Received: 11 April 2018 | Revised: 28 January 2019 | Accepted: 15 February 2019

DOI: 10.1002/qre.2467

RESEARCH ARTICLE

WILEY

Power approximations for failure-time regression models

Thomas H. Johnson | Rebecca M. Medlin | Laura Freeman

Institute for Defense Analyses,
Alexandria, Virginia, USA

Correspondence
Thomas H. Johnson, Institute for Defense Analyses, Alexandria, VA, USA.
Email: thjohnso@ida.org

Abstract

Reliability experiments determine which factors drive product reliability. Often, the reliability or lifetime data collected in these experiments tend to follow distinctly non-normal distributions and typically include censored observations. The experimental design should accommodate the skewed nature of the response and allow for censored observations, which occur when products do not fail within the allotted test time. To account for these design and analysis considerations, Monte-Carlo simulations are frequently used to evaluate experimental design properties. Simulation provides accurate power calculations as a function of sample size, allowing researchers to determine adequate sample sizes at each level of the treatment. However, simulation may be inefficient for comparing multiple experiments of various sizes. We present a closed-form approach for calculating power, based on the noncentral chi-squared approximation to the distribution of the likelihood ratio statistic for large samples. The solution can be used to rapidly compare multiple designs and accommodate trade-space analyses between power, effect size, model formulation, sample size, censoring rates, and design type. To demonstrate the efficiency of our approach, we provide a comparison to estimates from simulation.

KEYWORDS

censored data, design of experiments, power analysis, reliability, sample size determination

1 | INTRODUCTION

The reliability of products and systems is a concern for all fields of engineering. Reliability is the probability that a product or system will perform a required function without failure under stated conditions (ie, environmental and operating conditions) and for a stated period of time.[1] Reliability experiments typically focus on product lifetimes or time to failure, but cycles to failure are sometimes used instead of time. Reliability engineers use reliability experiments to develop models that estimate product lifetimes at use conditions. Depending on the goal of the experiment, the factors (independent variables) included in the experimental design may reflect varying use conditions, treatments designed to increase the lifetime of a product (ie, design for reliability), or stress factors that increase the probability of failure (ie, accelerated life tests [ALT]). Characteristics that differentiate reliability experiments are skewed response variables and censored observations.

There are numerous examples of reliability experiments in the literature. Much of the reliability experiment literature focuses on employing classical design techniques. An early example comes from Zelen[2] in which he considers replicated factorial experiments to determine the effect of voltage and temperature on the lifespan of a glass capacitor. The design for reliability literature focuses on classical experimental design.[3] Condra[4] uses classical experimental design to design experiments focused on improving product reliability. McCool[5] provides an example of using a 2 × 2 factorial experiment to determine if treating glassy polymers used in dental restorations improves resistance to fracture. Bullington[6] uses a resolution IV fold-over Plackett-Burman design to determine which of 11 factors influence the lifetime of an industrial thermostat, but the analysis assumed normally distributed lifetimes. These classical approaches fail to directly account for the skewed response and censoring. Accounting for these aspects is required to assess design properties and determine if the selected sample size is adequate to address the goals of the experiment.

Early papers on sizing reliability experiments that account for the skewed nature of the response focus on estimating the hazard function at a given time or a particular quantile of the data, ignoring the experimental factors. Meeker and Nelson[7] focus on estimating the quantiles for the Weibull distribution under Type I censoring. Meeker[8] emphasizes the importance of ensuring that the sample size in the experiment will deliver the desired precision, focusing on the ability to estimate the hazard function at a given time. These papers provide useful guides for selecting sample size for life tests under a fixed set of conditions, but they do not easily extend to experiments with two or more experimental factors.

Papers on reliability experiments including ALT have started to focus on determining the optimal setting for factors and allocation of test resources across factor levels. However, they fail to provide concrete recommendations on the required sample size. Meeter and Meeker[9] develop optimal ALT under the assumption of nonconstant scale parameters. Escobar and Meeker[10] developed ALT plans for tests with two or more experimental factors, including practical considerations such as avoiding accelerating variables that interact. Zhang and Meeker[11] developed Bayesian optimum and compromise test plans for ALT. However, these methods still focus on the objective of estimating with precision a particular quantile of the lifetime distribution, rather than sizing the experiment around the ability to determine what factors drive product reliability, which is often quantified using statistical power for model parameters.

Freeman and Vining[12] and Kensler et al[13] take a different approach and examine classical factorial designs with subsampling and blocking. Medlin et al[14] considers split-plot designs. Their focus was on the implications of nonrandomized designs on the analysis and resulting power. However, they did not extend their analysis to experimental design and sample size recommendations.

While power is common in classic experiment design evaluation, discussion of power is nearly nonexistent in current reliability research. In numerous papers, Meeker et al[7-11,15] focus on precision around a quantile estimate or hazard function and alluded to using Monte Carlo analysis for a more detailed review of design properties. This approach makes sense for quality control applications because the focus is on ensuring that products under any use conditions meet a specified lower bound. However, for most reliability experiments, power to detect the effect of a factor on the product lifetime is the exact reason the test is being conducted. For example, in both the Bullington[6] and McCool[5] experiments, the goal of the experiments was to determine if the treatment increased reliability. In Design for Reliability applications, power is a useful metric for determining the adequacy of the test to support sound design decisions.

Monte Carlo simulation is a flexible and accurate approach for estimating power, among other design properties, but it can be computationally inefficient, especially when comparing multiple experiments of various sample sizes and accounting for often unknown inputs including effect size, and the correct model formulation. A closed-form approximation of power allows a researcher to quickly compare several different sized experiments and evaluate robustness to unknown planning parameters such as the effect the factor will have on the failure time and related censoring rate.

In Self,[16] the authors show how to calculate an approximate power for likelihood ratio tests on coefficients within a generalized linear model. Their technique accommodates any model within the exponential family of distributions that can be arranged into the generalized linear model canonical form. Popular examples of these models include logistic, Poisson, and gamma regression models. Failure-time regression models share many qualities of a generalized linear model, but they cannot be arranged into the canonical form. For this reason, the approach by Self does not readily apply.

In this paper, we build on the technique of Self to enable power calculations for failure-time regression models involving both censored and exact-failure observations. This type of data and analysis is common when working with reliability experiments. We focus on a very common censoring scheme observed in reliability data: Type I, fixed, right-censoring. The discussion focuses on experimental designs for reliability involving categorical factors. We derive the power approximation equations for the most commonly used log-location-scale failure-time models: the lognormal and Weibull models. To demonstrate the accuracy and efficiency of our approach, we compare our calculations with the estimates generated using Monte Carlo simulation.

2 | FAILURE-TIME REGRESSION MODELS

Consideration is given to failure-time regression models within the log-location-scale family that includes the lognormal and Weibull models. Assume an experimental design has $i = 1, 2, 3 \dots, k$ unique design points and $\boldsymbol{m}_i^T$ is the ith unique row of the model matrix $\boldsymbol{M}$. The regression model has the form

$$\log(T_i) = \boldsymbol{m}_i^T \boldsymbol{\beta} + \sigma \epsilon_i. \tag{1}$$

Under this formulation, the location parameter $\mu_i = \boldsymbol{m}_i^T \boldsymbol{\beta}$ depends on the experimental factors, while the scale parameter σ does not. The lognormal model assumes $T_i \sim \text{LOGN}(\mu_i, \sigma)$ and $\epsilon_i \sim \text{NORM}(0, 1)$. The Weibull model assumes $T_i \sim \text{WEIB}(1/\sigma, e^{\mu_i})$ and $\epsilon_i \sim \text{SEV}(0, 1)$. For notational simplicity, we denote the pdf $f(t, \mu, \sigma)$ and cdf $F(t, \mu, \sigma)$ for the lognormal or Weibull distribution in location-scale form as $f_{t,\mu}$ and $F_{t,\mu}$ for $t > 0$. When referring to a specific distribution, we indicate this with a superscript, such as $f_{t,\mu}^{logn}$ or $F_{t,\mu}^{weib}$. We omit σ from the notation because in the power calculation we assume σ is a known constant.

Fixed, Type I, right-censoring prohibits the observation of all potential failure times. Let t_c be the fixed censoring time, and let T_{ij} be the jth failure time under the ith condition, where $j = 1, 2, 3, \dots, n_i$. The number of samples recorded under the ith condition is n_i, and the total sample size is $\sum_{i=1}^k n_i = N$. The censoring indicator $\delta_{ij} = 1$ if $T_{ij} < t_c$ and $\delta_{ij} = 0$ otherwise. Thus, T_{ij} is only observed if $\delta_{ij} = 1$ (or equivilently if $T_{ij} < t_c$).

The model coefficients $\boldsymbol{\beta}$ and scale parameter σ are estimated using maximum likelihood estimation. The log-likelihood for a sample of N independent observations with both right-censored and exact-failure observations is given by

$$l_{n\sigma}(\boldsymbol{\beta}) = \sum_{i=1}^k \sum_{j=1}^{n_i} \log\left[f_{T_{ij},\mu_i}\right]\delta_{ij} + \log\left[1 - F_{t_c,\mu_i}\right]\left(1 - \delta_{ij}\right), \tag{2}$$

and is dependent on $\boldsymbol{\beta}$ through $\mu_i = \boldsymbol{m}_i^T \boldsymbol{\beta}$.

Often, likelihood procedures are used to test the significance of a single model coefficient or group of model coefficients. One can split $\boldsymbol{\beta}$ into two vectors, where one vector, $\boldsymbol{\psi}$, contains the coefficients under test, and the other, $\boldsymbol{\lambda}$, includes all other coefficients that are not under test, referred to as nuisance coefficients. In a similar fashion, the model matrix $\boldsymbol{M}$ can be split into a test matrix $\boldsymbol{Z}$ and nuisance matrix $\boldsymbol{X}$, so that the location parameter under the ith condition is now expressed as

$$\mu_i = \boldsymbol{x}_i^T \boldsymbol{\lambda} + \boldsymbol{z}_i^T \boldsymbol{\psi}, \tag{3}$$

where $\boldsymbol{x}_i^T$ and $\boldsymbol{z}_i^T$ are the ith row of $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively. The hypothesis test under consideration is the null $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$, while the alternative is $H_A : \boldsymbol{\psi} \neq \boldsymbol{\psi}_0$. The likelihood ratio statistic is

$$2\left[l_{n\sigma}\left(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\lambda}}\right) - l_{n\sigma}\left(\boldsymbol{\psi}_0, \widehat{\boldsymbol{\lambda}}_0\right)\right], \tag{4}$$

where $\widehat{\boldsymbol{\psi}}$ and $\widehat{\boldsymbol{\lambda}}$ are the maximum likelihood estimates of the true parameters $(\boldsymbol{\psi}, \boldsymbol{\lambda})$ for the full model, $\boldsymbol{\psi}_0$ are the hypothesized values, and $\widehat{\boldsymbol{\lambda}}_0$ are the maximum likelihood estimates of the nuisance coefficients under the reduced model.

3 | A REVIEW OF THE SELF, MAURITSEN, AND OHARA APPROACH

The Self[16] power approximation assumes that the likelihood ratio statistic for large samples is asymptotically equivalent to a random noncentral chi-squared variable with p degrees of freedom and noncentrality parameter γ. One can solve for the noncentrality parameter and estimate power, by equating the expected value of the noncentral chi-square random variable to an approximation of the expected value of the likelihood ratio statistic. The expected value of the noncentral chi-square random variable is $p + \gamma$. The expected value of the likelihood ratio statistic can be approximated by taking the expected value of the lead terms in its asymptotic expansion.

A substantial body of research has been devoted to the asymptotic expansion of the likelihood ratio statistic. The motivation behind the research at the time was to improve the likelihood ratio test statistic so that it more closely follows a central chi-squared distribution under the null hypothesis, leading to accomplishments such as Bartlett's correction.[17] With similar asymptotic expansions in mind, except this time about the alternative hypothesis, Self[16] decomposed the likelihood ratio statistic into three terms and applied the expectation

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\psi}\boldsymbol{\lambda}}\left\{2\left[l_{n\sigma}\left(\widehat{\boldsymbol{\psi}},\widehat{\boldsymbol{\lambda}}\right)-l_{n\sigma}\left(\boldsymbol{\psi}_0,\widehat{\boldsymbol{\lambda}}_0\right)\right]\right\} &= \mathrm{E}_{\boldsymbol{\psi}\boldsymbol{\lambda}}\left\{2\left[l_{n\sigma}\left(\widehat{\boldsymbol{\psi}},\widehat{\boldsymbol{\lambda}}\right)-l_{n\sigma}(\boldsymbol{\psi},\boldsymbol{\lambda})\right]\right\} \\ &\quad - \mathrm{E}_{\boldsymbol{\psi}\boldsymbol{\lambda}}\left\{2\left[l_{n\sigma}\left(\boldsymbol{\psi}_0,\widehat{\boldsymbol{\lambda}}_0\right)-l_{n\sigma}(\boldsymbol{\psi}_0,\boldsymbol{\lambda}_0^*)\right]\right\} \\ &\quad + \mathrm{E}_{\boldsymbol{\psi}\boldsymbol{\lambda}}\left\{2\left[l_{n\sigma}(\boldsymbol{\psi},\boldsymbol{\lambda})-l_{n\sigma}(\boldsymbol{\psi}_0,\boldsymbol{\lambda}_0^*)\right]\right\}. \end{aligned} \tag{5}$$

Let the first, second, and third term on the right-hand side of Equation 5 be denoted as $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$, respectively, such that Equation 5 becomes $p+\gamma=\mathcal{A}-\mathcal{B}+\mathcal{C}$. Note, the expectation operator $E_{\boldsymbol{\psi}\boldsymbol{\lambda}}\{\cdot\}$ is taken with respect to the true parameters ($\boldsymbol{\psi}$ and $\boldsymbol{\lambda}$).

The first term ($\mathcal{A}$) in the decomposition permits an asymptotic expansion exactly as shown in Lawley.[18] Lawley's expansion included many higher order terms, but Self only retained the first two, resulting in $\mathcal{A}=p+q$, where p is the number of coefficients under test (length of $\boldsymbol{\psi}$), and q is the number of nuisance coefficients (length of $\boldsymbol{\lambda}$).

The second term ($\mathcal{B}$) does not yield a simple form like the first. Here, $\boldsymbol{\lambda}_0^*$ is the limiting value of the null coefficients. Self explains that $\widehat{\boldsymbol{\lambda}}_0$ is generally not a consistent estimator of $\boldsymbol{\lambda}$. It will converge to some value $\boldsymbol{\lambda}_0^*$ which is defined as the solution to the equation

$$\mathbf{0}=\lim_{n\to\infty}n^{-1}E_{\boldsymbol{\psi}\boldsymbol{\lambda}}\left\{\left.\frac{\partial l_{n\sigma}(\boldsymbol{\psi},\boldsymbol{\lambda})}{\partial\lambda_r}\right|_{(\boldsymbol{\psi}_0,\boldsymbol{\lambda})}\right\},\ r=1,\ 2,\ 3,\ \ldots q. \tag{6}$$

In $\mathcal{B}$, it is possible to calculate $l_{n\sigma}(\boldsymbol{\psi}_0,\boldsymbol{\lambda}_0^*)$, but not $l_{n\sigma}\left(\boldsymbol{\psi}_0,\widehat{\boldsymbol{\lambda}}\right)$. For this reason, Lawley[18] showed that a Taylor Series expansion can be used for approximation

$$\mathcal{B}=\mathrm{E}_{\boldsymbol{\psi}\boldsymbol{\lambda}}\left\{\boldsymbol{U}_{\lambda}^T\mathcal{I}_{\lambda\lambda}^{-1}\boldsymbol{U}_{\lambda}\right\}, \tag{7}$$

where $\boldsymbol{U}_{\lambda}$ is size $q\times1$, the rth element of $\boldsymbol{U}_{\lambda}$ is

$$\left.\frac{\partial l_{n\sigma}(\boldsymbol{\psi},\boldsymbol{\lambda})}{\partial\lambda_r}\right|_{\left(\boldsymbol{\psi}_0,\boldsymbol{\lambda}_0^*\right)}, \tag{8}$$

and $\mathcal{I}_{\lambda\lambda}^{-1}$ is a $q\times q$ matrix with rth-sth element equal to

$$\mathrm{E}_{\boldsymbol{\psi}\boldsymbol{\lambda}}\left\{\left.\frac{\partial^2 l_{n\sigma}(\boldsymbol{\psi},\boldsymbol{\lambda})}{\partial\lambda_r\partial\lambda_s}\right|_{\left(\boldsymbol{\psi}_0,\boldsymbol{\lambda}_0^*\right)}\right\}^{-1}, \tag{9}$$

where $s=1,\ 2,\ 3,\ \ldots,\ q$. Although Equation 7 permits a closed-form solution for generalized linear models within the canonical form, which was the focus of the work by Self, it does not allow a closed-form solution for failure-time regression models with censoring.

The third term ($\mathcal{C}$) can be calculated exactly for both generalized linear models in the canonical form as shown in Self and failure-time regression models with Type I, fixed, right-censored data as we will show in this paper. To approximate the power of the likelihood ratio test, one must first solve for γ and then calculate the probability of exceeding the critical value by referring to the noncentral chi-square distribution. The approach of Self, when applied to generalized linear models in the canonical form, provides a closed-form solution for the noncentrality parameter as

$$\gamma=\mathcal{A}-\mathcal{B}+\mathcal{C}-p. \tag{10}$$

The power is

$$1-\chi^2_{\mathrm{cdf}}\left\{\chi^2_{1-\alpha},p,\gamma\right\}, \tag{11}$$

where $\chi^2_{1-\alpha}$ is the $(1-\alpha)\times100$th quantile of the central χ^2 distribution, and χ^2_{cdf} is the noncentral chi-square distribution function with p degrees of freedom and noncentrality parameter γ.

4 | PROPOSED APPLICATION TO FAILURE-TIME REGRESSION MODELS

The objective of this paper is to provide a computationally efficient power approximation for the purposes of evaluating reliability experiments that avoids using costly numerical techniques. The closed-form solution in Equation 10 is not attainable for failure-time models due to the second term ($\mathcal{B}$). Numerically solving $\mathcal{B}$ is possible but would undermine our objective. We propose simplifying the calculation of the noncentrality parameter by omitting $\mathcal{A}$, $\mathcal{B}$, and p from Equation 10, which will enable a computationally efficient, closed-form solution.

Previous work suggests this simplification may be permissible. Self[16, 38, p.] discuss "the dominance of the [$\mathcal{C}$] term in the calculation of the noncentrality parameter," and states, "In our experience, the term [$\mathcal{A}-\mathcal{B}-p$] is usually very close to zero." The same discovery was made in work by Self,[19] Shieh,[20] and Brown.[21] Applying the simplification, the noncentrality parameter becomes $\gamma = \mathcal{C}$, which can be calculated for failure-time regression models in closed form. We investigate the consequences of this simplification later using a simulation study. Our results show that $\mathcal{A}$ and $\mathcal{B}$ are negligible for the conditions that are considered in the study.

Proceeding with this assumption, we develop the general formula for $\mathcal{C}$ for failure-time regression models. From Equation 5, $\mathcal{C}$ is given by

$$\begin{aligned}\mathcal{C} &= 2\mathrm{E}_{\boldsymbol{\psi}\boldsymbol{\lambda}}\{l_{n\sigma}(\boldsymbol{\psi},\boldsymbol{\lambda}) - l_{n\sigma}(\boldsymbol{\psi}_0,\boldsymbol{\lambda}_0^*)\} \\ &= 2\mathrm{E}_{\boldsymbol{\psi}\boldsymbol{\lambda}}\Big\{\log\big[f_{T_{ij},\mu_i}\big]\delta_{ij} + \log\big[1-F_{t_c,\mu_i}\big](1-\delta_{ij}) - \log\big[f_{T_{ij},\mu_i^*}\big]\delta_{ij} + \log\big[1-F_{t_c,\mu_i^*}\big](1-\delta_{ij})\Big\}.\end{aligned} \tag{12}$$

Recall that $T_{i1}, T_{i2}, \dots, T_{in_i}$ is lognormal or Weibull distributed with location parameter μ_i and scale parameter σ. δ_{ij} is a function that equals 1 if $T_{ij} \le t_c$ and equals 0 if $T_{ij} > t_c$. Additionally, $\mu_i^* = \boldsymbol{x}_i^T\boldsymbol{\lambda}_0^* + \boldsymbol{z}_i^T\boldsymbol{\psi}_0$.

Evaluation of the expectation operator in Equation 12 is analytically tractable, but the details of this calculation are omitted. The resulting expressions for the lognormal and Weibull regression models are, respectively, shown in Equations 13 and 14.

$$\mathcal{C} = \sum_{i=1}^{k} n_i\left\{\frac{2f_{t_c,-\mu_i}^{logn}(\mu_i^*-\mu_i)}{t_c^{-1-\frac{2\mu_i}{\sigma^2}}} + \frac{F_{t_c,\mu_i}^{logn}(\mu_i-\mu_i^*)^2}{\sigma^2} + \left(2-2F_{t_c,\mu_i}^{logn}\right)\log\frac{2-2F_{t_c,\mu_i}^{logn}}{2-2F_{t_c,\mu_i^*}^{logn}}\right\} \tag{13}$$

$$\mathcal{C} = \sum_{i=1}^{k} n_i\left\{-\frac{2}{\sigma}e^{-\frac{\mu_i^*}{\sigma}}F_{t_c,\mu_i}^{weib}\left(-e^{\frac{\mu_i}{\sigma}}\sigma + e^{\frac{\mu_i^*}{\sigma}}(\mu_i-\mu_i^*+\sigma)\right)\right\} \tag{14}$$

5 | EXAMPLE: LOGNORMAL REGRESSION MODEL

For proof of concept, consider a reliability experiment designed for a batch of $N = 240$ products. A $2\times2\times3$ full factorial experiment exposes the products to 12 unique treatment combinations, resulting in 20 products per condition. Each product is exposed until failure or up to 10 days; thus, the right censoring time is $t_c = 10$. The three categorical factors are $\mathcal{F}_1$ (two-level factor), $\mathcal{F}_2$ (two-level factor), and $\mathcal{F}_3$ (three-level factor).

The effect size of interest is defined in terms of a change in probability of failure at a particular time of interest, t_p. The engineer identifies day 7 as a crucial juncture in the product's lifetime ($t_p = 7$) and anticipates 80% of the products will fail by this time. At $t_p = 7$, the engineer would like to detect a 10 percentage point change in probability of failure due to an exposure factor. Thus, the nominal failure rate is $\overline{p} = .8$, and the effect size is $\Delta_p = .1$. This results in a lower and upper probability of failure equal to $p_1 = .75$ and $p_2 = .85$, respectively. Based on inspection of historical test data, the engineer assumes the failure times follow a lognormal distribution with scale parameter $\sigma = 2$. The effect sizes, in terms of the location parameter, are obtained by solving for μ_{p_1} in $p_1 = F_{t_p,\mu_{p_1}}^{logn}$, and by solving for μ_{p_2} in $p_2 = F_{t_p,\mu_{p_2}}^{logn}$, resulting in $\mu_{p_1} = .60$ and $\mu_{p_2} = -.13$.

Assuming a main effects model, and using a sum-to-zero contrast scheme, the effect size in terms of alternative coefficient vector is

$$\begin{aligned}\boldsymbol{\beta}^T &= \left[\beta_{int}\ \ \beta_{\mathcal{F}_1}\ \ \beta_{\mathcal{F}_2}\ \ \boldsymbol{\beta}_{\mathcal{F}_3}\right] \\ &= \left[\frac{\mu_{p_1}+\mu_{p_2}}{2}\ \frac{\mu_{p_1}-\mu_{p_2}}{2}\ \frac{\mu_{p_1}-\mu_{p_2}}{2}\ \frac{\mu_{p_1}-\mu_{p_2}}{2}\ 0\right] \\ &= \left[\,.23 \quad .36 \quad .36 \quad .36 \quad .00\right].\end{aligned} \tag{15}$$

The coefficient vector $\boldsymbol{\beta}$ is size 5×1 with first element corresponding to the intercept, second element corresponding to $(\mathcal{F}_1)$, third element corresponding to $(\mathcal{F}_2)$, and fourth and fifth elements corresponding to $(\mathcal{F}_3)$.

We illustrate the power calculation for $\mathcal{F}_1$, where the hypothesis test of interest is $H_0 : \psi = 0$ versus $H_A : \psi \neq 0$. Splitting the alternative coefficient vector into two parts, we have

$$\boldsymbol{\lambda}^T = \left[\beta_{int} \quad \beta_{\mathcal{F}_2} \quad \boldsymbol{\beta}_{\mathcal{F}_3}\right],\ \psi^T = \beta_{\mathcal{F}_1}. \tag{16}$$

The location parameter at each unique exposure condition is $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\lambda} + \boldsymbol{z}_i^T\psi$ for $i = 1, 2, \ldots, 12$.

The limiting value of the null coefficients $(\boldsymbol{\lambda}^*{}_0)$ is obtained by fitting a lognormal regression model to the alternative data. The alternative data are the failure times that represent the perfect fit to the alternative coefficients. We denote these failure times as T_{ij}^* and solve for them by setting ϵ_i equal to zero in Equation 1. This results in $T_{ij}^* = e^{\mu_i}$ for all $i = 1, 2, \ldots, 12$ and $j = 1, 2, \ldots 20$. Using standard failure-time model fitting software, such as JMP's parametric survival model fitting platform or the R function "survreg," fit the reduced model to T_{ij}^*. The fitted coefficients are equal to $\boldsymbol{\lambda}_0^*$. Then, calculate $\mu_i^* = \boldsymbol{x}_i^T\boldsymbol{\lambda}_0^*$ for $i = 1, 2, \ldots, 12$.

Using Equation 13, the noncentrality parameter is $\gamma = \mathcal{C} = 7.59$. The critical value, which is found by referring to the central chi-squared distribution with $p = 1$ degrees of freedom and $\alpha = 0.05$, is equal to 3.84. Finally, referencing the noncentral chi-squared distribution, power is equal to 0.79. The power calculations for $\mathcal{F}_2$ and $\mathcal{F}_3$ follow this same procedure.

6 | VERIFICATION STUDY

This study serves as a limited verification of our approach. It compares power estimates of the proposed method with that of Monte Carlo simulation. Comparisons are made over a variety of benign conditions, including changes in model type, experimental design, sample size (N), nominal failure rate ($\overline{p}$), effect size (Δ_p), and the scale parameter (σ). The verification study conditions are shown in Table 1 and are varied according to a 40-run D-optimal experimental design. Power is calculated for each of the 40 simulation conditions.

Calculation of the effect size follows the same procedure described in the earlier example. The effect size for each run in the simulation D-optimal experiment is calculated in terms of the location parameter by solving for μ_{p_1} in $p_1 = F_{t_p,\mu_{p_1}}$, and by solving for μ_{p_2} in $p_2 = F_{t_p,\mu_{p_2}}$, where the model, $\overline{p}$, Δ_p, and σ are specified by each row in the simulation D-optimal experiment. The coefficients $\boldsymbol{\psi}$ and $\boldsymbol{\lambda}$ are calculated from μ_{p_1} and μ_{p_2}. The particular time of interest (t_p) and fixed right censoring time (t_c) are set constant and equal to 7 and 10 days, respectively.

The simulation study calculates power for two factorial experiments. The first is a 2^2 experiment, which has two two-level factors, and the second is a 2^6, which has six two-level factors. The experiments are replicated to meet the specified sample size (N). The hypothesis test $\boldsymbol{\psi}_0 = \mathbf{0}$ is applied to the first factor in each experiment, and the power of this test is presented.

The outputs of the simulation study are the approximated power ($\mathcal{P}$) and Monte Carlo power ($\mathcal{P}_{mc}$) using 10 000 iterations. $\mathcal{P}_{mc}$ is calculated as follows. For each iteration of the simulation, the random failure times are generated, a

TABLE 1 Simulation conditions

Model	Design	N	$\overline{p}$	Δ_p	σ
Lognormal	2^2	64	0.50	0.06	0.50
Weibull	2^6	128	0.80	0.12	1.00
		256			3.00

failure-time regression model is fit to these data, a likelihood ratio test is conducted, and the resulting P-values are recorded. Power is the proportion of iterations that have a P-value less than the significance level ($\alpha = .05$).

Tables 2 and 3 present the results of the simulation study. The difference in power ($\mathcal{P} - \mathcal{P}_{mc}$) between the proposed method and traditional Monte Carlo method highlights the similarity between the two approaches. The mean and standard deviation of ($\mathcal{P} - \mathcal{P}_{mc}$) across all simulation conditions are $-.005$ and 005, respectively.

As a screening analysis, we fit a main effects linear model to the absolute value of ($\mathcal{P} - \mathcal{P}_{mc}$), which we denote as $|\mathcal{P} - \mathcal{P}_{mc}|$, to identify the influential simulation study conditions. The results of this analysis indicate that the influential conditions, in descending order from most to least influential, are design, $\bar{p}$, model, n, σ, and Δ_p. The 2^6 experiment resulted in larger $|\mathcal{P} - \mathcal{P}_{mc}|$ than the 2^2 which may suggest that additional nuisance parameters decrease the accuracy of the approximation method. Larger values of $\bar{p}$ led to smaller $|\mathcal{P} - \mathcal{P}_{mc}|$, and the lognormal model tended to have smaller $|\mathcal{P} - \mathcal{P}_{mc}|$.

7 | EFFICIENCY STUDY

The benefit of our approximation method relies on its computational efficiency. The previous section served as a limited verification of our approach for a broad set of conditions. This section narrows its focus to the condition specified in the example problem (Section 5) to closely assess accuracy relative to computation timeliness.

Computational efficiency requires consideration of the trade-off between accuracy and timeliness. The accuracy and timeliness of our approximation approach for a fixed set of input parameters are constant because the computation is deterministic. The method's accuracy is limited by the Taylor series approximation to the distribution of the likelihood ratio statistic. In contrast, Monte Carlo methods improve in accuracy and increase in computation time as the number of simulation iterations increases. The following study attempts to compare the computation efficiency between the two methods.

TABLE 2 Lognormal simulation results

	Model	Design	N	$\bar{p}$	Δ_p	σ	$\mathcal{P}$	$\mathcal{P} - \mathcal{P}_{mc}$
1	Lognormal	2^2	64	0.5	0.06	0.5	0.090	−0.002
2	Lognormal	2^2	64	0.5	0.06	3	0.086	−0.001
3	Lognormal	2^2	64	0.5	0.12	1	0.207	−0.007
4	Lognormal	2^2	64	0.8	0.06	1	0.136	−0.002
5	Lognormal	2^2	64	0.8	0.12	3	0.403	0.001
6	Lognormal	2^2	128	0.5	0.12	3	0.348	−0.002
7	Lognormal	2^2	128	0.8	0.06	3	0.223	−0.002
8	Lognormal	2^2	128	0.8	0.12	0.5	0.691	−0.003
9	Lognormal	2^2	256	0.5	0.06	1	0.207	−0.012
10	Lognormal	2^2	256	0.5	0.12	0.5	0.649	−0.002
11	Lognormal	2^2	256	0.8	0.06	3	0.394	−0.003
12	Lognormal	2^6	64	0.5	0.06	1	0.088	−0.007
13	Lognormal	2^6	64	0.5	0.12	0.5	0.215	0.009
14	Lognormal	2^6	64	0.8	0.06	0.5	0.137	−0.004
15	Lognormal	2^6	128	0.5	0.06	0.5	0.131	−0.001
16	Lognormal	2^6	128	0.5	0.12	1	0.361	−0.010
17	Lognormal	2^6	128	0.8	0.06	1	0.226	−0.001
18	Lognormal	2^6	128	0.8	0.12	3	0.674	−0.006
19	Lognormal	2^6	256	0.5	0.12	3	0.597	−0.007
20	Lognormal	2^6	256	0.8	0.06	0.5	0.403	−0.001

TABLE 3 Weibull simulation results

	Model	Design	N	$\bar{p}$	Δ_p	σ	$\mathcal{P}$	$\mathcal{P}-\mathcal{P}_{mc}$
21	Weibull	2^2	64	0.5	0.06	1	0.085	−0.006
22	Weibull	2^2	64	0.8	0.06	1	0.109	−0.002
23	Weibull	2^2	64	0.8	0.12	0.5	0.314	−0.005
24	Weibull	2^2	128	0.5	0.06	1	0.122	0.000
25	Weibull	2^2	128	0.5	0.12	0.5	0.401	−0.004
26	Weibull	2^2	128	0.8	0.12	1	0.524	−0.009
27	Weibull	2^2	128	0.8	0.12	3	0.497	−0.004
28	Weibull	2^2	256	0.5	0.06	3	0.176	0.002
29	Weibull	2^2	256	0.8	0.06	0.5	0.311	0.002
30	Weibull	2^2	256	0.8	0.12	1	0.815	−0.003
31	Weibull	2^6	64	0.5	0.06	0.5	0.092	−0.012
32	Weibull	2^6	64	0.5	0.12	3	0.178	−0.023
33	Weibull	2^6	64	0.8	0.06	3	0.105	−0.017
34	Weibull	2^6	64	0.8	0.12	1	0.293	0.001
35	Weibull	2^6	128	0.5	0.06	3	0.112	−0.010
36	Weibull	2^6	128	0.8	0.06	0.5	0.179	−0.003
37	Weibull	2^6	256	0.5	0.12	1	0.598	−0.005
38	Weibull	2^6	256	0.8	0.06	1	0.293	−0.005
39	Weibull	2^6	256	0.8	0.12	0.5	0.833	−0.001
40	Weibull	2^6	256	0.8	0.12	3	0.781	−0.005

Consider a Monte Carlo simulation setup for the earlier example. Recall the example problem used the approximation technique to calculate power for the test on the coefficient: $\beta_{\mathcal{F}_1}$. For this same construct, we now compute power using Monte Carlo simulation. To evaluate simulation accuracy and timeliness, we repeat the Monte Carlo power computation by varying the number of simulation iterations: 100, 500, 1000, 5000, 10 000, 15 000, 20 000, and 25 000. We estimate Monte Carlo power by taking the proportion of iterations that have a *P*-value less than the significance level ($\alpha = .05$). For each iteration value considered, we replicate the computation 20 times, generating a distribution of Monte Carlo power values at a given iteration setting. We facilitate the accuracy assessment by establishing a "baseline" metric that represents a reasonably good estimate of the true power. We define the "baseline" as the mean Monte Carlo power estimate from 5 million simulation iterations (equal to 0.78820). The accuracy of the approximation estimate is defined as the difference between the "baseline" and the approximation estimate from the previous example section. This difference is $0.78820 - 0.78688 = 0.00132$. The time it takes to compute all steps of the approximation estimate, given our computer specifications and best coding efforts, is 0.25 seconds.

The accuracy of the Monte Carlo estimate varies with the number of iterations. For a given number of iterations, let "error" represent the difference between the "baseline" and the Monte Carlo power estimate for a single iteration. Then, the accuracy is defined as the root-mean-squared-error (RMSE). For illustration, for the 100 iteration condition, 20 "errors" are computed, and the RMSE is calculated by taking the square root of the mean of the squared "errors." Table 4 presents the RMSE for each number of simulation iterations and presents the mean time it takes to compute one simulation iteration.

The results indicate for the 1000 iteration condition that the approximation method is roughly one order of magnitude more timely than the Monte Carlo approach (0.25 seconds versus 4.1 seconds) and one order of magnitude more accurate (.001 versus.01). At first glance, 4.1 seconds may appear trivial, but computation time can quickly compound. For example, in practice, the computation is performed separately for each coefficient in the model, which equates to nine coefficients for the corresponding main effect and two-factor interaction model.

Sample size determination problems are often set up to solve for the sample size that provides at minimum the prespecified level of confidence and power. This requires an iterative numerical technique. Assuming the technique

TABLE 4 Efficiency comparison

Method	Iterations	RMSE	(95% CI)	Mean Time (Sec)
Power Approx.	1	Abs. Error = 0.0013	–	0.25
Monte Carlo	100	0.0363	(0.0278, 0.0525)	0.46
Monte Carlo	500	0.0155	(0.0119, 0.0224)	2.85
Monte Carlo	1000	0.0105	(0.0081, 0.0152)	4.12
Monte Carlo	5000	0.0052	(0.0040, 0.0076)	19.91
Monte Carlo	10000	0.0038	(0.0029, 0.0055)	39.73
Monte Carlo	15000	0.0032	(0.0024, 0.0046)	59.52
Monte Carlo	20000	0.0030	(0.0023, 0.0044)	78.30
Monte Carlo	25000	0.0025	(0.0019, 0.0037)	98.01

requires five iterations, and given nine model coefficients, in practice, the sample size determination computation using Monte Carlo could take 3 minutes (4 seconds × 9 coefficients × 5 iterations) compared with 11 seconds for our approximation method (0.25 seconds × 9 coefficients × 5 iterations). These durations are further exacerbated in practice when one chooses to explore other inputs parameters, such as different experimental designs, effect sizes, and confidence levels. Such in-depth exploration could reasonably take hours.

8 | CONCLUSIONS

The proposed method provides power estimates for sizing reliability experiments by adapting the approximation technique that Self developed in 1992.[16] Their method was developed out of necessity due to computing resources at that time. Today, it is quite reasonable to use simulation to calculate a single estimate of power, but when numerous power estimates are needed (ie, when comparing experimental designs, effect sizes, confidence levels), the approximation technique becomes cost effective. As an illustration, our Efficiency Study demonstrated a scenario where the approximation technique was an order of magnitude more accurate and timely than Monte Carlo simulation.

The Verification Study demonstrated the accuracy of our method for a variety of benign conditions: large samples, small effect sizes, and full factorial experiments with few factors and levels. Our approximation method retains one of three terms in Self's decomposition of the likelihood ratio statistic. The term $\mathcal{C}$ is calculated exactly and is presented for both the lognormal and Weibull regression models. We omitted $\mathcal{A}$ and $\mathcal{B}$ from the formulation, because they could not be solved in closed form. Nonetheless, the Verification Study demonstrated the accuracy of our approach.

Future work may include the extension to more general families of failure-time distributions, such as the generalized gamma, generalized F, or log-logistic models. The method could also be applied to interval or left-censoring. Although this paper focused on multilevel categorical covariates in the failure-time regression model, one might also consider adapting the approach to accommodate continuous factors, polynomial, or mixed polynomial model effects. And finally, one could also explore the accuracy of our method for alternative experimental designs, such as fractional factorial experiments.

REFERENCES

1. ANSI/GEIA Standard. 0009 reliability program standard for systems design, development, and manufacturing. 2008.
2. Zelen M. Factorial experiments in life testing. *Dent Tech*. 1959;1(3):269-288.
3. Crowe D, Feinberg A. *Design for Reliability*. 11 Boca Raton, FL: CRC Press; 2001.
4. Condra L. *Reliability Improvement with Design of Experiment*. Madison Avenue, New York, NY: CRC Press; 2001.
5. McCool JI, Baran G. The analysis of 2× 2 factorial fracture experiments with brittle materials. *J Mater Sci*. 1999;34(13):3181-3188.
6. Bullington RG, Lovin S, Miller DM, Woodall WH. Improvement of an industrial thermostat using designed experiments. *J Qual Technol*. 1993;25(4):262-270.
7. Meeker WQ, Nelson W. Weibull variances and confidence limits by maximum likelihood for singly censored data. *Dent Tech*. 1977;19(4):473-476.

8. Meeker WQ, Escobar LA, Hill DA. Sample sizes for estimating the Weibull hazard function from censored samples. *IEEE Trans Reliab*. 1992;41(1):133-138.
9. Meeter CA, Meeker WQ. Optimum accelerated life tests with a non-constant scale parameter. *Dent Tech*. 1994;36(1):71-83.
10. Escobar LA, Meeker WQ. Planning accelerated life tests with two or more experimental factors. *Dent Tech*. 1995;37(4):411-427.
11. Zhang Y, Meeker WQ. Bayesian methods for planning accelerated life tests. *Dent Tech*. 2006;48(1):49-60.
12. Freeman LJ, Vining GG. Reliability data analysis for life test designed experiments with sub-sampling. *Qual Reliab Eng Int*. 2013;29(4):509-519.
13. Kensler JLK, Freeman LJ, Vining GG. Analysis of reliability experiments with random blocks and subsampling. *J Qual Technol*. 2015;47(3):235-251.
14. Medlin R, Freeman L, Kensler J, Vining G. Analysis of a split-plot reliability experiment with subsampling. Submitted to Quality and Reliability Engineering International (Under Review); 2019.
15. Meeker WQ, Escobar LA. *Statistical Models for Reliability Data*. NJ: John Wiley & Sons; 1998.
16. Self SG, Mauritsen RH, Ohara J. Power calculations for likelihood ratio tests in generalized linear models. *Biometrics*. 1992;48(1):31-39.
17. Bartlett MS. Approximate confidence intervals. II. More than one unknown parameter. *Biometrika*. 1953;40(3/4):306-317.
18. Lawley DN. A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika*. 1956;43(3/4):295-303.
19. O'Brien RG, Shieh G. *A Simpler Method to Compute Power for Likelihood Ratio Tests in Generalized Linear Models*. Dallas, Texas: Annual Joint Statistical Meetings of the American Statistical Association:1998.
20. Shieh G. On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics*. 2000;56(4):1193.
21. Brown BW, Lovato J, Russell K. Asymptotic power calculations: description, examples, computer code. *Stat Med*. 1999;18(22):3140.

AUTHOR BIOGRAPHIES

Thomas H. Johnson is a research staff member with the Institute for Defense Analyses, Alexandria, VA, where he supports the Live Fire Test and Evaluation Task providing expertise in statistics. His areas of emphasis include sample size determination, reliability experiments, and acceptance sampling plans. He focuses on operational tests for personal protect equipment, as well as tests for Army helicopters. He received a BS degree from Boston University, and MS and PhD degrees from Old Dominion University, all in Aerospace Engineering.

Rebecca M. Medlin is a research staff member with the Institute for Defense Analyses, Alexandria, VA, where she supports the Air Warfare Test and Evaluation Task providing expertise in statistics. She focuses on the operational test of radars and electronic warfare systems. Her areas of expertise include design of experiments and reliability analysis. She received her MS and PhD degrees in statistics from the Virginia Tech, Blacksburg, VA.

Laura Freeman is an assistant director with the Institute for Defense Analyses, Alexandria, VA, where she leads the Test Science Task providing support to the Director, Operational Test and Evaluation on the use of statistics in test and evaluation. Her areas of statistical expertise include designed experiments, reliability analysis, and industrial statistics. She focuses on operational tests for Air Force fixed wing aircraft. In addition, she has a background in aerospace engineering. She received a BS degree in Aerospace Engineering, and MS and PhD degrees in statistics from the Virginia Polytechnic Institute and State University, Blacksburg, VA.

How to cite this article: Johnson TH, Medlin RM, Freeman L. Power approximations for failure-time regression models. *Qual Reliab Engng Int*. 2019;35:1666–1675. https://doi.org/10.1002/qre.2467