Let’s practice what we preach: Planning and interpreting simulation studies with design and analysis of experiments

Hugh CHIPMAN1* and Derek BINGHAM2

1Department of Mathematics and Statistics, Acadia University, Wolfville, Nova Scotia, Canada
2Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada

Key words and phrases: ANOVA; data reduction; factorial and fractional factorial designs; simulation study; tables; Taguchi robust parameter design.

MSC 2020: Primary 62K99; secondary 62K25.

Abstract: Statisticians recommend design and analysis of experiments (DAE) for evidence-based research but often use tables to present their own simulation studies. Could DAE do better? We outline how DAE methods can be used to plan and analyze simulation studies. Tools for planning include cause-and-effect diagrams and factorial and fractional factorial designs. Analysis is carried out via analysis of variance, main effect and interaction plots, and other DAE tools. We also demonstrate how Taguchi robust parameter design can be used to study the robustness of methods to a variety of uncontrollable population parameters.

Résumé: Les statisticiens prônent le recours aux plans et analyse d’expériences (DAE) en recherche factuelle, mais le plus souvent, ils se contentent de tableaux pour présenter leurs propres études de simulation. Les auteurs de ce travail examinent l’apport et l’amélioration que pourrait apporter l’adoption de l’approche (DAE). Pour y répondre, ils présentent une description de la façon dont les méthodes DAE peuvent être employées pour planifier et analyser les études de simulation. Les outils de planification proposés comprennent les diagrammes de cause à effet et les plans factoriels et fractionnaires. L’analyse préconisée repose sur l’ANOVA, les graphiques d’effets principaux et d’interactions et d’autres outils DAE. Les auteurs montrent également comment les plans robustes de Taguchi peuvent être utilisés pour étudier la robustesse des méthodes à divers paramètres de population non contrôlables.

1. INTRODUCTION

It would be difficult to impossible to find a contemporary issue of a statistical or data sciences journal without a simulation study. Simulation studies typically generate random realizations of samples, systematically varying population or model parameters across realizations. At each combination of parameters, the performance of models, inferential methods, or algorithms is evaluated empirically (e.g., significance level, power, predictive ability, etc.). In other words, a designed experiment is performed on the statistical procedure. So, why not use well-established design and analysis of experiments (DAE) techniques to design simulation studies and synthesize the results?

We illustrate how tools for the design of physical experiments can yield efficient simulation studies and lead to easy-to-interpret results. This approach offers the potential to discover deeper interpretations from simulation studies than would be evident from simply looking
at tabulated results. The benefits are greatest for larger studies, where four, five, or more factors are varied, and where the layout of the table of reported results could either reveal or hide relationships. Meta-models such as analysis of variance (ANOVA) are designed for identification and interpretation of important effects.

Using a designed experiment can save time while giving accurate and interpretable results. In ANOVA, effect estimates use all experimental runs, giving much smaller standard errors than a “one cell at a time” analysis. This, in turn, implies that far fewer replications of the study are needed than an analysis of a single cell’s Monte Carlo error would suggest. Fractional factorial experiments offer time savings in comparison to full factorial designs and permit the systematic study of more factors. We also illustrate Taguchi’s robust parameter design (TRPD), a design and analysis tool that focuses on identifying and controlling the effects of random variation. For studies of statistical procedures, a TRPD approach helps us to find models and methods whose performance is stable across uncontrollable factors such as unknown population parameters, sampling schemes, and violations of model assumptions.

The message that statisticians would do well to use their own tools for their own studies is not new. Hoaglin & Andrews (1975) argued for better reporting of computation-based results in statistics. Gelman, Pasarica & Dodhia (2002) recommended that statisticians use graphs instead of tables to report their research results. Several articles have used DAE tools to report their research results, for example, Culpepper & Aguinis (2011). Others, such as Skrondal (2000), Harwell, Kohli & Peralta (2017), and Morris, White & Crowther (2019), provide general advice in the planning and analysis of simulation studies. Both Skrondal (2000) and Harwell, Kohli & Peralta (2017) argue for the use of a factorial design approach for planning and a “meta-model” for analysis. Skrondal (2000) is particularly similar to our approach, although generalized linear models (GLMs) are favoured over the simpler ANOVA we demonstrate. The factorial design approach is in contrast to the focus on Monte Carlo error associated with a single table entry (e.g., Morris, White & Crowther, 2019). Dolatsara et al. (2021) advocate for the use of experimental design to systematically study the effect of data preparation and model choice in supervised learning applications.

The rest of the article is outlined as follows. In Section 2, we motivate the problem by introducing a simulation study originally presented in Krishnamoorthy, Mallick & Mathew (2011). They examine the effect of five different factors on the type I error rate of statistical hypothesis tests. Section 3 presents our general approach to the planning, execution, and analysis of a simulation study. We return in Section 4 to the motivating example, identifying some new insights. Section 5 presents an end-to-end example in which we design, execute, and analyze a study of the performance of two statistical learning algorithms. We conclude the article with discussion of some related issues.

2. MOTIVATING EXAMPLE

To demonstrate advantages of using DAE in planning and analyzing simulation studies, we consider a previously published study. Krishnamoorthy, Mallick & Mathew (2011, hereafter KMM) examined the performance of four hypothesis-testing procedures for the mean of a log-normal distribution, using censored data. Table 1 summarizes the five factors varied in the study. All tests were at level $\alpha = 0.05$.

KMM’s results are reproduced in Table 2. For each entry, an unspecified number of realizations of a sample from the log-normal distribution were generated under the null hypothesis. Type I error rates were estimated by the proportion of the simulated datasets for which the method rejects $H_0$. KMM give qualitative statements summarizing patterns in the table.

The simulation study can be thought of as a full factorial experiment with five factors. There are a total of $4 \times 3 \times 3 \times 4 \times 3 = 432$ combinations of $\text{method} \times \text{tail} \times n \times p_0 \times \sigma$. Each run of the $3^3 4^2$ full factorial design corresponds to one of the 432 cells in Table 2.
TABLE 1: KMM study: description and levels of the factors.

| Factor | Description                                      | Levels |
|--------|--------------------------------------------------|--------|
| Method | Hypothesis testing procedure                     | (AN = asymptotic normality of the MLE, SL = signed log-likelihood ratio test, GV = generalized variable method, MS = modified signed log-likelihood ratio test) |
| Tail   | Tail of the hypothesis test                       | L = left, R = right, T = both tails |
| n      | Sample size                                      | 20, 30, 50 |
| p0     | Proportion of left-censored observations         | 0.20, 0.30, 0.50, 0.70 |
| sigma  | Standard deviation of lognormal population       | 1, 2, 3 |

Since the simulation study is a designed experiment, it can be analyzed as such. Table 3 displays an ANOVA for a fourth-order model based on all runs with “type I error rate” (recorded as a value between 0 and 100) as the response. The remaining 72 degrees of freedom are treated as residual, although they could be assigned to the fifth-order interaction, leaving no estimate of residual error. Factors n, p0, and sigma are numeric, but treated as categorical in the ANOVA.

Even a cursory inspection of the ANOVA table indicates many significant and some large effects, including high-order interactions. For example, the two-way interaction method:tail is very large, indicating that the way in which the type I error rates vary across tail type varies by method. Such interactions are difficult to spot by inspecting tabulated results, and often rely on how the factors are arranged in rows and columns as in Table 2. In Section 4, we demonstrate how ANOVA gives insight into the study.

3. DESIGN AND ANALYSIS OF A SIMULATION STUDY

Broadly speaking, an experiment consists of five stages: (i) problem definition, (ii) planning, (iii) execution, (iv) analysis, and (v) conclusions (e.g., MacKay & Oldford, 2000). Originally developed for physical experiments, these stages can be adapted to simulation studies. Our description is general, but the ideas can be adapted to almost any setting. In this section, we use the KMM study and our statistical learning study (Section 5) for illustration.

The problem definition (stage i) usually involves the definition of a Quantity of Interest (QoI) that measures some aspect of performance. The QoI can either be calculated individually for every run or be an estimate formed using all replicates with the same experimental settings. In the KMM study, the QoI is the estimated type I error rate, expressed as a percentage in Table 2. The QoI is an estimate of an unknown parameter, such as the type I error rate. Multiple QoIs may be appropriate for a single study. In our statistical learning study, the problem is to compare the predictive accuracy of two statistical learning models. In that case, the QoI is $R^2$ measured on a test set.

A first step in planning (stage ii) a simulation study is to establish the list of factors to include in the experiment. A large number of potential factors should be identified and then reduced. A reduced set of factors will lead to an experimental design that can be run in a reasonable amount of time. A useful tool for organizing the factors that might impact a QoI is a cause-and-effect diagram (Ishikawa, 1982). In physical experiments, the different factors impacting the performance of a process are usually grouped as branches.
| n  | p₀  | Method | σ = 1 | σ = 2 | σ = 3 |
|----|-----|--------|-------|-------|-------|
|----|-----|--------|-------|-------|-------|
| 20 | 0.2 | GV     | 5.4   | 4.4   | 5.0   |
|    |     | AN     | 2.1   | 10.5  | 8.2   |
|    |     | SL     | 4.6   | 6.9   | 5.9   |
|    |     | MS     | 5.3   | 4.2   | 4.7   |
| 0.3|     | GV     | 5.3   | 4.4   | 5.8   |
|    |     | AN     | 2.1   | 11.0  | 8.3   |
|    |     | SL     | 4.4   | 5.8   | 5.1   |
|    |     | MS     | 5.3   | 3.6   | 4.3   |
| 0.5|     | GV     | 5.9   | 3.6   | 4.4   |
|    |     | AN     | 2.2   | 10.8  | 8.5   |
|    |     | SL     | 4.8   | 5.9   | 5.6   |
|    |     | MS     | 4.9   | 2.9   | 3.6   |
| 0.7|     | GV     | 6.2   | 1.6   | 4.3   |
|    |     | AN     | 3.9   | 0.3   | 1.7   |
|    |     | SL     | 5.2   | 4.0   | 4.3   |
|    |     | MS     | 4.4   | 3.0   | 3.7   |
| 30 | 0.2 | GV     | 5.4   | 4.7   | 5.4   |
|    |     | AN     | 2.8   | 9.6   | 7.1   |
|    |     | SL     | 4.6   | 6.0   | 5.5   |
|    |     | MS     | 5.1   | 4.4   | 4.7   |
| 0.3|     | GV     | 5.6   | 4.6   | 5.4   |
|    |     | AN     | 2.5   | 9.2   | 6.9   |
|    |     | SL     | 5.0   | 6.1   | 5.4   |
|    |     | MS     | 5.5   | 3.6   | 4.4   |
| 0.5|     | GV     | 6.0   | 3.9   | 5.2   |
|    |     | AN     | 2.4   | 9.8   | 7.6   |
|    |     | SL     | 5.2   | 3.9   | 5.1   |
|    |     | MS     | 4.8   | 3.3   | 3.9   |
| 0.7|     | GV     | 6.6   | 2.4   | 4.8   |
|    |     | AN     | 3.9   | 4.2   | 2.3   |
|    |     | SL     | 4.5   | 4.0   | 4.2   |
|    |     | MS     | 4.6   | 2.7   | 3.2   |
| 50 | 0.2 | GV     | 5.3   | 5.3   | 4.8   |
|    |     | AN     | 2.8   | 8.0   | 6.0   |
|    |     | SL     | 4.3   | 5.9   | 5.5   |
|    |     | MS     | 5.1   | 4.7   | 4.9   |
| 0.3|     | GV     | 6.7   | 4.4   | 4.9   |
|    |     | AN     | 2.6   | 8.3   | 6.0   |
|    |     | SL     | 4.9   | 5.9   | 5.8   |
|    |     | MS     | 5.0   | 4.7   | 4.8   |
| 0.5|     | GV     | 6.1   | 4.1   | 5.2   |
|    |     | AN     | 2.7   | 8.6   | 6.5   |
|    |     | SL     | 4.9   | 5.8   | 5.1   |
|    |     | MS     | 5.0   | 3.9   | 4.3   |
| 0.7|     | GV     | 7.9   | 1.5   | 4.8   |
|    |     | AN     | 3.5   | 5.6   | 4.0   |
|    |     | SL     | 4.8   | 4.5   | 4.7   |
|    |     | MS     | 4.7   | 4.1   | 4.2   |

Note: Entries are percentages on a 0–100 scale.

DOI: 10.1002/cjs.11719
Table 3: KMM study: ANOVA table for the 432-run full factorial experiment.

|                     | Df | Sum Sq | Mean Sq | F value  | Pr(>F) |
|---------------------|----|--------|---------|----------|---------|
| method              | 3  | 555.1  | 185.0   | 629.713  | < 2e-16 *** |
| tail                | 2  | 332.3  | 166.1   | 565.361  | < 2e-16 *** |
| n                   | 2  | 11.4   | 5.7     | 19.353   | 1.88e-07 *** |
| p0                  | 3  | 2.7    | 0.9     | 3.117    | 0.031319 *  |
| sigma               | 2  | 99.2   | 49.6    | 168.785  | < 2e-16 *** |
| method:tail         | 6  | 2258.0 | 376.3   | 1280.705 | < 2e-16 *** |
| method:n            | 6  | 50.4   | 8.4     | 28.571   | < 2e-16 *** |
| method:p0           | 9  | 35.4   | 3.9     | 13.374   | 2.64e-12 *** |
| method:median       | 6  | 137.7  | 23.0    | 78.107   | < 2e-16 *** |
| tail:n              | 4  | 11.6   | 2.9     | 9.867    | 2.02e-06 *** |
| tail:p0             | 6  | 21.7   | 3.6     | 12.290   | 1.79e-09 *** |
| tail:median         | 4  | 90.1   | 22.5    | 76.693   | < 2e-16 *** |
| n:median            | 6  | 1.0    | 0.2     | 0.545    | 0.771926 |
| n:median            | 4  | 11.9   | 3.0     | 10.139   | 1.45e-06 *** |
| p0:median           | 6  | 80.7   | 13.5    | 45.777   | < 2e-16 *** |
| method:tail:n       | 12 | 48.3   | 4.0     | 13.693   | 3.39e-14 *** |
| method:tail:p0      | 18 | 60.1   | 3.3     | 11.370   | 1.54e-14 *** |
| method:tail:median  | 12 | 298.1  | 24.8    | 84.525   | < 2e-16 *** |
| method:n:p0         | 18 | 5.0    | 0.3     | 0.936    | 0.539622 |
| method:n:median     | 12 | 13.4   | 1.1     | 3.813    | 0.000169 *** |
| method:p0:median    | 18 | 103.3  | 5.7     | 19.528   | < 2e-16 *** |
| tail:n:p0           | 12 | 0.9    | 0.1     | 0.268    | 0.992458 |
| tail:median         | 8  | 5.1    | 0.6     | 2.170    | 0.039827 *  |
| tail:p0:median      | 12 | 37.3   | 3.1     | 10.591   | 1.08e-11 *** |
| n:p0:median         | 12 | 10.8   | 0.9     | 3.050    | 0.001627 ** |
| method:tail:n:p0    | 36 | 6.8    | 0.2     | 0.640    | 0.928695 |
| method:tail:n:median| 24 | 18.4   | 0.8     | 2.606    | 0.000950 *** |
| method:tail:p0:median| 36 | 129.9  | 3.6     | 12.281   | < 2e-16 *** |
| method:n:p0:median  | 36 | 22.2   | 0.6     | 2.098    | 0.003836 ** |
| tail:n:p0:median    | 24 | 6.1    | 0.3     | 0.866    | 0.644092 |
| Residuals           | 72 | 21.2   | 0.3     | Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

on the cause-and-effect diagram. In physical experiments, these branches correspond to the following major categories: methods, machines, people, materials, and measurement. In a simulation study, the factors can instead be categorized under methods, population, population design, simulation design, and validation design, as depicted in Figure 1. In a standard cause-and-effect diagram, the population design settings would usually be given by substructures within the population design. We have found it easier to visualize by separating population and population design.

Figure 1 provides a framework to list potential factors, by filling in the branches. We illustrate this using the KMM study and the corresponding cause-and-effect diagram in Figure 2. Under
FIGURE 1: A generic cause-and-effect diagram for simulation experiments.

FIGURE 2: Cause-and-effect diagram for the KMM simulation study.
methods, for example, the list of statistical methods to be compared is likely clear from the context (here, AN, ML, SL, and GV are compared). The tail of the test can also be listed under method. Similarly, population is a factor when there is more than one population model from which simulations are to be drawn. In the KMM study, data might arise from a lognormal population or a gamma distribution, although KMM considered only the lognormal. Under the population design branch, population parameters (e.g., mean, variance, etc.) are potential factors. KMM considered the amount of left censoring and the error variance as population parameters. Together, the population and population design branches determine the distributions from which the data are simulated.

Different types of studies will obviously have different kinds of populations. The KMM study is a population of univariate measurements (e.g., survival times arising from a lognormal distribution). Our statistical learning study simulates an observational data scenario in which the response variable and predictor variables are observed, and arise from some joint distribution.

Facing any experimenter is the issue of sampling from the population. The simulation design has factors that are external to the population, such as the sample size (KMM considered \( n = 20, 30, \) and 50). The validation design typically would entertain similar factors as the simulation data, but may consider other population parameters if robustness is being investigated under validation samples that differ from the training set. In KMM, there is no validation design and no corresponding factors for this branch. In our statistical learning study, a validation set is used to evaluate predictive performance, and so the number of observations in the validation set is a factor that could be considered in the validation design.

The cause-and-effect diagram is a first step at identifying and organizing the factors but typically lists more factors than can be accommodated in most physical experiments. The factors thought to be of most interest are then selected from the factors listed on the cause-and-effect diagram, while the remaining factors are fixed at nominal levels in subsequent stages. For example, KMM generated data only from a lognormal population, rather than varying the population model.

Next is the selection of the factor level settings for the simulation study. For some factors, this task is straightforward. If model is a factor, then the factor levels are simply the list of models included in the study. For a numeric factor such as sample size or amount of censoring, the choice in levels needs to be realistic in real-world settings and also should be spaced far enough apart so that potentially meaningful differences in the QoI can be observed. We see later in the KMM example that the sample sizes used may have been too close.

With a list of unique levels for each factor, the simplest experimental design is a full factorial, containing all possible combinations of all factor levels. For example, in the KMM study, a full factorial design could be formed with all \( 4 \times 3 \times 3 \times 4 \times 3 \) possible combinations of the levels of the five factors. As the number of factors and the number of levels grow, the number of runs of the simulation design increases quickly. When full factorial designs become too time consuming, we recommend fractional factorial designs (e.g., Box, Hunter & Hunter, 2005). In the illustrations considered in this article, we consider only two-level fractional factorial designs; however, there are fractional factorial designs with different numbers of levels (e.g., see Mukerjee & Wu, 2006; Wu & Hamada, 2009).

With the choice of level settings and factor level combinations (i.e., the simulation experimental design) in hand, it is tempting to move directly to stage iii and execute the entire simulation study. Before doing so, there are still some details to address.

First, one must decide how many replicates of the simulation design to conduct. The QoI is often an aggregate over the replicates (e.g., an average or percentage), so variation in the QoI will be reduced by using more replicates. The type of the QoI and its corresponding Monte Carlo error will be instrumental in choosing the number of replicates. For instance,
in the KMM study, a single replicate of one cell of the table gives a binary outcome (reject or do not reject $H_0$), and so one would probably want to use more than 1000 replicates to obtain a sufficiently small Monte Carlo error for the type I error. We show later in the statistical learning study that a single replicate can be sufficient. We also argue that by fitting a meta-model to all experimental runs, even a small number of replicates may give sufficient accuracy.

Second, it is frequently worth running a small pilot study to see if the simulations can be executed as planned. The pilot study also allows evaluation of the choice of each factor level settings and develops an estimate of the computational resources required to complete the study. For example, in the KMM study, if only sample sizes of 20 and 30 were considered, it is likely that the differences in the QoI would be too small to reach concrete conclusions. A small pilot study would likely detect this and lead to different choices of factor level settings.

Following the pilot study, the simulation design is run (stage iii). This is a conceptually simple step, performing the desired simulations. A single instance of the simulation may include simulating data for estimating a model, simulating a validation set, model fitting, making predictions or conducting hypothesis tests, and computing the QoI. For studies requiring long computation times, parallel computing may be required.

The analysis (stage iv) is the study of the relationship between the QoI and the factors under study. As with the analysis of designed experiments, we recommend ANOVA models and associated tools as the primary techniques used in this stage. GLMs may be appropriate for some QoIs. The usual “statistical toolbox” for analysis, including model diagnostics, hypothesis testing, transformation, prediction intervals, etc., will be useful. Indeed, the purpose of a careful simulation design is to make the analysis efficient, interpretable, and also relatively easy to do.

ANOVA (or other meta-models) offer the opportunity to reduce uncertainty in the QoI beyond replication. For example, in a $2^k$ factorial design replicated $n$ times, we would consider fitting an ANOVA model via the regression formulation $Y = X\beta + \epsilon$, where $Y$ is the QoI, $X$ is a vector of covariate levels (coded as ±1) corresponding to the factors under study, $2\beta$ is the vector of effects, and $\epsilon$ is the Monte Carlo error in the study, with $\text{Var}(\epsilon) = \sigma^2$. The effect estimates (elements of $2\hat{\beta}$), which represent the mean change in $Y$ when the corresponding term of $X$ goes from −1 to +1, have standard error given by $\text{se}(\hat{\beta}_j) = 2\sigma/\sqrt{n2^k}$, while the mean for a single cell (across $n$ replicates) would have standard error of $\sigma/\sqrt{n}$. The $2^k$ term represents a reduction in the estimation uncertainty of the effect, due to the meta-model. For example, a study with a 64-run design will have effects that have standard errors of effect estimates that are smaller than cell estimate standard errors by a factor of $\sqrt{64}/2 = 4$. Associating standard errors with effects (and a meta-model) instead of with individual cells implies that fewer replicates are needed to control estimation error.

TRPD is a modelling and optimization tool useful when some factors are under the control of researchers (e.g., the choice of inferential procedure) and other factors are uncontrollable (e.g., an unknown population variance). In TRPD, we seek combinations of the “control” factors that give results which are robust across the levels of the uncontrollable (or “noise”) factors. For an overview of TRPD, see Nair et al. (1992).

The conclusions (stage v) should flow naturally from the analysis. A convenient feature of ANOVA models is the ability to identify important factors and, in the case of factors with more than two levels, determine which factor levels are significantly different. For significant main effects and interaction effects, we recommend using plots to observe the magnitude and nature of the effects.
We propose that the tools outlined above be applied to the design of simulation studies. In the following sections, we illustrate how the five experimental stages can aid in understanding the performance of methods in two examples.

4. RETURN TO MOTIVATING EXAMPLE

We revisit the KMM study, considering the five stages of experimentation outlined in Section 3. One complete cycle is demonstrated in Section 4.1, including a TRPD analysis. Additional analysis of higher order interactions in Section 4.2 gives further insights. Our conclusions are compared with KMM’s original findings in Section 4.3.

4.1. The Five Stages in the KMM Example

(i) Problem definition

The aim of the KMM study was to investigate the merits of four hypothesis testing procedures for the lognormal mean, in the presence of left-censoring, using simulation. As such, the problem definition is identified, and the chosen QoI for the simulation is the type I error rate. As in KMM, we will assume that it is preferable to be at or slightly below the nominal 5% level.

(ii) Planning

Since we did not plan the experiment, a full description of this stage is not available. In particular, we did not have a full list of potential variables to include in the experiment. The cause-and-effect diagram for the $3^34^2$ full factorial design in the simulation study was presented in Figure 2. The cause-and-effect diagram includes a few factors that were held fixed, such as the distributional family (lognormal) and population mean (set at 0), but KMM would likely have considered them as potential factors.

(iii) Execution

We would normally recommend a pilot study to investigate whether the level settings give meaningful differences in the QoI. It also allows the experimenters to predict the run time of computations for the full suite of simulations.

If a pilot study had been carried out, it is possible that the $\text{AN}$ method would have been identified as having much larger variation in type I error than any other method. Reading across any $\text{AN}$ row of Table 2 reveals this.

An alternate way to run a pilot study would be to carry out a small number of replicates at all 432 level combinations of the full factorial, rather than the several thousand used in KMM.

(iv) Analysis

Simple examination of main effect and two-way interaction plots (not shown) indicate the poor type I error performance of the $\text{AN}$ method, including a mean type I error rate of 7.6%. In light of the terrible performance of the $\text{AN}$ method, we removed the 108 runs with $\text{method} = \text{AN}$ and analyzed the remaining 324 runs as a $3^44^1$ full factorial design.

It is common to analyze simulation studies by visual inspection of the tabulated QoI (e.g., Table 2). Detection of all but the largest effects can be difficult and depends heavily on the placement and nesting of factors within the columns and rows of the table. Instead, we propose using ANOVA, or related methods, for the analysis.

Table 4 shows a summary of an ANOVA for the fourth-order model using the 324 runs. In comparison to the ANOVA of the 432 runs (Table 3), fewer interactions are now significant, and the estimate of residual variance (MSE, the mean square for residuals) is considerably smaller than for the original fourth-order ANOVA. Routine diagnostics indicate no problems with the ANOVA model, so we did not transform the response or report an analysis of a binomial GLM.
Main effects and two-factor interactions

Main effects are displayed in Figure 3a. Of note are the following:

- The largest main effect is tail.
- Main effect magnitudes are small, and the main effect of n is negligible.
- Response means are close to the nominal level of $\alpha = 5\%$.
- For the numeric factors n, p0, sigma, effects are monotone or almost so. For example, the mean response increases as sigma increases from 1 to 3.

### Table 4: KMM study: ANOVA table for 324-run full factorial experiment excluding method = AN.

|            | Df | Sum Sq | Mean Sq | F value | Pr(>F)   |
|------------|----|--------|---------|---------|----------|
| method     | 2  | 11.34  | 5.67    | 29.152  | 5.16e-09 *** |
| tail       | 2  | 52.96  | 26.48   | 136.192 | < 2e-16 *** |
| n          | 2  | 0.59   | 0.30    | 1.521   | 0.228797 |
| p0         | 3  | 4.07   | 1.36    | 6.973   | 0.000548 *** |
| sigma      | 2  | 8.52   | 4.26    | 21.913  | 1.73e-07 *** |
| method:tail| 4  | 215.13 | 53.78   | 276.611 | < 2e-16 *** |
| method:n   | 4  | 0.46   | 0.11    | 0.588   | 0.672858 |
| method:p0  | 6  | 1.56   | 0.26    | 1.337   | 0.259421 |
| method:.sigma| 4  | 7.46   | 1.86    | 9.590   | 8.81e-06 *** |
| tail:n     | 4  | 0.54   | 0.13    | 0.694   | 0.599625 |
| tail:p0    | 6  | 43.31  | 7.22    | 37.123  | < 2e-16 *** |
| tail:sigma | 4  | 1.18   | 0.30    | 1.518   | 0.211780 |
| n:p0       | 6  | 0.84   | 0.14    | 0.719   | 0.635852 |
| n:sigma    | 4  | 2.41   | 0.60    | 3.102   | 0.023800 * |
| p0:sigma   | 6  | 9.12   | 1.52    | 7.818   | 6.76e-06 *** |
| method:tail:n | 8  | 7.58   | 0.95    | 4.875   | 0.000194 *** |
| method:tail:p0 | 12 | 16.33  | 1.36    | 6.998   | 4.00e-07 *** |
| method:tail:sigma | 8  | 35.23  | 4.40    | 22.651  | 7.55e-14 *** |
| method:n:p0 | 12 | 4.57   | 0.38    | 1.957   | 0.050457 . |
| method:n:sigma | 8  | 2.47   | 0.31    | 1.587   | 0.153641 |
| method:p0:sigma | 12 | 3.87   | 0.32    | 1.660   | 0.106537 |
| tail:n:p0  | 12 | 0.50   | 0.04    | 0.214   | 0.997052 |
| tail:n:sigma | 8  | 2.08   | 0.26    | 1.335   | 0.249705 |
| tail:p0:sigma | 12 | 0.99   | 0.08    | 0.423   | 0.946606 |
| n:p0:sigma | 12 | 1.50   | 0.12    | 0.642   | 0.795650 |
| method:tail:n:p0 | 24 | 4.54   | 0.19    | 0.974   | 0.513875 |
| method:tail:n:sigma | 16 | 7.83   | 0.49    | 2.517   | 0.006996 ** |
| method:tail:p0:sigma | 24 | 21.98  | 0.92    | 4.710   | 2.51e-06 *** |
| method:n:p0:sigma | 24 | 6.93   | 0.29    | 1.485   | 0.120822 |
| tail:n:p0:sigma | 24 | 2.52   | 0.10    | 0.539   | 0.948027 |
| Residuals   | 48 | 9.33   | 0.19    |         |           |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
The presence of large and significant second-, third-, and fourth-order interactions (ANOVA in Table 4) strongly suggests that analysis cannot stop after examining the main effects. The two largest two-factor interactions, method:tail and tail:p0, are plotted in Figure 3b,c. While other interaction effects are significant, they are much smaller. Large interaction effects are seen as nonparallel lines. The largest two-factor interaction is method:tail (Figure 3b). All three methods exhibit considerable range in type I error rate over the three different tails. For example, when method = MS, the mean type I error rate is just above 6% for a left-tailed test, while it is well below 4% for a right-tailed test. Thus, a change in tail from left to right produces an interaction effect for the MS method of about 2.5%. This effect will be additive with the effects involving other terms and could push predictions of type I error rate much higher or lower.

Figure 3c shows the tail:p0 interaction effects. The type I error rate for left- and right-tailed tests varies with the censoring proportion p0. The downward-sloping dashed line indicates that right-tailed tests become increasingly conservative as p0 increases. The nearly flat solid line indicates that for two-tailed tests, type I error rate is stable and close to 5%. Ideally, the choice of method would be one that results in a type I error rate close to 5% across changes of factors that are not under the control of the experimenter. This desire for robustness leads us to consider the analysis of the KMM study in the context of TRPD.

**Taguchi robust parameter design**

The interpretation above of the tail:p0 interaction suggests that the two-tailed tests are robust to changes in p0. The idea of TRPD is to identify experimental factors that can be controlled, and choose levels of those factors so that the response has a value equal to a target, with small variation. Here is a quick primer on TRPD in this setting:

- The factors method and tail are control factors, which can be chosen by an analyst.
- The factors n, p0, and sigma are noise factors, which are largely beyond the control of the analyst. One might argue that n can be controlled, but for illustrative purposes, we assume that samples are expensive, and so changing n is unlikely to be controllable.
- The target value for the response in this example is the nominal 5% level.
- The idea of TRPD is to choose a level of the control factors so that responses are close to the target and are robust to the uncontrollable (noise) factors. Large interactions between control and noise factors provide an opportunity to achieve such robustness.

The previously identified tail:p0 interaction in (Figure 3c) can be interpreted in the context of TRPD. The tail of the hypothesis test is the control factor since the user could choose that. The user cannot control the censoring level p0, making it a noise factor. We want to choose a tail so that no matter what p0 is, our Type I error rate is close to the desired 5%. Choosing tail = T gives a more robust response (a line closer to horizontal).

Another way to examine robustness is to lump together the noise effects and look at the variation in response at each combination of the two control factors method and tail. Figure 4 does this, showing histograms of the response for each of the nine combinations of method and tail. We see the following:

- Two-tailed tests give type I error closer to the desired 5%, no matter what method is used. The histograms in the right column are centred around 5% and have small variation.
- If a left-tailed test is desired, the SL method is preferred over the other methods. It is slightly conservative with levels just below 5% and less variation in level than the other methods. The other methods are too liberal and have some levels well above 5%.
- If a right-tailed test is desired, GV and MS are conservative (sometimes quite so), while SL is liberal. All exhibit considerable variation in type I error, but the conservative tests may be preferred.
3.5 4.0 4.5 5.0 5.5 6.0 6.5

(a)

Factors

Type I error

GV

MS

SL

L

R

T

20

30

50

0.2

0.3

0.5

0.7

1

2

3

FIGURE 3: KMM study: (a) Estimated main effects and (b,c) select two-factor interactions. The vertical axis is the response, “Type I error rate”, which should have a nominal level of 5%, indicated by the horizontal blue line.

(b)

(c)

FIGURE 4: KMM study: Histograms of response within each combination of method:tail. Columns of the grid correspond to tail and rows correspond to method, as indicated by titles above each histogram.

These findings relate to both the centre and spread of the distribution of the response values. TRPD seeks levels of the control factor as a way of controlling both a target (i.e., a measure of location) and variation.

(v) Conclusions

The TRPD analysis has identified some important conclusions, in addition to the previous conclusions that the AN method is terrible, and some two-way interactions are larger than the fairly small main effects. We postpone further discussion of conclusions until Section 4.3.

DOI: 10.1002/cjs.11719
The Canadian Journal of Statistics / La revue canadienne de statistique
4.2. Higher Order Analysis and TRPD Again

Having identified and examined main effects and two-way interactions and considered robustness, an analysis would often be complete at this point. Higher order effects are often small. We have illustrated the five stages. In this example, however, it is worthwhile to press on further. Large third- and fourth-order interactions (ANOVA in Table 4) may explain additional variation in type I error.

To visualize three-way interactions, we combine the two control factors method and tail into a single factor with nine method/tail levels (GV/L, GV/R, GV/T, MS/L, …, SL/T) and plot the new combined control factor against the third factor and the response. The layout is similar to the two-way interaction plots in Figure 3, but with nine lines for the nine control factor combinations. Line type and colour represent method and tail, respectively.

The previous ANOVA table (Table 4) identifies the two largest three-factor interactions as method:tail:p0 and method:tail:σ. These interactions are visualized in Figure 5. Non-parallel lines indicate sizeable interactions between control and noise factors. The fact that control by noise interactions are large suggests that TRPD can be used to identify sources of uncontrollable variation, giving more insight into the methods. Here, we want to choose a method and tail so that no matter what n, p0, and σ are, our Type I error rate is close to the desired 5%.

Some interpretations of these results (not previously seen in the histograms) are as follows:

- From the method:tail:p0 interaction plot (Figure 5a), both the GV and the MS methods have poor type I error for large p0 (i.e., large amounts of censoring). The solid and dashed lines have an upward slope for left-tail (red) and downward slope for right-tail tests (green).
- From the method:tail:p0 interaction plot (Figure 5a), the SL method does not change its level as censoring changes (dotted lines are horizontal). However, the left-tailed SL test (red dotted) is conservative and the right-tailed SL test (green dotted) is liberal.
- From the method:tail:σ interaction plot (Figure 5b), the MS and SL methods have better type I errors for σ = 1 than σ = 2 or 3. The GV method seems less sensitive to σ. That is, the three solid lines are closer to horizontal.
In the KMM study, some fourth-order effects are also large. These are studied in the Supplementary Material.

4.3. Comparison to the Findings of KMM

KMM summarize their findings of this study over two paragraphs of text. Below, we quote their findings and compare them to our conclusions (in italic). We then identify some additional findings not given in KMM.

1. “The test based on the asymptotic normality of the MLE (denoted by AN in Table 1) seems to be the worst among all tests ....Thus, the test AN should be avoided in applications.” This was overwhelmingly clear and we dropped runs corresponding to method = AN in our analysis. A pilot study might have identified and eliminated AN before carrying out the full experiment.

2. “Performance depends on $\sigma$, $p_0$ and tail.” We observe this also, although we additionally note that the dependence is complex with higher order interactions that would be difficult to see by visual inspection of the results. Some effects containing $n$ are statistically significant, but are relatively small contributions.

3. “No test emerges as the clear winner in all the scenarios.” We come to the same conclusion (e.g., see Figure 4).

4. “For left-tailed testing, the SLRT appears to exhibit the most satisfactory performance, even though it is conservative in some cases. Both the GV approach and the MSLRT appear to be quite liberal in most cases, especially when $\sigma$ is large.” Figure 4 also confirms all these conclusions.

5. “For right-tailed testing, it is the GV approach that provides the most satisfactory solution.” GV is good for right-tail tests, but so is MS (Figure 4).

6. “For 2-tailed testing, all the three procedures provide satisfactory performance, even though the SLRT is liberal and the MSLRT is conservative in a few cases. The GV approach appears to be quite satisfactory for 2-sided testing, for all the cases considered in the simulation.” Our findings are similar.

7. “As the sample size gets large, there is very little difference between the GV approach and the MSLRT. In several instances, the MSLRT is not as satisfactory as the SLRT.” Table 3 indicates that only one significant effect involves $n$, namely method:tail$n$. KMM’s finding is based partly on a separate run not included in the study, with $n = 100$. Over the range $n = 20, 30, 50$, the effect of sample size is relatively small. As previously noted, we would expect that the relatively small effect of $n$ might have been detected in a pilot study, and potentially suggested including $n = 100$ instead of one of the other settings (likely $n = 30$).

Did we discover anything KMM did not? Perhaps:

- Type I error rates vary in a complex way over the five factors. This is evidenced by large and significant interaction effects, even at the third and fourth order.
- After excluding method = AN, left-tail tests are, on average, a bit liberal. Right-tail tests are a bit conservative, and two-tailed tests are, on average, at the correct level.
- Sample size $n$ has a relatively small effect, compared to the other factors. This may be in part due to the narrow range of levels ($n = 20, 30, 50$). Indeed, the authors report some additional results for $n = 100$, not contained in the main study. One might hope that with sufficiently large samples, type I error rates would approach the nominal level.
- KMM identify only GV as the preferred method for right-tail tests. We find that method = MS is just as good (Figure 4).
• By systematically analyzing the response, we were able to identify the important factors. In this case, it turned out that there were significant and important high-order interactions, many involving method and tail. While this complicated the interpretation somewhat, the end result was an analysis in which we are confident we have found the most important effects.

• The presence of large interactions between the four factors method, tail, sigma, and p0 (see the Supplementary Material) suggests that some combinations of method and tail may be more stable across uncontrollable factors. As sigma and p0 increase, one-sided tests deviate more from the nominal level (“cone shape” of Y vs. p0 in Figure 1 of the Supplementary Material, with more deviations for large sigma).

5. EXAMPLE: PLANNING A STATISTICAL LEARNING SIMULATION STUDY FROM SCRATCH

In this section, the five stages of experimentation are illustrated using a study that investigates the predictive accuracy of two statistical learning models. The purpose of the illustration is to highlight how the five stages can be considered in practice. A fractional factorial design is introduced as a way to reduce the number of runs.

5.1. The Five Stages, with a Fractional Factorial Experiment

(i) Problem definition

We examine the predictive accuracy of possible models for supervised learning. A variety of response variables could be used to measure predictive accuracy. We focus on $R^2$, the percent variation explained, evaluated on large test sets.

(ii) Planning

To begin, a cause-and-effect diagram (Figure 6) was constructed to organize the list of potential factors. While one can imagine many potential factors to include, an experiment based solely on those in the Figure 6 can get large. Choices must be made to keep the run size manageable. To keep things simple, for example, two statistical learning procedures (lasso and random forests) were compared and only a normal population was used. We eventually chose seven factors, each with two levels, and excluded or held fixed all other factors (e.g., fixed coefficients were used and outliers were not investigated). The factors are summarized in Table 5.

We now discuss the choice of levels for each of the factors. In all simulations, training and test data are generated from a linear model with normal errors:

$$y = X'\beta + \epsilon, \quad \epsilon \sim IID \ N(0, \sigma^2),$$

where $X$ is a $p$-vector of predictor variables formed from $q$ measured variables ($q < p$). The predictors in the linear model will include linear terms equal to individual measured variables and product terms formed by multiplying together two measured variables. Thus, there is a total of $p = q + \binom{q}{2} = q(q + 1)/2$ predictor variables. Some elements of the coefficient vector $\beta$ are exactly 0, corresponding to inactive predictors.

The two statistical learning models are the following:
Table 5: Statistical learning study: description and levels of the factors.

| Factor | Description                                      | Levels                |
|--------|--------------------------------------------------|-----------------------|
| model  | Statistical learning model                       | Lasso, random forest  |
| n      | Training sample size                             | 250, 1000             |
| q      | Number of measured variables                     | 20, 50                |
| ENE    | Expected number of active effects                | 10, 20                |
| beta.mu| Coefficient values of active effects             | 1, 3                  |
| sigma  | Error standard deviation                         | 0.5, 2                |
| x.cor  | Correlation parameter for predictors             | 0, 0.8                |

1. Least squares regression using the lasso and potential predictors consisting of all possible linear terms and product terms (total of \(q(q + 1)/2\) predictors);
2. Random forest regression using \(q\) measured variables as predictors.

Since the form of the lasso model matches the data-generating mechanism, we expect it to outperform random forests.

Both procedures have tuning parameter(s) that must be chosen. In the case of lasso regression, there is a single penalty parameter \(\lambda\), which will be chosen by running 10-fold cross-validation within the training set. For random forests, only the \(m_{try}\) parameter (the proportion of
predictors that are randomly sampled when growing each branch of a tree) is important. Pilot runs determined that holding fixed \( n_{\text{tree}} = 500 \) and unbounded \( \text{maxnodes} \) were reasonable choices. The \( \text{mtry} \) parameter will be chosen from a grid of proportion values (0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 1), using the out-of-bag data.

The sample size of the training set is \( n \). The nonzero elements of \( \beta \) will all be equal to \( \beta_{\text{mu}} \). The error standard deviation \( \sigma \) corresponds to \( \sigma \) in (1). The \( q \) measured variables are simulated as multivariate normals with mean vector 0 and covariance matrix \( \Sigma \). The diagonal elements of \( \Sigma \) are always 1, giving measured variables that are standard normal variates. The factor \( x.\text{cor} \) controls the correlation between the measured variables, with the correlation between variables \( i \) and \( j \) given by \( \Sigma_{i,j} = \Sigma_{j,i} = \exp(\log(x.\text{cor})|i - j|) \). This decaying correlation gives a (maximum) correlation of \( x.\text{cor} \) between measured variables with adjacent indices (e.g., \( |i - j| = 1 \)) and smaller correlation for other pairs of measured variables. This formulation of the measured variables results in product terms having mean 0 and variance 1, so that scaling of the predictors is not needed.

Some elements of the regression coefficient vector \( \beta \) were set to zero. Rather than having a fixed pattern, this was done randomly. An “effect heredity” prior distribution (Chipman, 1996) specifies a prior probability \( \pi \) that a main effect will be active, assuming that activities of main effect are independent of each other. Conditional on whether “parent” main effects \( A \) and \( B \) are active, the probability that the \( AB \) product term is active is given as

\[
P(AB \text{ active} | \text{activity of } A, B) = \begin{cases} 0 & \text{if both } A, B \text{ inactive}, \\ c_1 \pi & \text{if exactly one of } A, B \text{ active,} \\ c_2 \pi & \text{if both } A, B \text{ active.} \end{cases}
\]

For given values of \( q, \pi, c_1, \) and \( c_2 \), the expected number of active effects of each type (i.e., main effects, two-way interactions with two active parents, etc.) can be calculated. In our experiment, \( \pi, c_1 \) and \( c_2 \) were chosen so that

- a total of \( \text{ENE} \) effects are expected to be active,
- \( \text{ENE}/2 \) main effects are expected to be active,
- \( \text{ENE}/4 \) two-way interactions with two active parents are expected to be active, and
- \( \text{ENE}/4 \) two-way interactions with one active parent are expected to be active.

The factor \( \text{ENE} \) is set at a value chosen as part of the experimental design. For given values of \( \text{ENE} \) and \( c_2 \), and using the 50/25/25 split of active effects specified above, the parameters \( \pi, c_1, c_2 \) take unique values and can be easily calculated. Derivations follow Bingham & Chipman (2007) and are sketched in the Supplementary Material. We draw from the prior for each training set, giving random numbers of active main effects and one- and two-parent two-factor interactions.

Instead of running a full factorial \( 2^7 \) design (with 128 runs), we introduced fractional factorial designs as a way of saving computational effort. In this example, we chose a design with 32 runs. Fractionating a two-level, full factorial design is a well-studied problem and is a standard topic in introductory DAE texts (e.g., Box, Hunter & Hunter, 2005). A two-level fractional factorial is typically constructed by selecting rows based on “generators,” which are identities that apply to the columns of the design matrix. To simplify notation, we label the seven factors in our experiment as \( n = A, q = B, \text{ENE} = C, \beta_{\text{mu}} = D, \sigma = E, x.\text{cor} = F, \) and \( \text{model} = G \), and assume each two-level factor is coded with levels \(-1\) and \(+1\). The generators we will use are \( ABCE = +1 \) and \( BCDF = +1 \) (this is a resolution IV design). Of the 128 runs of the full factorial (without replication), 32 runs satisfy both these conditions, making this a quarter fraction of the full factorial design. For these runs, the generators will imply aliases...
between effects. That is, an effect estimate will actually correspond to the sum of several effect 
estimates. For example, the AB interaction will be aliased with the CE interaction (implied by 
the first generator). The AB interaction will also be aliased with other higher order terms. The 
generators imply the following aliasing patterns in this case:

1. Main effects A–F are each aliased with two different three-way interactions, and higher 
order terms. Main effect G is aliased with terms of order 5 and higher.
2. Two-way interactions that do not involve G are aliased in pairs, e.g., AB = CE.
3. Two-way interactions involving G are aliased with four-way interactions and higher order 
interactions.

This fractional factorial was chosen because the factor model (G) is of interest as the only 
control factor in TRPD. Similar to the analysis of the KMM study in Section 4, control by noise 
interactions give insight into the robustness of the models with respect to uncontrollable factors. These generators enable estimation of the main effect for G and two-way interaction effects 
involving G, assuming that fourth and higher order interactions are negligible. This does come 
at a cost, namely the inability to disentangle two-way interactions not involving G.

(iii) Execution

Pilot studies were used to prototype the code and investigate factor levels. Random forests 
required more compute time, so during prototyping we temporarily set ntree = 50. Table 5 
contains the levels settings for the final experiment. The original choices for the training sample 
size were \( n = 1000 \) and 10,000, but they were large enough that non-meaningful differences in 
the response variable was observed. Instead, we changed the levels to \( n = 250 \) and 1000.

The 32-run quarter fraction design was executed. Every run was a completely independent 
draw from the population model. Runs were parallelized, using two cores of an eight-core laptop 
and taking under 30 min to execute.

(iv) Analysis

The response analyzed was a logistic transform of the test \( R^2 \). That is, we defined \( R^2 = 1 - \frac{SSE}{SSTo} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \), taking sums over the test set, and analyzed as 
our response

\[
\log \left( \frac{R^2}{1 - R^2} \right),
\]

using an ANOVA model. Preliminary analysis of an untransformed \( R^2 \) gave unsatisfactory 
residual diagnostics and predictions of \( R^2 \).

An advantage of using two-level designs is that the simpler half-normal plot (Daniel, 1959) 

can replace an ANOVA table as the primary analysis technique. In balanced designs like the ones 
we have discussed, all effects are estimated by contrasts with equal variance. Effect estimates 
for terms that are inactive will all be normally distributed with mean 0. The half-normal plot 
is a quantile–quantile plot of absolute effect estimates against corresponding quantiles of a 
half-normal distribution (i.e., the absolute value of a standard normal). Effects that are small will 
fall on a straight line, while large effects will be larger than this line (and lie to the right of the 
plots in our figures).

The half-normal plot in Figure 7 suggests that between 6 and 10 effects are important. The 
large effects are labelled, while the remaining smaller, unlabelled effects appear to form a linear 
trend. The six largest effects are model, sigma, beta.mu, sigma:model, x.cor, and
FIGURE 7: Statistical learning study: Half-normal plot of estimated effects. Effects are the estimated mean change in response when the corresponding two-level factor changes from $-1$ to $+1$.

FIGURE 8: Statistical learning study: (a) main-effect plot and (b,c) select interaction plots.

The four large main effects are aliased with three-way interactions, while the two interactions involving model are aliased with four-way interactions.

The dominant effect is model, with other large effects including beta.mu, sigma and x.cor, and interactions beta.mu:model and sigma:model. These effects are represented visually in Figure 8a–c.

Although the response is a transformation of $R^2$, it is still a “larger the better” measure. With this in mind, the largest main effects and their interpretations are
• model, with lasso performing much better than random forests;
• sigma, with better predictive accuracy when the training set has less noise ($\sigma = 0.5$);
• beta.mu, with better predictive accuracy when the coefficients are larger ($\beta_{\mu} = 3$); and
• x.cor, with better predictive accuracy when the measured variables are correlated ($r = 0.8$). This is likely because the test set has the same correlation patterns among the measured variables, and the performance measure is predictive accuracy, not selection of the correct subset of predictors.

In the interaction plots, the largest effects involve model and one other term. Keeping in mind that a large interaction will appear as a large change in slope, rather than mean level, the interaction effects and their interpretations are

• beta.mu:model, with lasso models seeing a larger gain in predictive accuracy than random forests when the signal levels are high ($\beta_{\mu} = 3$). That is, the slope of the lasso line is much steeper than that of random forests;
• sigma:model, with lasso models seeing a larger loss in predictive accuracy than random forests when the noise levels are high ($\sigma = 0.5$). That is, the slope of the lasso line is negative and larger than the slope of the random forests line.

(v) Conclusions
In this experiment, the lasso is the clear winner. This should not be surprising since the form of the population model matches the lasso regression. Considering TRPD, interaction plots (Figure 8b,c) show less variation in the performance of random forests than in the lasso, across levels of the noise factors $\beta_{\mu}$ and $\sigma$. But even though the lasso model has more variable performance, it is always superior to random forests. The better performance (higher $R^2$) of lasso is more important than its increased sensitivity to noise factors, and so for this study, TRPD turns out to be a secondary consideration.

5.2. Other Remarks
Other experimental designs and their analysis were considered but not reported here. These include two replicates of the above design and a $2^7$ full factorial with one or two replicates. Half-normal plots and ANOVA tables in the Supplementary Material demonstrate that for this application, the conclusions from these other three designs would be effectively the same.

Fractionated designs will imply that some effects are aliased. In this case, the aliasing did not impact the conclusions. In cases where aliasing was a problem, additional runs could be later added to reduce aliasing due to fractionation.

A different fractional factorial could have been chosen, with fewer aliases between the two-way interaction terms. Doing so would treat effects more symmetrically, rather than prioritizing effects involving the only control factor, method (labelled G). For example, the generators ABCF = +1 and ABDEG = +1 give a resolution-4 design but with just 6 out of 21 two-way interactions aliased with a two-way interaction (compared to 15 of 21 in the design we chose). However, TRPD analysis of such designs is more difficult. See Nair et al. (1992) discussion of crossed arrays for further details of this trade-off.

It may seem ridiculous to have so few runs and no replication. Indeed, replication and a full factorial may be reasonable and are explored in the Supplementary Material. We chose to present the 32-run results to underscore the potential to discover main results with far less computation than usual.
6. CONCLUSION

This is not rocket science. The tools we have used would usually be covered in an undergraduate course in DAE, with perhaps the exception of TRPD. We hope that readers will decide that it is obvious (and straightforward) to design and analyze statistical studies using statistical tools and that there are real benefits to doing so.

A systematic approach like this formalizes some of the steps that would normally be carried out, such as identifying relevant factors to consider and appropriate levels. But other parts are not seen in most studies. Even if a full factorial experiment is run, an ANOVA combined with graphical exploration of effects should make it easier to discover all the important effects. In circumstances where there are many factors that are important to consider in the study, fractional factorial designs offer the possibility of exploring them all, rather than artificially eliminating some factors in order to design a study that can be run. The TRPD framework formalizes the common objective that, when choosing among the models being compared, it is desirable to pick a method that is stable (i.e., robust) across the scenarios we cannot control.

6.1. Randomization Restrictions to Reduce Noise, Blocking, and Random Effects

Our treatment (and both examples) implicitly assume that the random errors are independent and identically distributed across runs. For example, in the statistical learning study, every run in our design drew a separate, independent realization of data from the population model. In that example, a natural temptation would be to pair together runs that are the same except for the learning method used (lasso vs. random forests). So, the 32 runs used in our example could be grouped into 16 pairs, with 16 different datasets (rather than 32). Both methods would be applied to the 16 datasets, yielding 32 runs. One might argue that it would be a fairer comparison of the two models to have them analyze exactly the same data. This would be equivalent to adding 16 blocks to the design. Such a design could be analyzed via block terms or a random effects model. Essentially, we would be analyzing the performance difference between the two models as a function of the other six factors in the study.

Randomization restrictions will complicate analysis and may not be worth the extra effort. The biggest win is likely to come from the use of a meta-model, which shifts attention from standard errors on cells to (much smaller) standard errors on effect estimates. Randomization restrictions and an analysis that accounts for them may further reduce standard errors in some cases.

6.2. Take-Homes

We conclude with a short list of recommended techniques:

- Use cause-and-effect diagrams to list many candidate factors and do not be afraid to consider many.
- Plan a full or fractional factorial experiment.
- Carry out a pilot study.
- Use fewer replicates than you think you need. You can always add more later.
- Use meta-models like ANOVA to identify the important effects, including interactions. Effect estimates will have smaller standard errors than individual table cells.
- If factors can be characterized as controllable and uncontrollable, use TRPD to explore control factor settings that give desired values of the QoI with low variation.

ACKNOWLEDGEMENTS

We are grateful to Kalimuthu Krishnamoorthy for providing additional detail on the motivating example, to Pritam Ranjan, Andrea Benedetti, Nathaniel Stevens, and Andrew Gelman for
discussions on earlier versions of this work, and to Jock MacKay for suggesting the use of DAE to analyze simulation studies, during the first author’s doctoral research. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

DATA AVAILABILITY STATEMENT

All computations were carried out in R (R Core Team, 2022). The DoE.base package (Grömping, 2018) was used for half-normal plots. Code for the analysis of the KMM study and the statistical learning simulation study is provided as the Supplementary Material. Code (including any updates made after publication) is also available in a public repository on GitHub, at https://github.com/hughchipman/TablesAsDesigns.

REFERENCES

Bingham, D. R. & Chipman, H. A. (2007). Incorporating prior information in optimal design for model selection. Technometrics, 49, 155–163.

Box, G. E., Hunter, W. H., & Hunter, S. (2005). Statistics for Experimenters, 2nd ed., John Wiley & Sons, New York.

Chipman, H. (1996). Bayesian variable selection with related predictors. Canadian Journal of Statistics, 24, 17–36.

Culpepper, S. & Aguinis, H. (2011). Using analysis of covariance (ANCOVA) with fullible covariates. Psychological Methods, 16, 166–178.

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

Dolatsara, H. A., Chen, Y.-J., Leonard, R. D., Megahed, F. M. & Jones-Farmer, L. A. (2021). Explaining Predictive Model Performance: An Experimental Study of Data Preparation and Model Choice. Big Data. ahead of print, http://doi.org/10.1089/big.2021.0067

Gelman, A., Pasarica, C., & Dodhia, R. (2002). Let’s practice what we preach: Turning tables into graphs. The American Statistician, 56, 121–130.

Grömping, U. (2018). R package DoE.base for factorial experiments. Journal of Statistical Software, 85, 1–41.

Harwell, M., Kohli, N., & Peralta, Y. (2017). Experimental design and data analysis in computer simulation studies in the behavioral sciences. Journal of Modern Applied Statistical Methods, 16, 3–28.

Hoaglin, D. C. & Andrews, D. F. (1975). The reporting of computation-based results in statistics. The American Statistician, 29, 122–126.

Ishikawa, K. (1982). Guide to Quality Control, 2nd ed., Asian Productivity Organization, Tokyo.

Krishnamoorthy, K., Mallick, A., & Mathew, T. (2011). Inference for the lognormal mean and quantiles based on samples with left and right type I censoring. Technometrics, 53, 72–83.

MacKay, R. J. & Oldford, R. W. (2000). Scientific method, statistical method and the speed of light. Statistical Science, 15, 254–278.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38, 2074–2102.

Mukerjee, R. & Wu, C. F. J. (2006). A Modern Theory of Factorial Design, Springer, New York.

Nair, V. N., Abraham, B., MacKay, J., Nelder, J. A., Box, G., Phadke, M. S., Kacker, R. N. et al. (1992). Taguchi’s parameter design: A panel discussion. Technometrics, 34, 127–161.

R Core Team (2022). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. Multivariate Behavioral Research, 35, 137–167.

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

Received 22 October 2021
Accepted 17 March 2022

DOI: 10.1002/cjs.11719