



Power Considerations in Designed Experiments

Luyao Lin¹ · Derek Bingham¹ · Ryan Lekivetz²

Published online: 18 November 2019
© Grace Scientific Publishing 2019

Abstract

Designed experiments for ANOVA studies are ubiquitous across all areas of scientific endeavor. An important decision facing experimenter is that of the experiment run size. Often the run size is chosen to meet a desired level of statistical power. The conventional approach in doing so uses the lower bound on statistical power for a given experiment design. However, this minimum power specification is conservative and frequently calls for larger experiments than needed in many settings. At the very least, it does not give the experimenter the entire picture of power across competing arrangements of the factor effects. In this paper, we propose to view the unknown effects as random variables, thereby inducing a distribution on statistical power for an experimental design. The power distribution can then be used as a new way to assess experimental designs. It turns out that using the proposed expected power criterion often recommends smaller, less costly, experimental designs.

Keywords Statistical minimum power specification · Power distribution · ANOVA

1 Introduction

When planning an experiment, a critical first question to be addressed is whether or not the allocated run size is sufficient to meet the aim of the investigation. A natural way to assess an experimental design is to see if it achieves a satisfactory level of power for the effects of interest (e.g., Fairweather [3] and Gerrodette [4]).

The power of a statistical test is the probability of correctly detecting the presence of a significant effect (In most of the discussion throughout, we follow the more

Part of special issue guest edited by Pritam Ranjan and Min Yang—Algorithms, Analysis and Advanced Methodologies in the Design of Experiments.

✉ Derek Bingham
dbingham@stat.sfu.ca

¹ Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

² JMP Division, SAS Institute, Cary, NC 27513, USA

colloquial description of power as a percentage). In ANOVA studies, the power is a function of the minimum effect size to be detected, significance level of the hypothesis test, system variability and the design characteristics (e.g., number of treatment levels for a factor, number of replicates, etc.). The effect size, Δ , is typically defined as the difference between the largest and smallest level of treatment effects. For effects involving two-level factors, the effect size is directly related to the coefficients in a least squares regression model and can be easily utilized to calculate power [1]. On the other hand, for explanatory variables with more than two levels, there are infinitely many choices for treatment effects under a given effect size. It turns out that these differences in the assignment of treatment effects can have a large impact on the power for detecting significant effects of the desired magnitude.

A common approach for computing power in ANOVA studies, where the sum-to-zero constraint is used, is to set the treatment effect for two of the factor-level settings so that their difference is the minimum size the experimenter wishes to detect, and the remaining treatment effects are set to zero. This gives the lowest power among all specifications for the experimenter's desired difference in treatment effects (i.e., the effect size) and is known as the *minimum power specification*.

While conservative, the minimum power specification provides a lower bound for the power [1] for the design. Using this approach ensures that the statistical power of the experimental procedure is at least the level chosen by the experimenter. On the other hand, it is highly unlikely that the treatment effects are those of the minimum power specification (i.e., two large treatment effects, while the remaining effects are exactly zero). The practical result of using this specification is being too conservative in the assessment of the design and requiring a potentially costly increase in the run size to achieve the desired lower bound for power, or removing levels and factors from consideration to meet a certain level for the power.

In this paper, a new approach for the assessment of power in designed experiments is proposed. Instead of using a specific pattern for the treatment effects, we propose using a distribution for the treatment means. From this, a distribution for the power is induced. Expected values and percentiles of the power distribution can then be employed to help a researcher evaluate an experiment based on desired statistical power. This allows for more flexibility in assessing the experimental setup, and frequently results in a reduction in the required run size versus the common minimum power specification. By considering the distribution of the treatment effects, a practitioner can get a better assessment for the likelihood of detecting a significant effect over a range of plausible specifications, rather than focusing on a specific, potentially extreme, case.

This paper is outlined as follows. In Sect. 2, basic concepts of power analysis for ANOVA models are reviewed and illustrated with a motivating example. The "power distribution" is then proposed in Sect. 3, followed by examples in Sect. 4. Concluding remarks are made in Sect. 5.

2 Power in ANOVA Studies

We begin by introducing some of the issues related to power analysis using an application from the literature [6] where the chemical behavior of different combinations of material was studied. This design is used throughout to help illustrate the main ideas. The experiment was performed as a completely randomized design with five factors with two-, three- or four-level settings (see Table 1). The outcome of the experiment is the measurement of chemical species present in a closed container containing multiple materials. The corresponding main effects ANOVA model can be written as

$$y_{ij} = \mu + \tau_{A,i_A} + \tau_{B,i_B} + \tau_{C,i_C} + \tau_{S,i_S} + \tau_{T,i_T} + \epsilon_{ij}, \quad i = (i_A, i_B, i_C, i_S, i_T)' \quad (1)$$

where i is an index vector denoting the experimental treatment, the elements of i indicate the level setting for each factor, y_{ij} is the j th response for treatment i , μ is the overall mean, and $\epsilon_{ij} \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$ is the measurement error. Let $\tau_A, \tau_B, \tau_C, \tau_S$ and τ_T be the vectors of treatment effects with, for example, $\tau_A = (\tau_{A,1}, \tau_{A,2}, \tau_{A,3}, \tau_{A,4})'$.

For standard ANOVA models, there is an identifiability issue due to over-parameterization (e.g., see Dean et al. [2], Section 3.4). That is, not all elements of the τ 's are simultaneously estimable. To address this issue, constraints are placed on the model parameters. Throughout, the sum-to-zero constraints (e.g., see Dean et al. [2]) are assumed (i.e., the treatment effects for each factor sum to zero). Of course, there are other types of constraints (e.g., baseline constraints in Wu and Hamada [7]) that can be applied. It is worth noting that the hypothesis test for equality of treatments remains the same across different types of constraints, but the practical interpretation may be different.

Let $\theta = (\mu, \tau_{A,1}, \tau_{A,2}, \tau_{A,3}, \tau_{B,1}, \tau_{B,2}, \tau_{C,1}, \tau_{C,2}, \tau_{S,1}, \tau_{T,1})'$ denote the vector of estimable parameters for a main effects model. This parametrization defines the model uniquely as the zero-sum constraint implies that $\tau_{A,4} = -\tau_{A,1} - \tau_{A,2} - \tau_{A,3}$, $\tau_{B,3} = -\tau_{B,1} - \tau_{B,2}$ and $\tau_{C,3} = -\tau_{C,1} - \tau_{C,2}$. For the experimental design in Table 2 [6], the corresponding model matrix, \mathbf{X} , and column vector of responses, \mathbf{Y} , are used to obtain the ordinary least squares estimate of θ (i.e., $\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$). To test the significance of, for instance, factor A under $H_0: \tau_A = \mathbf{0}$ versus $H_a: \tau_A = \tau_A^a$, as an example, the F test statistic is

$$F_A = \frac{(\hat{\tau}_{A,1}, \hat{\tau}_{A,2}, \hat{\tau}_{A,3}) \Sigma_{\hat{\tau}_A}^{-1} (\hat{\tau}_{A,1}, \hat{\tau}_{A,2}, \hat{\tau}_{A,3})^T / \nu_1}{\sigma^{-2} (\mathbf{Y}^T \mathbf{Y} - \hat{\theta}^T \mathbf{X}^T \mathbf{X} \hat{\theta}) / \nu_2}, \quad (2)$$

Table 1 Material factors [6]

Factor	Treatment
Material A	a_0, a_1, a_2, a_3
Material B	b_0, b_1, b_2
Material C	c_0, c_1, c_2
Spacing S	s_1, s_2
Temperature T	t_1, t_2

Table 2 24-run design from Wendelberger et al. [6]

Run	A	B	S	C	T	Run	A	B	S	C	T	Run	A	B	S	C	T
1	0	0	0	0	0	9	2	2	0	1	0	17	1	1	1	0	1
2	0	1	0	1	0	10	3	0	0	0	1	18	1	2	1	1	0
3	0	2	0	2	1	11	3	1	0	1	1	19	2	0	1	0	0
4	1	0	0	1	1	12	3	2	0	2	1	20	2	1	1	1	1
5	1	1	0	2	0	13	0	0	1	1	1	21	2	2	1	2	1
6	1	2	0	0	0	14	0	1	1	2	0	22	3	0	1	1	0
7	2	0	0	2	0	15	0	2	1	0	1	22	3	1	1	2	0
8	2	1	0	0	1	16	1	0	1	2	1	24	3	2	1	0	0

where $\hat{\tau}_A$ is the least squares estimate of τ_A , ν_1 and ν_2 are the degrees of freedom, and $\Sigma_{\hat{\tau}_A}$ is the variance–covariance matrix of $(\hat{\tau}_{A,1}, \hat{\tau}_{A,2}, \hat{\tau}_{A,3})'$.

With \mathbf{X} and τ_A^a specified, along with the error variance, σ^2 , and significance level for the test, α , the calculation of power for factor A makes use of the probability distributions of F_A under H_0 and H_a in the following steps: (1) find the critical value, F_A^* , under the null hypothesis of the equality of treatment effects for factor A , from an F distribution with degrees of freedom $\nu_1 = 3$, $\nu_2 = 24 - 10 = 14$ under H_0 ; (2) compute the power as $P(F_A > F_A^* | H_a)$, where F_A follows a non-central F distribution of F_A , with non-centrality parameter $\phi_A = (\tau_{A,1}^a, \tau_{A,2}^a, \tau_{A,3}^a) \Sigma_{\hat{\tau}_A}^{-1} (\tau_{A,1}^a, \tau_{A,2}^a, \tau_{A,3}^a)'$ [5] under H_a (not all treatment effects are the same). Note that we have chosen factor A somewhat arbitrarily, and the calculations for the other factors in the material study can be made similarly.

Many different choices are available for τ_A^a under the sum-to-zero constraint and effect size Δ_A . Figure 1 depicts four examples for τ_A^a that satisfy the sum-to-zero constraints with $\Delta_A = 2$. In the far left panel of the figure, the first and last effects are set to -1 and $+1$, respectively, while the other two effects are set to zero. This corresponds to an extreme case where the effects are overall “minimal” (having as many effects of zero as possible). The second panel demonstrates a “maximal” effect pattern (having large absolute effects) where the first two effects are specified as -1 and the last two effects are set to $+1$. The specifications for the final two scenarios are examples of not-so-extreme treatment effects.

The different specifications have consequences in the assessment of the power for the experiment. In practice, the minimum power specification is commonly used since it offers a lower bound on the power for the specified effect sizes (e.g., Δ_A). While using the minimum power specification is tempting, the corresponding effects specification is unlikely to occur in practice and can lead to larger recommended run sizes than necessary. For instance, with effect size $\Delta_A = 2$ and measurement error variance $\sigma^2 = 1$, the power of factor A for design D (Table 2) under the minimum power specification is around 70%, which suggests more trials should be performed if one wishes to detect an effect size of 2 with higher power. In fact, obtaining a power over 90% under this specification requires a design that is twice as costly. On the other hand, if the experimenter uses the arrangement in the second panel of Fig. 1 (i.e., $\tau_A^a = (-1, -1, 1, 1)'$), the power for factor A is above 95%, indicating that

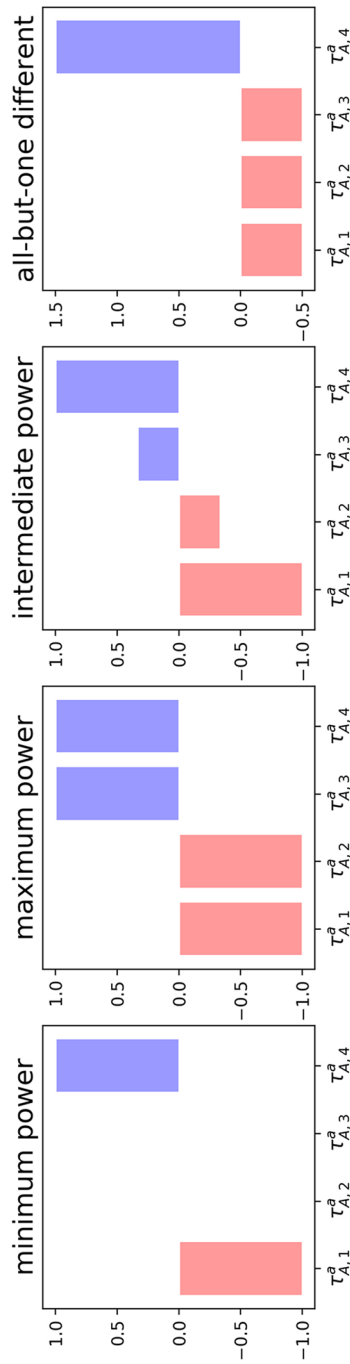


Fig. 1 Example of specifications for τ_A^α : the minimum power specification (71.50% power, first panel from the left); the maximum power specification (95.86% power, second panel from the left); intermediate power specification with even spaced τ_A^α (76.40% power, third panel from the left) and all-but-one same specification τ_A^α (88.47% power, fourth panel from the left)

the 24-run design would be sufficient. The patterns in the final two panels of Fig. 1 (i.e., evenly spaced and all-but-one same specifications) produce power of 76.40% and 88.47%, respectively.

The act of declaring power derived from any one specification of τ_A^a is potentially substantial. It also raises the question of whether performing the most costly experiment is the right thing to do in practical settings. It is desirable to incorporate the uncertainty of τ_A^a to produce power calculations of typical arrangements for the effects, and to move away from being overly conservative with specifications that describe only unlikely scenarios. In the next section, new methodology for assessing statistical power in designed experiments is proposed to address this issue.

3 The Power Distribution

In this section, we propose a new way to assess statistical power. Our approach is to define statistical power as a random variable, and instead of producing one fixed number, the probability distribution of the statistical power (the power distribution) is used to reflect possible specifications of the treatment effects. With the power distribution in hand, the expected value and percentiles of the power distribution, for example, can be used to assess experimental designs or to aid in run size selection. The proposed methodology can be employed used in place of, or as a complement to, the traditional power calculation.

It turns out that integration of the density for the power distribution is analytically intractable, and an approach is developed in Sect. 3.1 for fast numerical evaluation to obtain the expected value of power (or expected power). Exploration of the entire distribution of power requires Monte Carlo methods which can also be accomplished quickly and is discussed in Sect. 4.1 (Details on numerical evaluation and considerations for computing resources are presented in ‘‘Appendix C.’’) Throughout Sect. 3, the material example from Sect. 2 serves as an illustration of our methodology, and results are compared to that of different specifications to demonstrate the benefits of adopting the proposed approach.

We begin the introduction of power distribution by considering a general ANOVA setting. For a factor with k levels, denote by $\tau^a = (\tau_1^a, \tau_2^a, \dots, \tau_k^a)'$ the vector of treatment effects under the alternative hypothesis. To perform power calculations, the effect size, Δ , the residual variance, σ^2 and the significance level, α must be specified. Without loss of generality, it is assumed that $\tau_1^a, \tau_2^a, \dots, \tau_k^a$ are in non-decreasing order, where τ_1^a and τ_k^a are the minimum and maximum of all treatment effects, respectively. That is to say, $\tau_k^a - \tau_1^a = \Delta$, and $\tau_1^a \leq \tau_2^a \leq \dots \leq \tau_k^a$.

A multivariate probability distribution, $f_\tau(\cdot)$ with domain $[\tau_1^a, \tau_k^a]$, is used to reflect the experimenter’s belief about the size and potential arrangements of the remaining effects in τ^a (i.e., the elements of τ^a other than the extreme setting, τ_1^a and τ_k^a). The sum-to-zero constraint can be readily imposed by applying a shift to sampled τ^a under the factor effects model. That is, the constrained vector of treatment effect, denoted by τ^{a*} , is $(\tau_1^a - \delta, \tau_2^a - \delta, \dots, \tau_i^a - \delta, \dots, \tau_k^a - \delta)'$ with scalar $\delta = \sum_i \tau_i^a / k$. As a result, the statistical power function $\eta(D, \tau^a) = P(\text{reject } H_0 | H_a : \tau = \tau^{a*} \text{ is true})$ is a random variable whose

distribution will be used in assessing the power under all possible values of τ^a within the domain of $f_\tau(\cdot)$. In our framework, the minimum power specification is a special case that can be defined through $f_\tau(\cdot)$ by assigning point mass probability of 1 to $\{\tau_1^a = -\Delta/2, \tau_2^a = \tau_3^a = \dots = \tau_{k-1}^a = 0, \tau_k^a = +\Delta/2\}$, and probability of 0 to all other arrangements. Generally, $f_\tau(\cdot)$ puts a distribution on the model space for the treatment effects, and the minimum power specification is just one of many possible realizations from $f_\tau(\cdot)$.

Returning to the material example, we now examine the power distribution for factor A with the 24-run near-orthogonal design proposed in Wendelberger et al. [6] (Table 2) as an illustration. The first step is to specify the effect size of interest, the residual variance and the significance level. Here, we use $\Delta_A = 2$, $\sigma^2 = 1$ and $\alpha = 0.05$. Next, the distribution $f_\tau(\cdot)$ must be selected. We consider four specifications for $f_\tau(\cdot)$ (see histograms shown in the top row of Fig. 2). The first distribution (top left in Fig. 2) is a truncated normal distribution with mean zero and scale $\sigma_\tau = 0.33$. This distribution can be viewed as a relaxation of the minimum power specification where the density peaks at $\tau = 0$ and decreases as the effects deviate from 0. The second one is an example of asymmetric distribution (top row, second panel in Fig. 2); specifically, a truncated normal distribution on $[-1, 1]$ with $\mu_\tau = 0.5$ and $\sigma_\tau = 0.3$ is chosen. The third distribution (top row, second from the right in Fig. 2) is a uniform distribution on $(-1, 1)$, which can be thought of the uninformative choice. The final distribution (top right in Fig. 2) is mixture of normals with mean ± 1 and scale $\sigma_\tau = 0.3$ on $(-1, 1)$. It is chosen as a relaxed version of the maximum power specification.

From the four different distributions, samples of the treatment effects τ_A^a are drawn. For each sampled τ_A^a , the power for factor A can be calculated as described in Sect. 2. The corresponding histograms for power are presented in the second row in Fig. 2. The power using the conventional minimum power specification is indicated by the dotted black line, and the maximum power obtained by letting $\tau_A^a = (-1, -1, 1, 1)'$ is indicated by dashed blue line. The mean of the power distribution is identified by the solid red line.

Of course, under all four $f_\tau(\cdot)$'s, the power distribution (bottom row in Fig. 2) indicates that the power is always between the minimum and maximum power specification. When $f_\tau(\cdot)$ assigns less mass in the area near $\tau = 0$ (left to right), the power distribution shifts to the right indicating more power. That is, the probability of having higher power increases, and the expected value of the power distribution also becomes larger. This shift in the power distribution coincides with the intuition that effects with magnitudes near zero tend to lead to lower power.

Different metrics can be extracted from the power distribution that can be used to improve the overall assessment of statistical power for a design over and above the minimum power specification. For example, if an experimenter does not want to be too conservative or optimistic, and chooses a uniform distribution for $f_\tau(\cdot)$, they can use the expected power (slightly above 80% in the above example) for the experiment. Next, we discuss a method for fast evaluation of expected power.

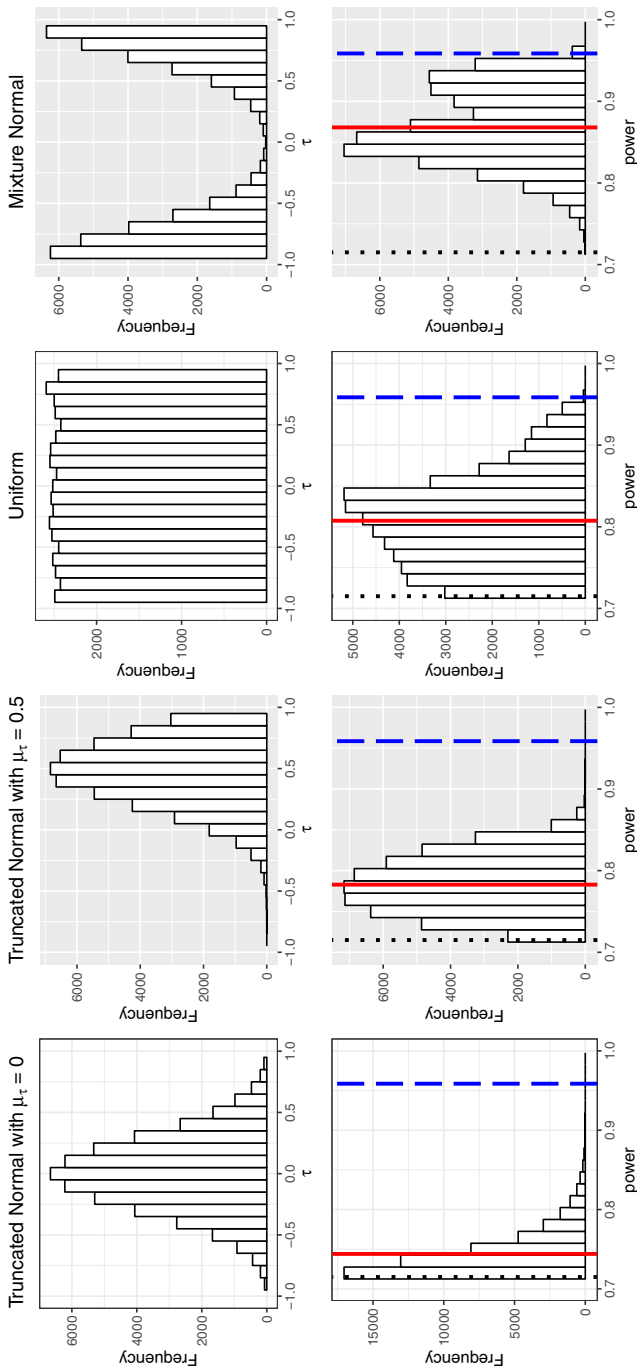


Fig. 2 Histograms for treatment effects (top row) and the power (bottom row) of factor A in the material experiment, with treatment effects drawn from (i) truncated normal centered at 0 (first column), (ii) truncated normal centered at 0.5 (second column), (iii) uniform (third column) and (iv) mixture of normals (fourth column). The effect size Δ_A is set to 2, and a sample size of 50,000 is used for each distribution. Dotted, solid and dashed lines correspond to the minimum, mean and maximum of the random power, respectively

3.1 Expected Power

Denote the power under a design, D , as $\eta(D, \boldsymbol{\tau}^a)$. Having defined the power distribution, we can evaluate the expected power and use this to assess a design. The expected value of the power distribution is

$$E[\eta(D, \boldsymbol{\tau}^a)] = \int \eta(D, \boldsymbol{\tau}^a) f_{\boldsymbol{\tau}}(\boldsymbol{\tau}^a) d\boldsymbol{\tau}^a. \tag{3}$$

Generally, this integral will not be tractable. The following theorem provides a way to evaluate the expected power of a design D given any distribution $f_{\boldsymbol{\tau}}(\cdot)$ for the treatment effects.

Theorem 4 *The expected power, $E[\eta(D, \boldsymbol{\tau}^a)]$, for a k -level factor and distribution $f_{\boldsymbol{\tau}}(\cdot)$ for the treatment effects can be obtained as*

$$E[\eta(D, \boldsymbol{\tau}^a)] = 1 - \sum_{s=0}^{\infty} I\left(\frac{(k-1)c}{n_e + (k-1)c} \middle| \frac{k-1}{2} + s, \frac{n_e}{2}\right) \cdot g(s), \tag{4}$$

where $I(\cdot)$ is the regularized incomplete beta function, c is the critical value of the test, $H_0: \boldsymbol{\tau} = \mathbf{0}$ versus $H_a: \boldsymbol{\tau} = * \boldsymbol{\tau}^a$ (sum-to-zero constrained version of $\boldsymbol{\tau}^a$), n_e is the degrees of freedom for the residuals, and

$$g(s) = \int \frac{e^{-\phi/2} (\phi/2)^s}{s!} \cdot f_{\boldsymbol{\tau}}(\boldsymbol{\tau}^a) d\boldsymbol{\tau}^a, \tag{5}$$

where ϕ is the non-centrality parameter for the non-central F distribution that the test statistic follows under H_a .

Proof See ‘‘Appendix A’’. □

It is worth noting that the above theorem applies to normal ANOVA models in general, and not just to the main effect model. To maintain the desired effect size, the domain of $f_{\boldsymbol{\tau}}(\cdot)$ needs to be studied beforehand for $\boldsymbol{\tau}^a$ ’s that require more than one sum-to-zero constraint, such as the interaction effects in two-way ANOVA models. One of many approaches is to first use Monte Carlo method to generate factor effects and then to only keep candidate that satisfies all constraints.

For simplicity, denote $I^*(s)$ as the incomplete beta function in Eq. (4),

$$I\left(\frac{(k-1)c}{n_e + (k-1)c} \middle| \frac{k-1}{2} + s, \frac{n_e}{2}\right),$$

since the other quantities (k , c and n_e) are fixed for a given design. The expected power then becomes

Table 3 Expected power for the near-orthogonal design in Table 2 with $\Delta_A = \Delta_B = \Delta_C = 2$, $\sigma^2 = 1$

Pattern for τ^a	Factor A (%)	Factor B (%)	Factor C (%)
Minimum power	71.50	89.82	89.82
Expected power ($\mu_\tau = 0, \sigma_\tau = 0.33$)	74.78	90.78	90.78
Expected power ($\mu_\tau = 0.5, \sigma_\tau = 0.33$)	78.39	92.16	92.16
Expected power (uniform)	80.83	92.44	92.44
Expected power (mixture of normals)	86.74	94.39	94.39
All-but-one same	88.47	96.41	96.41
Maximum power	95.86	96.41	96.41

Distributions for τ considered include truncated normal with $\mu_\tau = 0$ and $\mu_\tau = 0.5$, uniform and mixture of normals. Deterministic patterns for the treatment effects include minimum/maximum power specification and all-but-one the same pattern

$$E[\eta(D, \tau^a)] = 1 - \sum_{s=0}^{\infty} I^*(s) \cdot g(s). \quad (6)$$

where an infinite series indexed by s needs to be evaluated.

The analytic form of the summation in Eq. (4) is not tractable. Fortunately, fast numerical integration can be performed for $g(s)$ via Monte Carlo. The sum can then be approximated by truncating the index, s , at some large number M (see ‘‘Appendix C’’ for theoretical justification and numerical illustrations).

Evaluation of the expected power for all factors in the material study, with distributions for the $f_\tau(\cdot)$ ’s, is shown in Table 3 with $M = 30$. The minimum power and maximum power are also included for comparative purposes. It can be seen from the results that the expected power is often much higher than the minimum value. Realistically, this implies that a design might be much more effective than the minimum power specification would lead one to believe. The expected power provides an additional assessment of designs and shows the utility of cost-effective designs which are otherwise considered inadequate if only minimum power is used. It is worth noting that as the probability of having treatment effects near zero gets lower, the expected power increases. In practice, it is unlikely to have exactly two of the treatment effects being large and all of the remaining treatment effects having magnitudes near zero. Thus, assessing a design using a uniform distribution for $f_\tau(\cdot)$ and computing the expected power can give a more sensible assessment of statistical power.

Aside from the expected value, other quantities of the power distribution might also be of great interest to experimenters. One might want to know, for example, what is the probability that the power is above 80%? We will address this type of question next, by examining the entire power distribution.

3.2 Other Metrics

Besides the expected value, one might also be interested in percentiles to gain more insight on the distribution of $\eta(D, \tau^a)$. Questions of interest may include (1) what is the chance that the power of design D for detecting factor A is greater than 80%? or (2) with 95% confidence, what is lower threshold for the power?

3.2.1 Tail Probabilities

To address the first question, it is important to translate it into a probability statement. The probability of power being greater than some threshold β can be expressed as $P(\eta(D, \tau^a) > \beta)$, which is the tail probability of power distribution evaluated at β . Using the notation from Eq. (4), this tail probability can be written as

$$P(\eta(D, \tau^a) > \beta) = E(\mathbf{1}_{\{\eta(D, \tau^a) > \beta\}}) = \int_{\eta(D, \tau^a) > \beta} f_{\tau}(\tau^a) d\tau^a.$$

The above integral is typically analytically intractable. However, we will demonstrate how to approximate tail probability via Monte Carlo with the material example.

Figure 2 displays random samples of τ_A^a drawn from the three distributions we have considered. With these samples, an empirical power distribution is then constructed. We revisit these histograms in the first row of Fig. 3, followed by corresponding empirical CDFs in the second row. We can address the question on the chance of power for factor A being over 80%. From the empirical CDFs in the second row of Fig. 3, it can be seen that the mixture of normals (right column) for τ_A^a suggests a chance as high as 0.9 of having a power over 80%, while the truncated normal distribution (left column) suggests merely a 0.1 chance. An intuitive interpretation can be made by recalling that the truncated normal distribution is considered as a relaxation of the minimum power specification, and the mixture of normals (right column) serves as a relaxation of the maximum power specification. The uniform distribution (middle column) produces an intermediate chance of almost 0.5.

3.2.2 Upper Percentile of the Power Distribution

For the second question raised, we are interested in finding the lower threshold for the power function with a given confidence level. This question can be addressed by the upper percentile of the power distribution. The upper percentile of order p for $\eta(D, \tau^a)$ is defined as

$$q_p = \max\{b \in \mathcal{R} : \mathbb{P}(\eta(D, \tau^a) \leq b) \leq 1 - p\}. \quad (7)$$

The lower threshold for the power function with a 0.95 confidence level is therefore $q_{0.95}$, and it can be obtained from empirical CDFs as ones shown in Fig. 3. Alternatively, one could use simulation to perform numerical integration and search for the appropriate q_p that satisfies

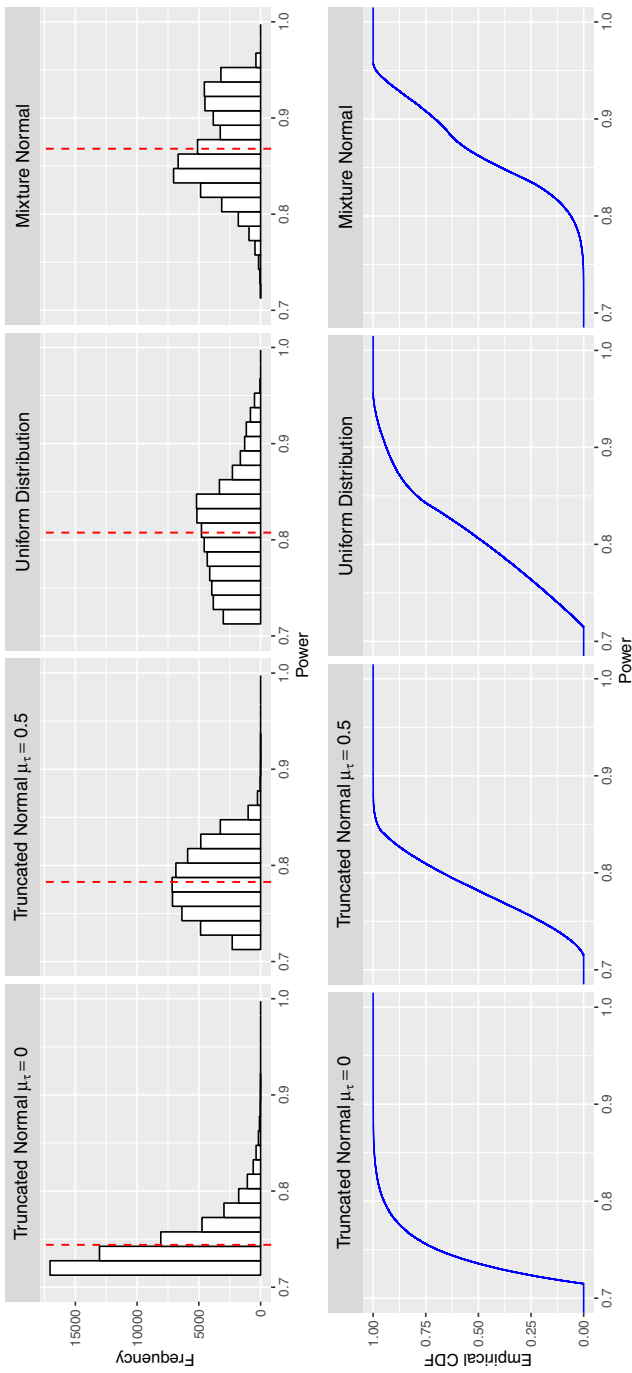


Fig. 3 Histograms and empirical CDFs for the power of factor A in the material experiment, with treatment effects drawn from truncated normal distribution centered at 0 (first panel), truncated normal distribution centered at 0.5 (second panel), uniform distribution (third panel) and mixture of normals (right) with $\sigma_\tau = 0.33$ (last panel). The effect size Δ is assumed to be 2, and sample size of 50,000 is used for each distribution

Table 4 Upper percentiles of the power for factor A under different distributions

Distribution for τ^a	$q_{0.5}$ (%)	$q_{0.8}$ (%)	$q_{0.9}$ (%)	$q_{0.95}$ (%)
Truncated normal $\mu_\tau = 0$	74.03	72.31	71.88	71.68
Truncated normal $\mu_\tau = 0.5$	78.14	74.98	73.73	72.87
Uniform on $(-\Delta/2, \Delta/2)$	80.74	75.55	73.59	72.55
Mixture of normals $\sigma_\tau = 0.33$	85.86	82.14	80.19	78.62

$$E[\mathbf{1}(\eta(D, \tau^a) \leq q_p)] = \int \mathbf{1}(\eta(D, \tau^a) \leq q_p) \cdot f_\tau(\tau^a) d\tau^a \leq 1 - p.$$

Different q_p 's of the power for factor A (with $p = 0.5, 0.8, 0.9$ and 0.95) under different distributions for τ^a are shown in Table 4. As can be seen, more optimistic distributions on τ^a_A (e.g., mixture of normals) lead to a large lower bound on power across all confidence levels.

4 Examples: Choosing the Run Size

To illustrate the use of the power distribution in planning experiments, two examples are presented in this section. Full factorial designs are considered to gain insights on how the power distribution can be used to assess experimental design from a cost or run size perspective.

4.1 One-way ANOVA

At the beginning of an experiment, an experimenter is faced with balancing available resources with the desire to detect significant effects. In this section, a balanced one-way ANOVA model is used as a simple illustration of how the proposed power distribution method can be adapted to choose the run size.

Consider a one-way ANOVA model with a factor with k levels

$$y_{i,j} = \mu + \tau_i + \epsilon_{i,j}, \tag{8}$$

where $i = 1, 2, \dots, k, j = 1, 2, \dots, r$ where r is the number of replicates of a full factorial design. Suppose that the experimenter has an initial estimate of the random error and specifies that $\epsilon_{i,j} \sim N(0, \sigma^2)$.

To assess the power using the proposed approach, the distribution of the effects, the minimum size of the effect, Δ and the significance level, α , must be specified. In this example, a uniform distribution on $[-\Delta/2, \Delta/2]$ is assumed for τ^a . This assumes that other than the two effect levels that are set with a difference of Δ , the distribution of remaining effects is uniform in this range between. The effect size is chosen to be $\Delta = 1$ and 2 . Full factorial designs with factors having $k = 4, 5, 6, 7$ levels are considered using error an error variance of $\sigma^2 = 1$ and significance level $\alpha = 0.05$.

Table 5 Required run size for balanced one-way ANOVA models with $\Delta = 2$, $\tau \sim U[-1, 1]$ and $\sigma^2 = 1$

Factor levels	Pattern/quantity	Specified power level			
		90%	80%	70%	50%
4	Minimum power	36	28	24	16
	Upper percentile $q_{0.9}$	32	28	24	16
	Expected power	28	24	20	16
	Maximum power	20	16	16	12
5	Minimum power	45	35	30	25
	Upper percentile $q_{0.9}$	40	35	30	20
	Expected power	35	30	25	20
	Maximum power	25	20	20	15
6	Minimum power	60	48	42	30
	Upper percentile $q_{0.9}$	48	42	36	24
	Expected power	42	36	30	24
	Maximum power	24	24	18	18
7	Minimum power	70	56	49	35
	Upper percentile $q_{0.9}$	56	49	42	28
	Expected power	49	42	35	28
	Maximum power	28	28	21	21

For comparison, we consider four approaches for the assessment of power and the choice of experiment run size: (1) minimum power; (2) maximum power; (3) upper 90th percentile of the power distribution; and (4) expected value of the power distribution. These represent the most conservative and liberal choices as well as two other potential applications of the proposed methodology.

The minimum run size results when the experimenter under each criterion, for a desired power of at least 50%, 70%, 80% and 90%, is summarized in Tables 5 and 6. The minimum run sizes are shown in columns 3–6. Looking at column 3 (which corresponds to a desired power of 90%) of Table 5, for example, we see that a four-level factor requires at least 36 runs to achieve at least 90% power (i.e., minimum power specification) and 20 runs (i.e., maximum power for this run size) if one is not worried about being too optimistic. The upper 90th percentile of the power distribution criteria requires a 32-run design, and the expected power criteria requires 28-run design.

The same patterns are consistent across different effect sizes ($\Delta = 1$ and $\Delta = 2$) and across different numbers factor levels. Overall, using the power distribution instead of minimum power allows one to more thoroughly investigate the options facing the experimenter. In many cases, there is an opportunity to reduce the run size of the experiment. For each of the power thresholds in Tables 5 and 6, as the number of levels increases, the differences in the recommended run sizes of the proposed approach and the conventional minimum power specification generally get

Table 6 Required run size for balanced one-way ANOVA model with $\Delta = 1$, $\tau \sim U[-0.5, 0.5]$ and $\sigma^2 = 1$

Factor levels	Pattern/quantity	Specified power level			
		90%	80%	70%	50%
4	Minimum power	120	92	76	52
	Upper percentile $q_{0.9}$	116	88	72	52
	Expected power	100	76	64	44
	Maximum power	66	48	40	28
5	Minimum power	160	125	105	70
	Upper percentile $q_{0.9}$	145	115	95	65
	Expected power	120	95	80	55
	Maximum power	70	55	50	35
6	Minimum power	204	162	132	90
	Upper percentile $q_{0.9}$	174	132	114	78
	Expected power	144	108	90	66
	Maximum power	72	60	48	36
7	Minimum power	252	203	168	112
	Upper percentile $q_{0.9}$	196	154	126	81
	Expected power	161	126	105	70
	Maximum power	84	63	56	42

larger. A similar conclusion arises when comparing the expected power and the 90th percentile criteria. For example, looking at the 80% power threshold of Table 5, for a seven-level factor, one would be satisfied with a experiment of 42 runs based on expected power, while the upper 90% percentile suggests that 49 is the appropriate size. Either quantity can help one avoid being too optimistic (28 runs from maximum power) or too conservative (56 runs from minimum power). In cases with higher desired precision (i.e., a smaller Δ), more substantial gains can be observed for the run size using the proposed methodology when compared with the minimum power specification. We recommend using the expected power in settings where resources are in short supply as a guide for sample size selection as a balance between the two extremes. The researcher can now make an informed decision on which quantity and threshold are suitable for the individual setting, and the proposed methodology provides a more thorough, and pragmatic, understanding of a design’s statistical power.

4.2 Two-way ANOVA

The power distribution methods can be easily scaled up to studies with more than one factor. Consider a two-way ANOVA model with main effects and an interaction effect given by:

$$y_{ij} = \mu + \tau_{A,i_A} + \tau_{B,i_B} + \tau_{AB,i_{AB}} + \epsilon_{ij}, \quad i = (i_A, i_B, i_{AB})'$$

Table 7 Power for the two-way ANOVA model with full factorial designs, τ following a uniform distribution, experimental noise $\sigma^2 = 1$ and $\Delta = 1, 2$

Effect	Δ	Repli- cates (%)	Minimum (%)	Upper $q_{0.8}$ (%)	Expected (%)	Maximum (%)
Main (A)	1	2	20.85	22.48	25.41	38.69
		4	47.13	50.89	56.90	79.11
		6	67.98	72.19	77.87	94.46
	2	2	69.27	73.36	78.81	94.67
		3	92.17	94.30	96.25	99.83
		4	98.29	98.97	99.40	100.00
Main (B)	1	2	33.27	33.66	36.48	42.87
		4	67.85	68.46	72.45	80.72
		6	86.59	87.04	89.71	94.69
	2	2	89.09	89.50	91.86	96.01
		3	98.90	98.99	99.35	99.86
		4	99.90	99.92	99.95	100.00
Interaction	1	2	9.98	13.21	16.22	34.44
		4	20.42	31.09	40.03	80.34
		6	31.99	49.30	60.72	96.07
	2	2	29.81	45.88	56.70	93.83
		3	54.34	76.94	85.06	99.91
		4	73.12	91.84	95.17	100.00

where factors A and B both have three levels with $i_A \in \{1, 2, 3, 4\}$, $i_B \in \{1, 2, 3\}$, and the error term is $\epsilon_{ij} \stackrel{i.i.d.}{\sim} N(0, 1)$. The interaction effect τ_{AB} has 12 levels, with six degrees of freedom for full factorial designs under the sum-to-zero constraint.

Again, we begin by specifying the minimum size of all effects as $\Delta_A = \Delta_B = \Delta_{AB} = 1$ or 2, and the significance level α as 0.05. A uniform distribution for the appropriate domain imposed by the effect size and the sum-to-zero constraint is assumed for the remaining levels of τ_A^a , τ_B^a and τ_{AB}^a that are not fixed by the effect size. Note that for the interaction effect, there are three sum-to-zero constraints required. The domain of the uniform distribution for τ_{AB}^a is obtained through Monte Carlo method. In this illustration, full factorial designs with different numbers of replicates are considered. For the main effects τ_A and τ_B and interaction effect τ_{AB} , the power is assessed with the four distributions discussed in the one-way ANOVA examples, and compared with each other for different run sizes (Table 7).

Table 7 reveals that across different effect sizes and experimental run sizes, the expected power shows a noticeable improvement on the power level compared with the minimum power for all treatment effects. For example, if the desired power is 75% for the main effect of factor A with $\Delta_A = 2$, the expected power indicates that only two replicates (24 runs) are needed, while the minimum power asks for three replicates (36 runs). Such difference is even larger when smaller effect size is considered. For the interaction term under $\Delta_{AB} = 1$, a desired power of merely 40% for the interaction

requires four replicates based on expected power, while minimum power suggests more than six replicates are needed. If researchers use the power distribution to select run size instead of minimum power specification, for this simple two-way ANOVA model, they may be able to safely save at least one replicate of the full factorial (12 experimental runs).

5 Discussion

It is evident from the results shown in Sect. 5 that how one chooses to evaluate power has a significant impact on one’s choice of run sizes. In cases as simple as a full factorial design with one or two factors, to achieve a certain level of power, the most economic number of replicates for each treatment combination depends largely on assumptions of treatment effects.

The following questions are recommended for practitioners to go through in planning an experiment:

- What level of power is desired?
- What is the experimental budget?
- Is there any information regarding the potential arrangement of the treatment effect that can be used to infer a distribution for the τ^a ?
- What are the potential savings in run size after both the conventional and proposed methodologies for τ have been explored?

If the experimental budget is not of major concern, the minimum power specification can be adopted to protect experimenters against all possible arrangements of τ . However, we believe it is important for researchers to understand the “uncertainty” of power and evaluate any potential savings and related risks. On the other hand, as there are usually constraints on resources in designing an experiment, and viewing statistical power as a random variable indicates that a more economical run size may be suitable.

Finally, when choosing the distribution for τ , the uniform distribution can be viewed as the least informative choice. To address the risk related to poor choice of $f_\tau(\cdot)$, Table 8 presents an example where a one-way ANOVA model with factor of 4

Table 8 Comparison between suggested run size and true required run sizes for the one-way ANOVA model with factor of 4 levels, $\Delta = 1$ and $\sigma^2 = 1$

Desired power	Uniform	True distribution		
		Truncated norm	Mixture of normals	Minimum power
50%	44	44	40	52
70%	64	68	60	76
80%	76	80	72	92
90%	100	104	96	120

The expected power is used to meet the desired power

levels considered. The uniform distribution is used to assess the power distribution, with expected power, to make recommendations for the run sizes for different desired power levels. We then assess the risk by assuming the true distribution of τ to be either the truncated normal distribution or the mixture of normals as shown in Fig. 2. For all desired power levels and both true distributions, the “improper” uniform distribution suggests run sizes that deviate from the true required run sizes by at most one set of replicates, with savings of more than two sets of replicates when compared with minimum power specification. We recommend practitioners who are unsure about the distribution of τ to conduct similar analysis to explore the risk-return trade off, with the nature of the experiment in mind (whether additional experimental runs can be arranged easily, for example).

Appendix

A Proof of Theorem 4

Proof By definition, the expected value of $\eta(D, \tau^a)$ satisfies

$$\begin{aligned} 1 - E[\eta(D, \tau^a)] &= \int P(F_{k-1, n_e, \phi} \leq c | \tau^a) \cdot f_{\tau}(\tau^a) d\tau^a \\ &= \int \sum_{s=0}^{\infty} \frac{e^{-\phi/2} (\phi/2)^s}{s!} \cdot I\left(\frac{(k-1)c}{n_e + (k-1)c} \middle| \frac{k-1}{2} + s, \frac{n_e}{2}\right) f_{\tau}(\tau^a) d\tau^a. \end{aligned}$$

For any design D and value of τ^a , $\eta(D, \tau^a)$ as a power function is bounded between 0 and 1. Therefore, $1 - E[\eta(D, \tau^a)]$ is also bounded between 0 and 1. As both the cdf of F distribution and $f_{\tau}(\cdot)$ are nonnegative, by Fubini’s theorem, the integration and summation in the above equation are interchangeable. We have

$$\begin{aligned} 1 - E[\eta(D, \tau^a)] &= \sum_{s=0}^{\infty} I\left(\frac{(k-1)c}{n_e + (k-1)c} \middle| \frac{k-1}{2} + s, \frac{n_e}{2}\right) \cdot \int \frac{e^{-\phi/2} (\phi/2)^s}{s!} \cdot f_{\tau}(\tau^a) d\tau^a \\ &= \sum_s I\left(\frac{(k-1)c}{n_e + (k-1)c} \middle| \frac{k-1}{2} + s, \frac{n_e}{2}\right) \cdot g(s). \end{aligned}$$

□

B Multivariate Truncated Normal Distribution

One example for $f_\tau(\cdot)$ that is employed in this work for illustration purposes is the truncated normal distribution with truncation range $(-\Delta/2, \Delta/2)$, which implies that $\tau_1 = \Delta/2$ and $\tau_2 = -\Delta/2$. This is done to maintain effect size of Δ . If we further assume that $\tau_i, i = 3, \dots, k$, are mutually independent and have expected value of 0, we can write the joint probability density function for τ as

$$f_\tau(\boldsymbol{\tau}) = \prod_{i=3}^k \frac{\frac{1}{\sigma_\tau} f_z\left(\frac{\tau_i}{\sigma_\tau}\right) \cdot I_{[-\Delta/2 \leq \tau_i \leq \Delta/2]}}{\Phi\left(\frac{\Delta}{2\sigma_\tau}\right) - \Phi\left(-\frac{\Delta}{2\sigma_\tau}\right)} \tag{9}$$

$$= \left[\sigma_\tau \left(2\Phi\left(\frac{\Delta}{2\sigma_\tau}\right) - 1 \right) \right]^{2-k} \prod_{i=3}^k f_z\left(\frac{\tau_i}{\sigma_\tau}\right) \cdot I_{[-\Delta/2 \leq \tau_i \leq \Delta/2]},$$

where σ_τ is the scale parameter of the distribution, and $f_z(\cdot)$ is the pdf of standard normal distribution.

C Considerations for Numerical Approximations

As discussed earlier, truncation of an infinite sum and numerical method are employed to obtain the expected power in (6). That is

$$E[\eta(D, \boldsymbol{\tau}^a)] = 1 - \sum_{s=0}^{\infty} I^*(s) \cdot g(s) \approx 1 - \sum_{s=0}^M I^*(s) \cdot g(s).$$

Here, we provide some theoretical justification and a numerical illustration on the accuracy of the above approximation. First, we derive an upper bound of the approximation error. Note that the incomplete beta function is bounded above by the corresponding beta function, that is to say,

$$I\left(\frac{(k-1)c}{n_r + (k-1)c} \mid \frac{k-1}{2} + s, \frac{n_e}{2}\right) \leq B\left(\frac{k-1}{2} + s, \frac{n_e}{2}\right) \leq \frac{1}{\left(\frac{k-1}{2} + s\right) \cdot \frac{n_e}{2}}.$$

Furthermore, $g(s)$ can be viewed as the expected value of a Poisson probability mass function with parameter $\phi/2$ as

$$g(s) = \int \frac{e^{-\phi/2} (\phi/2)^s}{s!} \cdot f_\tau(\boldsymbol{\tau}^a) d\boldsymbol{\tau}^a = E_{\boldsymbol{\tau}^a} \left[\frac{e^{-\phi/2} (\phi/2)^s}{s!} \right].$$

Denote by ϕ_{\max} the maximum ϕ (which corresponds to the maximum power) given $f_\tau(\cdot), \sigma$ and the experimental setting, we can then assert that for $s > \lceil \phi/2 \rceil$,

$$g(s) \leq \frac{e^{-\phi_{\max}/2}(\phi_{\max}/2)^s}{s!},$$

by property from the Poisson distribution. Therefore, each term in the summation in equation (6) is bounded above by

$$\frac{1}{\left(\frac{k-1}{2} + s\right) \cdot \frac{n_e}{2}} \cdot \frac{e^{-\phi_{\max}/2}(\phi_{\max}/2)^s}{s!}.$$

It follows that the truncation error γ which is $1 - \sum_{s=0}^M I^*(s) \cdot g(s) - E[\eta(D, \tau^a)]$ satisfies

$$\begin{aligned} \gamma &= \sum_{s=M+1}^{\infty} I^*(s) \cdot g(s) \\ &\leq \sum_{s=M+1}^{\infty} \frac{2}{1+s} \cdot \frac{e^{-\phi_{\max}/2}(\phi_{\max}/2)^s}{s!} \\ &= 2(\phi_{\max}/2)^{-1} \sum_{s=M+2}^{\infty} \frac{e^{-\phi_{\max}/2}(\phi_{\max}/2)^s}{s!} \\ &\leq 2(\phi_{\max}/2)^{-1} \cdot \left(1 - \frac{\phi_{\max}/2}{M+3}\right) \cdot \frac{e^{-\phi_{\max}/2}(\phi_{\max}/2)^{M+2}}{(M+2)!} \end{aligned}$$

(upper bound for Poisson tail probability)

With this upper bound and some acceptable numeric error level γ , one can choose a large enough M ($M > \lceil \phi_{\max}/2 \rceil$) to construct a finite sum as approximation for the infinite sum, such that the numeric error of such approximation is within $[0, \gamma]$.

In the material example, factor A has $\phi_{\max} = 24$ with $\Delta_A = 2$. It can be easily calculated that when $M = \lceil \phi_{\max}/2 \rceil = 12$, the approximation error $\gamma = 0.002$ (relative error $\gamma/0.8074 \times 100\% = 0.25\%$), and when $M = 20$, γ drops to 0.00024 (relative error $\gamma/0.8074 \times 100\% = 0.03\%$). In Fig. 4, $I^*(s) \cdot g(s)$ and $\sum_{t=0}^M I^*(s) \cdot g(s)$ are plotted with $s \leq 30$ and $M \leq 30$, for each factor in the material example with the four illustrative distributions. It can be seen that when s is large enough, $I^*(s) \cdot g(s)$ decreases and approaches zero, which means that $\sum_{t=0}^s I^*(t) \cdot g(t)$ starts to become more steady and converge to its limit, $\sum_{t=0}^{\infty} I^*(t) \cdot g(t)$.

The computational resources required to complete numerical approximation of the expected power for the material example are presented in Table 9, together with approximation accuracy measures. Results are produced on a computer with 2.5 GHz Intel Core i7 processor and 6 GB 1600 MHz DDR3 memory. Different sizes of Monte Carlo in approximating $g(s)$ are considered for each factor. It can be seen that within reasonable computation time (no longer than two seconds), satisfactory approximation (standard deviation no larger than 0.005) of expected power can be achieved.

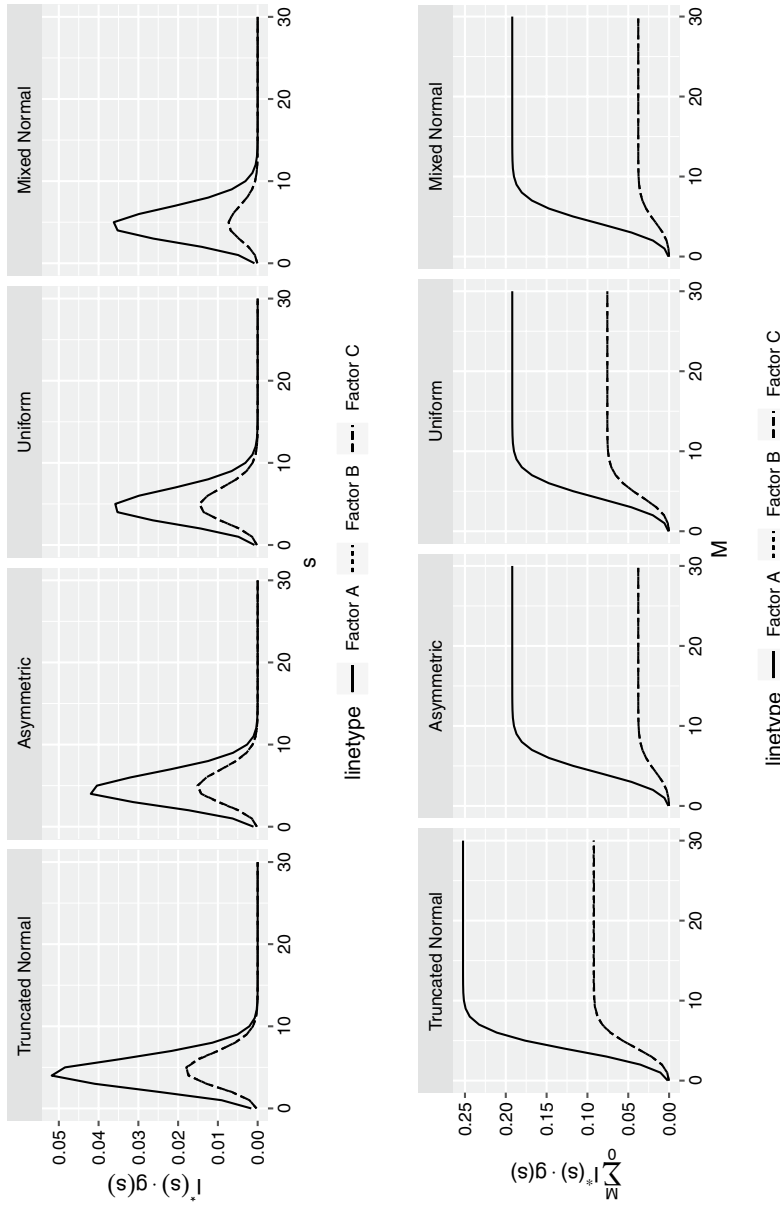


Fig. 4 Values of $f^*(s) \cdot g(s)$ (upper) and $\sum_0^M f^*(s) \cdot g(s)$ (lower) for $s \leq 30$ and $M \leq 30$. The distribution for r^d is assumed to be truncated normal with standard deviation $\sigma_r = 0.33$ and $\mu_r = 0$ (left), truncated normal with standard deviation $\sigma_r = 0.33$ and $\mu_r = 0.5$ (second to the left), uniform (third to the left) and mixture of normals (right), respectively

Table 9 Computation time and error

Monte Carlo size	Factor	Time (s)	Standard error
100	<i>A</i>	0.15	0.004
200	<i>A</i>	0.3	0.003
500	<i>A</i>	0.75	0.002
1000	<i>A</i>	1.5	0.001
100	<i>B</i>	0.15	0.001
200	<i>B</i>	0.3	0.0008
500	<i>B</i>	0.75	0.0005
1000	<i>B</i>	1.5	0.0003

References

1. Cohen J (1988) Statistical power analysis for the behavior science. Lawrence Erlbaum Associates, New York
2. Dean A, Voss D, Draguljić D et al (1999) Design and analysis of experiments, vol 1. Springer, New York
3. Fairweather PG (1991) Statistical power and design requirements for environmental monitoring. *Mar Freshw Res* 42(5):555–567
4. Gerrodette T (1987) A power analysis for detecting trends. *Ecology* 68(5):1364–1372
5. Kullback S, Rosenblatt HM (1957) On the analysis of multiple regression in k categories. *Biometrika* 44(1/2):67–83
6. Wendelberger JR, Moore LM, Hamada MS (2009) Making tradeoffs in designing scientific experiments: a case study with multi-level factors. *Qual Eng* 21(2):143–155
7. Wu CF, Hamada MS (2011) Experiments: planning, analysis, and optimization, vol 552. Wiley, New York

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.