using traditional residual diagnostics for the most-often-visited models. Some models may produce radically different diagnostics than others, but if so, that's worth knowing about, and may suggest a follow-up experiment to understand why there are differing explanations for the same results. Even with an orthogonal design, I agree with Jones that the diagnostic advantages of Daniel plots are for the most part overshadowed by more straightforward residual diagnostics based on effective but parsimonious models suggested by the analysis.

Conclusion

In rereading my article, I admit to being a little surprised by my strident tone when I recommend that we abandon Daniel plots and sentence them to a fiery demise in the nearest incinerator, or at least the recycling bin. But I stand by my observations that these are the wrong things to push at nonexpert users. Moreover, nonorthogonal cases seem likely to become more and more common, and for that reason, Daniel plots may eventually go the way of the slide rule and Yates's algorithm.

I know that change is tough, and it's hard to imagine a Daniel-plot-free world. But I can offer some insight based on my experience in vacating my office when I retired, and giving away most of my books and journals: You won't miss them as much as you think.

Additional References

Box, G. E. P. and Meyer, R. D. (1986). "An Analysis for Unreplicated Fractional Factorials". *Technometrics* 28, pp. 11–18.

Box, G. E. P.; Hunter, J. S.; and Hunter, W. G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd edition. Hoboken, NJ: Wiley.

Chipman, H.; Hamada, M.; and Wu, C. F. J. (1997). "A Bayesian Variable-Selection Approach for Analyzing Designed Experiments with Complex Aliasing". *Technometrics* 39, pp. 372–381.