Assurance for Sample Size Determination in Reliability Demonstration Testing

Kevin J. Wilson and Malcolm Farrow

School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne, UK

ABSTRACT
Manufacturers are required to demonstrate that products meet reliability targets. A way to achieve this is with reliability demonstration tests (RDTs), where a number of products are put on test and the test is passed or failed according to a decision rule based on the observed outcomes. There are various methods for determining the sample size for RDTs, typically based on the power of a hypothesis test following the RDT or risk criteria. Bayesian risk criteria approaches combine the choice of sample size with the analysis of the test data while relying on the specification of acceptable and rejectable reliability levels. In this article, we offer an alternative approach to sample size determination based on the idea of assurance. This approach chooses the sample size to provide a specified probability that the RDT will result in a successful outcome. It separates the design and analysis of the RDT, allowing different priors for the producer and consumer. We develop the assurance approach for sample size calculations in RDTs for binomial and Weibull likelihoods and propose appropriate prior distributions for the design and analysis of the test. In each case, we illustrate the approach with an example based on real data.

1. Introduction

Hardware products are required to have high levels of reliability, particularly those which provide safety-critical functions within larger systems. It is therefore crucial that the users of hardware products have confidence in the reliability claimed by the manufacturers. A typical and widely used approach to provide this confidence is through the use of a reliability demonstration test (RDT), in which a number of the hardware products are put on test and the number which fail or failure times are observed. If a statistic based on the observed outcomes meets some predefined threshold the test is said to be passed and the reliability of the product is demonstrated (Elsayed 2012).

An additional consideration affecting the test plan is the type of hardware product of interest. In this article, we classify hardware products into two types: products which are required to function on demand and can either work or fail, such as a parachute, and products which are required to function for a stated period of time, such as an engine. The design of an RDT for failure on demand products is given by the number of items put on test and the number which fail or failure times are observed. If a statistic based on the observed outcomes meets some predefined threshold the test is said to be passed and the reliability of the product is demonstrated (Elsayed 2012).

In a widely investigated set of approaches related to the approach in this article, risk criteria are considered. In these approaches, the risk to the producer and consumer of the product associated with incorrect conclusions from the RDT are evaluated. The first attempt was the classical risk criteria which defined the producer’s (consumer’s) risk as the probability of failing (passing) the test conditional on a chosen (un)acceptable value for the reliability (see, e.g., Martz and Waller 1982, chap. 10). To overcome the need to specify a value for the reliability of a passed (failed) test, average risk criteria (Easterling 1970), which condition on being above (below) a reliability threshold, and posterior risk criteria, which calculate the probability of (not) reaching a reliability target conditional on passing (failing) the RDT, have been proposed. See, for example, Chapter 10 of Hamada et al. (2008) and Hamada et al. (2014). As a result of
the structure of the approach, the prior distribution used in the posterior risk criteria is the same for the design and analysis of the RDT.

A related area is that of designing accelerated life tests. However, in this case, test designs are typically chosen for best inference of the underlying reliability of the product, for example, by minimizing the expectation of the posterior variance. For a recent example, see Lee et al. (2018). Another closely related area is Bayesian acceptance testing in quality assurance. We give an overview of this area and a discussion of the work in relation to this article in Section A of the supplementary materials.

Assurance has been proposed (Spiegelhalter, Freedman, and Parmar 1994; O'Hagan and Stevens 2001) as the correct Bayesian method for calculating sample sizes in clinical trials. It is based on the notion that the sample size should be chosen to meet a threshold for the probability that the trial leads to a successful outcome. If the analysis to be conducted following the trial is a frequentist hypothesis test for the superiority of a new treatment over the standard treatment, then the successful outcome would be rejection of the null hypothesis in the test. Spiegelhalter, Freedman, and Parmar (1994) considered simple structures of clinical trials and O'Hagan, Stevens, and Campbell (2005) extended the idea to trials involving binary data and nonconjugate prior distributions. Ren and Oakley (2014) detailed assurance calculations for clinical trials with time-to-event outcomes, considering exponential and Weibull survival distributions. Miller et al. (2018) compared assurance to traditional power calculations in clinical trials for rare diseases.

There is no limitation with assurance that the analysis following the trial be a frequentist test, however. O'Hagan and Stevens (2001) proposed assurance for sample size determination based on a Bayesian analysis following the test. In this case, a successful test was defined based on the posterior probability of the new treatment being superior to the standard treatment. They proposed to use different priors in the design and the analysis of the trial, representing the beliefs of different groups of people. Walley et al. (2015) used this approach in a case study set in early drug development, and discussed suitable design and analysis priors. Muirhead and Şoaia (2013) also proposed assurance for sample size determination based on a Bayesian analysis following the trial.

In this article, we consider the use of assurance for sample size determination outside the context of medicine for the first time. We consider RDTs for products which are required to function on demand and for products which are required to function for a stated length of time. In both cases, we develop an approach to select the number of items to test, based on the assurance, and the criterion for a successful RDT. We propose suitable structures for the prior distributions for the unknown model parameters and detail how historical data can be incorporated into the assurance calculation. In both the binomial and Weibull cases, the methodology developed is much more general than has been considered in the assurance literature previously. In particular, for Weibull random variables, the only work to date has considered a simple scenario of a two-sample hypothesis test. This article offers a substantial advance on this work, in the development of a suitable Bayesian analysis of the test results, in the incorporation of historical data into the design of the test and in the allocation of the sample size to different values of the design variable (in our case stress). In the binomial case, the development of suitable structures for analysis prior distributions for a Bayesian analysis based on skeptical and mixture priors is novel.

We regard the design of the RDT as a decision for the producer. The decision whether or not to accept the product is, however, the consumer's. If the consumer wishes to use a traditional consumer's-risk approach then, given the sample size chosen by the producer, the consumer can select a criterion to give the required consumer's risk. The consumer's risk is essentially the critical value for a test of the null hypothesis that the reliability is at the consumer's "acceptable level" against the alternative that the reliability is better than this. However, we also allow the consumer to use a Bayesian analysis and give particular attention to this. From the producer's point of view, in a traditional producer's-risk approach the producer considers the conditional probability that the test will be failed, that is, the consumer decides against accepting the product, given that the true reliability is at the producer's "acceptable level." However, in our assurance approach, the producer considers his or her marginal probability that the test will be failed, taking into account the producer's own probability distribution over the true reliability. If, as is likely to be the case often, the producer believes that the true reliability is probably better than the "acceptable level," then the producer's marginal probability of the test being failed will be less than the conditional producer's risk. In such a case, the producer could achieve an acceptable marginal risk with a smaller sample size. In any case, using the marginal probability gives a more realistic assessment of the risk to the producer.

Being able to incorporate the prior judgments of the producer into the design of the RDT, without also imposing them into the analysis of the test results, is therefore an important advantage of assurance over typical Bayesian risk-criteria-based approaches to RDT design. It is therefore appropriate, in the producer's choice of a design, to use the producer's prior beliefs, expressed through the design prior. Assurance also allows us to incorporate historical data into the design of the RDT, again without imposing these historical data into the analysis.

However, a successful outcome is that the consumer is sufficiently convinced of the reliability of the product by the results of the RDT. If the consumer also uses a Bayesian analysis then the consumer's posterior beliefs, after seeing the results of the RDT, will depend on the consumer's prior beliefs and these cannot be assumed to be the same as the producer's prior beliefs. We therefore allow a separate analysis prior to represent the consumer's beliefs. This feature is in common with the adversarial approaches of Lindley and Singpurwalla (1993) and Rios Insua et al. (2018). If the producer knows what the consumer's beliefs are then the producer can use this analysis prior directly when choosing a design. If the producer does not know the consumer's prior beliefs then the producer might assess a prior over the possible priors which the consumer might have or, more simply, use a conservative analysis prior for the purpose of the design.

We detail our approach to sample size determination using assurance for failure on demand data, using a binomial likelihood function, in Section 2, and for time to failure data, through a Weibull likelihood, in Section 3. We illustrate the approach in the context of examples, first for failure on demand data from...
emergency diesel generators in Section 4.1 and then for time to failure data on pressure vessels in Section 4.2. The article concludes with Section 5, which provides a summary of the approach and some areas for further work.

2. Binomial Reliability Demonstration Testing

2.1. Traditional Approaches

Suppose that we have $n$ items which we are to put on test and that $Y$ is the number which will fail the test. Also define $\pi$ to be the probability that an item survives the test. In reliability demonstration testing, for a given $n$, we define a maximum allowed number of failures $c$. If the actual number of failures in the test exceeds $c$ then the test is failed. If not, the test is passed.

Therefore, an RDT plan (Hamada et al. 2008) is given by the pair $(n, c)$. Provided that the failures are independent and identically distributed, the likelihood associated with the test is binomial

$$Y \mid \pi \sim \text{bin}(n, 1 - \pi).$$

Traditionally, an RDT would be analyzed using a hypothesis test where $H_0 : \pi = \pi_T$, the quantity $\pi_T$ is the target reliability and $H_1 : \pi > \pi_T$. For a $100(1 - \alpha)$% critical level, we would reject $H_0$ when $\Pr(Y \leq y \mid \pi = \pi_T) \leq \alpha$. That is, if

$$\sum_{y=0}^{c} \binom{n}{y} (1 - \pi_T)^{n-y} \pi_T^y \leq \alpha. $$

We choose $c$ to be the largest value of $y$ satisfying this inequality.

In the case of risk criteria, test plans are chosen to keep the probabilities of making the wrong conclusions from the RDT small. The two errors considered are named the producer’s risk, which is associated with a failed RDT for a product which is accepted. The two errors considered are named the consumer’s risk, which is associated with a passed RDT for an item which does not meet the reliability target. Classical risk criteria (Tobias and Trindade 1995), average risk criteria (Easterling 1970) and (Bayesian) posterior risk criteria (Hamada et al. 2014) have been proposed. The details of the posterior risk criteria for binomial reliability demonstration testing are given in Section B of the supplementary materials. A summary of the producer’s risk and consumer’s risk for each of the approaches is given in Table 1. In the table, $\pi_0$ is defined as the acceptable reliability level and $\pi_1$ is defined as the rejectable reliability level.

Often in reliability demonstration testing, test plans can involve running large numbers of very reliable items for long periods of time. Meeker and Escobar (2004) proposed using past observations $x_i$ in the analysis to reduce $n$. They named the new tests which incorporated past observations reliability assurance assurance tests.

| Approach | Producer’s risk | Consumer’s risk |
|----------|-----------------|-----------------|
| Classical | $\Pr(\text{Test is failed} \mid \pi = \pi_0)$ | $\Pr(\text{Test is passed} \mid \pi = \pi_1)$ |
| Average  | $\Pr(\text{Test is failed} \mid \pi \geq \pi_0)$ | $\Pr(\text{Test is passed} \mid \pi \leq \pi_1)$ |
| Posterior| $\Pr(\pi \geq \pi_0 \mid \text{Test is failed})$ | $\Pr(\pi \leq \pi_1 \mid \text{Test is passed})$ |

2.2. Assurance

The posterior risk criteria approach conflates two distinct processes: the choice of sample size for the test and the analysis to be undertaken once the test has been conducted. A result of this is that the prior used in the analysis stage is the same as that used in the design stage. This may not be appropriate.

A Bayesian alternative to the posterior risk criteria, which does not have these properties, is to use assurance (Spiegelhalter, Freedman, and Parmar 1994; O’Hagan, Stevens, and Campbell 2005; Ren and Oakley 2014). Assurance chooses a sample size based on the answer to the question, “What is the probability that the RDT is going to result in a successful outcome?” The probability of a successful test (Hamada et al. 2014; Lu, Li, and Anderson-Cook 2016) is given by

$$\Pr[\text{Successful test}] = \int_0^1 \Pr(Y \leq c \mid \pi) p(\pi) d\pi$$

$$= \int_0^1 \sum_{y=0}^{c} \binom{n}{y} (1 - \pi)^y \pi^{n-y} p(\pi) d\pi,$$

where $p(\pi)$ is the prior probability density for $\pi$ and $\Pr(Y \leq c \mid \pi)$ is the cumulative distribution function of $Y$ evaluated at $c$, which in this case is binomial. We would like this probability to be large. We may wish to choose a minimum acceptable value, $\gamma$, for the assurance.

Lu, Li, and Anderson-Cook (2016) considered assurance, which they called the acceptance probability, as part of a multicriteria optimization approach to the design and analysis of binomial RDTs. The other criteria considered were the posterior risk criteria and the cost of testing. A single prior distribution was used, taken from the point of view of the producer.

Suppose we have data from previous tests of the form $x_i$ for $i = 1, \ldots, I$ where $x_i \sim \text{bin}(n_i, 1 - \pi_i)$, and the probabilities of items surviving the test in each case can be thought of as coming from the same prior distribution $\pi_i \sim \text{beta}(a, b)$, with hyperparameters $a, b$. For example, they could be identical components produced in the same factory being tested at different locations. We can use these historical data in the design of the RDT using a hierarchical prior. That is, we suppose that our probability of interest also comes from the same prior distribution $\pi \sim \text{beta}(a, b)$ and we define a hyperprior distribution over $(a, b)$. Then, when we observe $x_i$, we learn about $(a, b)$, and update our prior for $\pi$. The assurance is now

$$\Pr[\text{Successful test} \mid x]$$

$$= \int_0^1 \Pr(Y \leq c \mid \pi) p(\pi \mid x) d\pi$$

$$= \int_0^1 \sum_{y=0}^{c} \binom{n}{y} (1 - \pi)^y \pi^{n-y} p(\pi \mid x) d\pi.$$

We can sample from the posterior distribution $p(\pi \mid x)$ using Markov chain Monte Carlo (MCMC) in the following way.

1. Generate $N$ posterior draws of $a, b$ of the form $a^{(j)}, b^{(j)}$ for $j = 1, \ldots, N$.

2. For $j = 1, \ldots, N$, draw $\pi^{(1)}, \ldots, \pi^{(N)}$ as $\pi^{(j)} \sim \text{beta}(a^{(j)}, b^{(j)})$. 



Using the draws of $\pi$ from the posterior distribution we can evaluate the probability of a successful test, via Monte Carlo integration, as

$$\Pr[\text{Successful test} \mid x] \approx \frac{1}{N} \sum_{j=1}^{N} \left[ \sum_{y=0}^{c} \frac{n!}{y!(n-y)!} (1 - \pi^{(j)})^{n-y} \pi^{n-y} \right].$$

For any sample size $n$, once we have a way to choose $c$, we can use assurance in this way to choose $n$. The critical number of failures $c$ is chosen based on the analysis to be carried out following the test. This flexible approach allows either a frequentist hypothesis test or a Bayesian analysis to be used to decide on the success or failure of the test. If a Bayesian approach is to be used to analyze the test data, the prior for the analysis can differ from that used here in the design.

### 2.3. Cutoff Choice

#### 2.3.1. Exact Binomial Test

Consider the binomial test in Section 2.1. We would reject $H_0$ given $Y = y$ if $\sum_{y=0}^{Y} \binom{n}{y}(1 - \pi)^y\pi^{n-y} \leq \alpha$, with, for example, $\alpha = 0.05$. Therefore, we simply select $c$ to be the largest value of $y$ for which this is true. An alternative to the exact binomial test would be to use a normal distribution to approximate the binomial distribution.

#### 2.3.2. Bayesian Approaches

Suppose that, in the analysis of the test result, we have a prior probability density $p_A(\pi)$ for $\pi$. This may be different from the design prior $p(\pi)$. We may choose a rule that, if the posterior probability that $\pi \leq \pi_T$ is small, then the test is passed. That is, the test is passed if $\Pr_A(\pi \leq \pi_T \mid Y = y) \leq \delta$, where $\Pr_A$ denotes the probability based on the analysis prior and, for example, $\delta = 0.05$.

That is, once we have conducted the RDT, we will have observed a number of failures, $Y = y$. We update the probability under the analysis prior $\Pr_A(\pi \leq \pi_T)$ to a posterior probability $\Pr_A(\pi \leq \pi_T \mid Y = y)$ in the standard way using Bayes’ theorem. If this posterior probability is less than a threshold, $\delta$, then the RDT is passed. The maximum number of allowed failures $c$ is then the largest value of $y$ for which $\Pr_A(\pi \leq \pi_T \mid Y = y) \leq \delta$. This relies on the following proposition, the proof of which is given in Section C of the supplementary materials.

**Proposition 1.** If the posterior probability $\Pr(\pi \leq \pi_T \mid Y = y)$ is controlled at or below the threshold $\delta$, then the posterior probability $\Pr(\pi \leq \pi_T \mid Y = y)$ at any $y \leq c$ is also controlled below $\delta$.

Thus, this approach is not based on risk criteria. We are not choosing the maximum allowed number of failures to keep the risk of making an error small, but so that the consumer’s probability, following the RDT, that the product does not meet the reliability target is small. The tail probability $\Pr(\pi \leq \pi_T \mid Y \leq c)$ as used in risk criteria is not appropriate as we do not observe $Y \leq y$ in a RDT. We observe $Y = y$.

So we choose $c$ to be the largest value of $y$ for which $\Pr_A(\pi \leq \pi_T \mid Y = y) \leq \delta$. This posterior probability, at the maximum allowed number of failures $c$, can be calculated (similar to Hamada et al. 2008) as

$$\Pr_A(\pi \leq \pi_T \mid Y = c) = \int_{0}^{\pi_T} f(c \mid \pi)p_A(\pi)d\pi = \int_{0}^{\pi_T} f(c \mid \pi)p_A(\pi)d\pi$$

where $f(c \mid \pi)$ is the probability mass function of $Y$ evaluated at $c$. For example, if the prior distribution for $\pi$ was chosen to be $\pi \sim \text{beta}(\alpha, \beta)$ then the posterior distribution is $\pi \mid Y = c \sim \text{beta}(\alpha + n - c, \beta + c)$ and $c$ is chosen using

$$\Pr_A(\pi \leq \pi_T \mid Y = c) = \frac{B_I(\pi_T; \alpha + n - c, \beta + c)}{B(\alpha + n - c, \beta + c)}$$

where $B_I(\cdot; \cdot)$ and $B(\cdot; \cdot)$ are the incomplete beta function and beta function, respectively. This approach produces an RDT plan, as the data have been used to calculate the sample size $n$ only and not to analyze the results of the test. A reliability assurance test plan can also be produced in this way by using the data $X = x$ here in the choice of $c$. This can be achieved using MCMC in the same way as when calculating $n$ from known $c$ in Section 2.2.

### 2.4. Choice of Priors

To use assurance to decide on the optimal test plan $(n, c)$ we need a design prior distribution for $\pi$, $p(\pi)$, to be used in the assurance calculation and, if a Bayesian analysis is to be undertaken following the test, a second prior distribution on $\pi$, $p_A(\pi)$, to be used in the analysis. In this section, we indicate some suitable prior distributions. However, the approach is general, and any sensible prior distributions may be used.

The design prior distribution to be used in the assurance calculation, $p(\pi)$, should represent the beliefs of the producer of the items on test. They take the risk associated with the failure of the test and so it is their probability that the test will be a success which will specify the sample size. The design prior distribution therefore needs to be defined in terms of quantities about which we could reasonably ask an engineer. Winkler (1967) described four methods for eliciting a prior distribution for the parameter of a binomial distribution. One of these is the equivalent prior sample method. With the usual conjugate beta prior distribution, $\pi \sim \text{beta}(a, b)$, we can write $a = mp$ and $b = m(1 - p)$ where $p$ is the engineer’s subjective expectation of $\pi$ and $m$ is the engineer’s assessment of the size of the “prior sample” on which this judgment is based. We give further details and alternatives in Section D of the supplementary materials. See also Garthwaite, Kadane, and O’Hagan (2005).

If required, to allow us to use historical data to inform the design prior we can formulate a hierarchical structure as follows. Suppose that we have $J$ samples of data on examples of the product. Sample $j$ has size $n_j$ and was collected under conditions $r_j$. The number of failures in sample $j$ is $y_j$. The data in the test will effectively be sample $j + 1$ and will correspond to conditions $r_{j+1}$. The reliability under conditions $r_j$ is $\pi_j$. For each $j$, we regard $y_j$ as an observation on the random variable $Y_j$ where the conditional distribution of $Y_j$ given $\pi_j$ is $\text{bin}(n_j, 1 - \pi_j)$. Then,
given the values of \( m \) and \( p \), we have \( \pi | m, p \sim \text{beta}(mp, m(1 - p)) \). Finally we give independent hyper-prior distributions to \( m \) and \( p \) with \( p \sim \text{beta}(a_p, b_p) \) and \( m \sim \text{gamma}(a_m, b_m) \). Guidance on the choice of the hyperparameters is given in Section D of the supplementary materials.

In the analysis following the test, basing the prior on the beliefs of the producer would typically be a controversial choice. The purpose of an RDT is to demonstrate to someone else, other than the producer, that the product meets a certain reliability. The analysis prior therefore represents the beliefs of the consumer. The consumer would not have access to the same historical data as the producer. They are also likely to be more skeptical as to the reliability of the product prior to the RDT than the producer. Therefore, it may be more reasonable in designing the test to suppose that a relatively conservative analysis prior will be used. Spiegelhalter, Abrams, and Myles (2004) suggested an approach to this: the skeptical prior. In the absence of being able to ask the consumer for their prior, a skeptical prior is a reasonable choice, as it assumes conservative beliefs for the consumer. The skeptical prior should be designed so that there is only a small prior probability that \( \pi > \pi_T \). So, for example, if a beta(\( \alpha, \beta \)) distribution is assumed for \( p_A(\pi) \), then we might choose \( (\alpha, \beta) \) to satisfy \( 1 - \frac{B(\pi_T; \alpha, \beta)}{B(\alpha, \beta)} = \delta \), where \( \delta = 0.05 \).

An alternative to a simple beta distribution which provides more flexibility in the choice of the consumer’s prior is to use a mixture distribution for the analysis prior. This could take the form

\[
p_A(\pi) = \sum_{k=1}^{K} q_k p_{A,k}(\pi),
\]

where \( p_{A,k}(\pi) \) is the density of mixture component \( k \), \( q_k > 0 \) for \( k = 1, \ldots, K \) and \( \sum_{k=1}^{K} q_k = 1 \). Some comments on how the number of components, \( K \), and the other parameters might be chosen are provided in Section D of the supplementary materials.

In this case, the posterior probability to assess whether the test is passed becomes

\[
\text{Pr}(\pi \leq \pi_T \mid Y = c) = \frac{\int_{Q}^{\pi_T} f(c \mid \pi) \sum_{k=1}^{K} q_k p_{A,k}(\pi) \, d\pi}{\int_{0}^{\pi_T} f(c \mid \pi) \sum_{k=1}^{K} q_k p_{A,k}(\pi) \, d\pi},
\]

and if each of the component distributions in the mixture is a beta distribution, then this can be expressed in terms of beta functions as previously.

Assuming that the producer does not know what the consumer’s prior is, another way to think of this is that the producer has a joint prior over \( \pi \) and the consumer’s prior, or, at least, the parameters of the consumer’s prior. For example, the producer might think that the consumer’s prior density is one of \( p_{A,1}(\pi), \ldots, p_{A,K}(\pi) \) and assign probability \( q_k \) to \( p_{A,k}(\pi) \). Given that the consumer actually has prior \( k \), this leads to a cut-off \( c_k \), exactly as for a single-component prior. So we now have

\[
\text{Pr}([\text{Successful test} \mid x] = \sum_{k=1}^{K} q_k \int_{0}^{\pi} \text{Pr}(Y \leq c_k \mid \pi) p(\pi \mid x) \, d\pi).
\]

In effect, the producer now has a probability distribution for the cut-off \( c \) which the consumer will use, with \( \text{Pr}(c = j) = u_j \), \( J_l \leq j \leq J_u \), where \( J_l \) and \( J_u \) are the minimum and maximum values taken by \( c \) and \( u_j = \sum_{k=j}^{K} q_k \). So

\[
\text{Pr}(\text{Successful test} \mid x) = \sum_{j=J_l}^{J_u} \left[ u_j \int_{0}^{\pi} \text{Pr}(Y \leq j \mid \pi) p(\pi \mid x) \, d\pi \right] = \sum_{j=0}^{J_u} U_j \delta_j,
\]

where \( \int_0^{\pi} \text{Pr}(Y = j \mid \pi) p(\pi \mid x) \, dx = \delta_j \) and \( U_j = \sum_{k=j}^{K} u_k \).

3. Weibull Reliability Demonstration Testing

3.1. Assurance

In binomial demonstration testing, we are interested in items which are required to function on demand. For items which are expected to function continuously for long periods of time Weibull demonstration testing is more appropriate. In this case, we would put \( n \) items on test, and record their failure times \( t_i \), for \( i = 1, \ldots, n \). We assume that the time to failure of items follows a Weibull distribution \( T \mid \rho, \beta \sim \text{Weibull}(\rho, \beta) \) with shape parameter \( \beta \), scale parameter \( \rho \), and probability density function \( f(t \mid \rho, \beta) = \rho \beta (t)^{\beta - 1} \exp \left[ -\rho(t)^{\beta} \right] \).

To decide on an RDT plan, we need to decide on a metric to assess the reliability of the product. Consider the failure time distribution \( F(\cdot) \), and suppose that we were interested in a time \( t_q \) such that \( F(t_q) = 1 - q \). Then \( t_q \) is the reliable life of the product for \( q \), that is, the time beyond which the proportion \( q \) of the items survive. We can specify some target \( t_{q,*} \) for the reliable life at \( q \). The test plan is the number of items to put on test, \( n \), and possibly a censoring time \( t_c \). The items will be tested until either they fail or the censoring time is reached.

However, for highly reliable items, testing until failure in normal operating conditions may not be feasible. In this case, accelerated testing can be done, in which items are tested at a much greater stress (e.g., temperature, pressure, vibration) than that of typical use. Suppose we are to conduct the test at stresses \( s_{\text{test}} = (s_{\text{test,1}}, \ldots, s_{\text{test,n}}) \) and that our target reliable life is specified under a (typically lower) stress \( s_\tau \). We can relate the failure times under different stresses using a link function,

\[
\log(\rho) = g(s, \theta),
\]

where the elements of \( \theta \) are the parameters of the relationship. We can express the reliable life in terms of the parameters of the Weibull distribution. Doing so gives

\[
t_q = \rho^{-\frac{1}{\beta}} \left[ -\log(q) \right]^\frac{1}{\beta}.
\]
The assurance is

\[ \Pr(\text{Successful test}) = \int_0^\infty \int_0^\infty \Pr(\text{Test passed} \mid \rho, \beta) \times p(\rho, \beta) \ d\rho \ d\beta, \]

where \( p(\rho, \beta) \) is a joint design prior density for \( \rho, \beta \). We could analyze the test results using a hypothesis test in a similar way to the binomial case. A suitable test in this context would be a maximum likelihood ratio test (Collett 1994) with the null hypothesis being \( q = q_{\text{rs}} \), and the alternative being \( q > q_{\text{rs}} \).

However, if the analysis following the RDT is to be Bayesian, one criterion which we could use to declare the test passed is whether the analysis posterior probability that \( q \geq q_{\text{rs}} \) under stress \( s_{\text{rs}} \), given the observations in the test, under stresses \( s_{\text{test}} \), is large. That is, the test is passed if \( \Pr_A(q \geq q_{\text{rs}} \mid t) \geq 1 - \delta \), where \( \delta = 0.05 \), for example. To evaluate this quantity we could use MCMC. Note that, to make inference about \( \theta \), we need to use more than one stress. Let \( t = (t_1, \ldots, t_n) \) be the observed failure or censoring times, \( z = (z_1, \ldots, z_n) \) be the censoring indicators, with \( z_i = 1 \) if the failure is observed and \( z_i = 0 \) if it is censored, and \( \rho_i = \exp[g(s_{\text{test}i}, \theta)] \) be the stress-dependent scale parameter for observation \( i \). The likelihood function is

\[ L(t \mid \rho, \beta) = \prod_{i=1}^n p(\rho, \beta) p(t_i, s_i) \exp[-(\rho_i t_i^\beta)]. \]

Suppose that the analysis prior distribution is given by \( p_{A}(\beta, \theta) \) and that we have observed failure times \( t \) under stresses \( s_{\text{test}} \). Then we could assess the RDT criterion above as follows.

1. Sample \( \beta_i(0), \theta_i(0), i = 1, \ldots, N_1 \), from their analysis posterior distribution using MCMC.

2. Evaluate \( \rho_i(0) = \exp[g(s_{\text{test}i}, \theta)] \) and then find \( q_i(0) = \rho_i(0)[-\log(q)]^{1/\theta} \).

3. Approximate the posterior probability as \( \Pr_A(q_i \geq q_{\text{rs}}) \approx \frac{1}{N_1} \sum_{i=1}^{N_1} I(q_i \geq q_{\text{rs}}) \), where \( I(q_i \geq q_{\text{rs}}) \) is an indicator function which takes the value 1 if \( q_i \geq q_{\text{rs}} \) and 0 otherwise.

4. Assess whether \( q \geq \Pr_A(q \geq q_{\text{rs}}) \geq 1 - \delta \).

A naïve approach to assess the assurance would be to sample \( M \) sets of parameters \( (\rho^{(k)}, \beta^{(k)}) \), \( k = 1, \ldots, M \) from the design prior distribution and, for each, sample \( N_2 \) sets of hypothetical data \( t^{(k)}, j = 1, \ldots, N_2 \) from the likelihood \( L(t \mid \rho^{(k)}, \beta^{(k)}) \). The assurance would then be given by the Monte Carlo approximation

\[ \Pr(\text{Successful test}) \approx \frac{1}{N_2 \times M} \sum_{k=1}^M \sum_{j=1}^{N_2} I(q_j \geq 0.95 \mid t^{(k)}), \]

where \( I(q_j \geq 1 - \delta \mid t^{(k)}) \) is an indicator variable which takes the value 1 if \( q_j \geq 1 - \delta \) and 0 otherwise. This would involve \( M \times N_1 \times N_2 \times n_{\text{max}} \) calculations to evaluate the assurance for all sample sizes in the range \( n = 1, \ldots, n_{\text{max}} \). The sample size \( n \) would be chosen to be the smallest value which meets a specified level of assurance.

A more efficient way to assess the assurance is to adapt the numerical scheme of Müller (1999) and Müller and Palmer (1996). In this case, we would proceed as follows:

1. Select a number \( M \) of sample sizes \( n_j \in [1, \ldots, n_{\text{max}}] \), for \( j = 1, \ldots, M \).

2. Simulate

\[ (\rho^{(j)}, \beta^{(j)}) \sim p(\rho, \beta), \]

\[ t^{(j)} \sim L(t \mid \rho^{(j)}, \beta^{(j)}), \]

where \( p(\rho, \beta) \) is the design prior, \( L(t \mid \rho, \beta) \) is the Weibull likelihood and \( t_{2j} \) is a vector of failure times of length \( n_j \).

3. Evaluate the vector \( x \), where

\[ x^{(j)} = I \left[ \Pr_A(q_j \geq q_{\text{rs}} \mid t^{(j)}_{2j}) \geq 1 - \delta \right], \]

where \( \Pr_A(q_j \geq q_{\text{rs}} \mid t^{(j)}_{2j}) \) is the posterior probability that \( q_j \geq q_{\text{rs}} \) based on the analysis prior. Note that the notation here is slightly different from that used previously.

4. We repeat Steps 2 and 3 a small number of times to obtain a proportion \( \hat{p}^{(j)} \) of successes for sample size \( n_j \).

5. A smoother is used to fit a curve to \( \hat{p}^{(1)}, \ldots, \hat{p}^{(M)} \). This is an estimate of the assurance.

This greatly reduces the number of computations required to evaluate the assurance. An alternative would be to use an augmented MCMC scheme (Müller 1999; Cook, Gibson, and Gilligan 2008). This is an area for future work.

Suppose that the producer has a censoring time for the RDT, \( t_C \), representing the maximum total time an item can be put on test. We can incorporate this censoring time into the numerical scheme by replacing any time \( t_i^{(j)} > t_C \) with \( t_C \) in Step 2, and using the correct contribution to the likelihood for the posterior probability in Step 3. If required, the producer can also evaluate the effect on the assurance of different choices of \( t_C \), as well as the sample size.

We are able to incorporate past observations into the assurance calculation. Suppose we have historical observations \( (t_{i,j}, s_{i,j}) \), which represent times to failure and stress levels of items \( j = 1, \ldots, n \) at locations \( i = 1, \ldots, m \), and these observations follow Weibull distributions \( t_{i,j} \mid \rho_{i,j}, \beta \sim \text{Weibull} (\rho_{i,j}, \beta) \) with scale parameter \( \rho_{i,j} \), where \( \log(\rho_{i,j}) = g(s_{i,j}, \theta_i) \), and a common shape parameter \( \beta \). We can use a hierarchical structure to learn about the parameters of the current test by giving \( \theta_i \) prior distributions \( p(\theta_i) \) with common hyperparameters \( p(\rho, \beta) \) having prior distribution \( p(\theta_i, \eta) \) and specifying a prior distribution on \( \rho, \beta \). We further suppose that the current test has the same prior structure with common \( p(\rho, \beta) \), \( p(\theta_i) \). We incorporate the historical information into the assurance calculation via

\[ \Pr(\text{Successful test} \mid \tilde{t}) = \int_0^\infty \int_0^\infty \Pr(\text{Test passed} \mid \rho, \beta) \times p(\rho, \beta \mid \tilde{t}) \ d\rho \ d\beta, \]

where \( p(\rho, \beta \mid \tilde{t}) \) is the joint design posterior distribution for \( \rho, \beta \) given the observations \( \tilde{t}_{ij} \). We evaluate this assurance in the same way as in the case with no previous observations, with the only change being that we sample \( \beta^{(j)}, \theta^{(j)} \) values from the design posterior distribution using MCMC rather than the design prior distribution in Step 2.
3.2. Choice of Priors

To detail the prior elicitation and specification, we need to specify the function \( g(\cdot, \cdot) \). This is part of the model specification, but for illustration we choose to use the general class
\[
g(s, \theta) = \alpha_0 + \alpha_1 s^k + \epsilon,
\]
where \( \alpha_0, \alpha_1 \) are intercept and slope terms and \( \epsilon \) is a location-specific random effect. This is known as the accelerating relationship in accelerated life testing. Commonly used accelerating relationships are linear, Arrhenius, Eyring, and logarithmic. The general form presented above includes the linear relationship when \( k = 1 \), the Arrhenius relationship when \( k = -1 \) and the logarithmic relationship when \( k = 1 \) and \( s \) is defined as the natural logarithm of the stress level. The Eyring relationship is an extension of the Arrhenius relationship and would require an extra term in the function \( g(s, \theta) \).

For the design prior we wish to elicit the beliefs of engineers about the reliability of the items on test. To do so, we need to ask them questions about observable quantities. As a result of the more complex model structure in the Weibull case than the binomial case, we split the elicitation and specification task into three stages, each of which is outlined below. However, before that, we outline a suitable structure for the prior. Consider the link function above. We choose to give the regression parameters \( (\alpha_0, \alpha_1) \) a bivariate normal prior distribution \( \alpha = (\alpha_0, \alpha_1)^T \sim \text{BNV}(\mu, \Sigma) \), for hyper-mean vector \( \mu = (\mu_0, \mu_1)^T \) and variance matrix \( \Sigma \) with diagonal elements \( (\sigma_{00}, \sigma_{11}) \) and off-diagonal element \( \sigma_{01} \). We give the location-specific effects zero-mean normal prior distributions with a common variance, \( \epsilon_i \sim N(0, \nu_e) \), where \( \nu_e \) is a hyperparameter to be chosen. We complete the specification by giving the Weibull shape parameter, \( \beta \), a suitable prior, \( \beta \sim \text{Gamma}(a_\beta, b_\beta) \).

The result is that there are eight hyperparameters to be specified in the design prior. We choose to do this using questions about quantiles of the lifetime distribution in terms of a hypothetical large future sample so that the empirical quantiles are, in principal, observable and, since the sample is large, aleatory uncertainty is dominated by epistemic uncertainty. We rescale the stress values so that \( s = 1 \) is a plausible value.

Stage 1: We ask the expert to suppose that stress is at a specified level \( s \). We ask the expert for their lower quartile, median and upper quartile for the reliable life for two different values of \( q \). For example, we ask them about the times by which 1/3 and 2/3 of items will have failed. These are
\[
\tau_q = \exp \left\{ - (\alpha_0 + \alpha_1 s^k + \epsilon_i) \right\} \left[ - \log(q) \right]^{1/\beta},
\]
for \( q = 1/3, 2/3 \). Specifically, we ask the expert to make judgments about the ratio
\[
\frac{\tau_{2/3}}{\tau_{1/3}} = \left[ \frac{\log(2/3)}{\log(1/3)} \right]^{1/\beta}.
\]

A value for this ratio gives a value for \( \beta \):
\[
\beta = \frac{\log[\log(2/3)/\log(1/3)]}{\log[\tau_{2/3}/\tau_{1/3}]}.
\]

By eliciting the expert’s quartiles for the ratio, we obtain quartiles for \( \beta \) and we can then choose \( a_\beta \) and \( b_\beta \) to match these.

Stage 2: We now ask the expert to consider two different locations, \( i, j \), and the reliable life at these locations for the same stress and the same value of \( q \), for example, \( q = 1/2 \). Then, if \( \tau_{q,i}(s) \) is the reliable life with probability \( q \) at location \( i \) and stress \( s \), we have
\[
\frac{\tau_{q,i}(s)}{\tau_{q,j}(s)} = \exp[\epsilon_i - \epsilon_j].
\]
The expert’s quartiles for this ratio lead to quartiles for \( \epsilon_i - \epsilon_j \), which has variance \( 2\nu_e \), and hence to a value for \( \nu_e \).

Stage 3: We have
\[
\log[\tau_{q,i}(s)] = - (\alpha_0 + \alpha_1 s^k + \epsilon_i) + \beta^{-1} \log[-\log(q)].
\]
If we choose \( q = e^{-1} \) then the second term on the right vanishes. In elicitation, the value \( q = 1/3 \) is more practical and gives a reasonable approximation since \( | \log[-\log(1/3)] | < 0.1 \), unless \( \beta \) is very small. Let \( M_p[\tau_{q,i}(s)] \) denote the expert’s \( p \)-quantile for \( \tau_{q,i}(s) \). Then \( \log[M_{1/2}[\tau_{1/3}(1)](s)] \) gives \( \alpha_0 + \alpha_1 \log[\tau_{1/2}(1)](s) \) and \( \log[M_{1/2}[\tau_{1/3}(3)](s)] \) for \( s \neq 1 \) gives \( \alpha_0 + \alpha_1 s^k \) so we obtain \( \alpha_0 \) and \( \alpha_1 \). For the variances we can use, for example, the expert’s quartiles. So, from \( \log[M_{1/4}[\tau_{1/3}(1)](s)] \) and \( \log[M_{3/4}[\tau_{1/3}(3)](s)] \), we obtain \( \sigma_{00}^2 + \sigma_{11}^2 + 2\sigma_{01} \nu_e \) and then, using two different stresses \( s_1 \) and \( s_2 \) we obtain \( \sigma_{00}^2 + \sigma_{11}^2 s_1^k + 2\sigma_{01} s_1^k \nu_e + \nu_e s_2 \). Thus, we obtain three simultaneous equations, from which we can find \( \sigma_{00}^2 \), \( \sigma_{11}^2 \), and \( \sigma_{01}^2 \).

In the case where we are to observe historical lifetime data in other locations which can be combined with the design prior distribution to form a design posterior distribution, we may wish to learn about the value of \( \nu_e \). In this case, we give \( \nu_e \) an inverse gamma prior distribution so \( \nu_e^{-1} \sim \text{Gamma}(a_\nu, b_\nu) \), where \( (a_\nu, b_\nu) \) are hyperparameters to be chosen. We modify Stage 2. The expert’s predictive distribution for \( T = (\epsilon_i - \epsilon_j)/\sqrt{2b_\nu/a_\nu} \) is a Student’s \( t \)-distribution on \( 2a_\nu \) degrees of freedom. To find both \( a_\nu \) and \( b_\nu \), we need two quantiles corresponding to probabilities \( q_1 \) and \( q_2 \) with \( q_1 \neq q_2 \) and \( q_i \neq 1/2 \) for \( j = 1, 2 \), for example, \( q_1 = 0.6 \) and \( q_2 = 0.8 \). Alternatively the expert might imagine a large number of such ratios \( \tau_{q,1}(s)/\tau_{q,2}(s) \) and give quartiles for the empirical upper quartile of the ratios.

For the analysis prior, we can assume the same prior structure. A skeptical analysis prior would have hyperparameters which give small prior probability to the event \( \tau_{q,\alpha} > \tau_{q,\alpha} \) under stress \( s_\alpha \). As in Section 2.4, another possibility is for the producer to assign probabilities to possible consumer’s prior distributions.

4. Examples

4.1. Example 1: A Binomial Reliability Demonstration Test

We consider an example from Martz, Kram, and Abramson (1996). The values used in this example are illustrative. However, for important quantities, we describe how the specification could be made in practice. We would like to demonstrate the reliability of an emergency diesel generator in a nuclear power plant, with a target reliability of \( \pi_T = 0.96 \). We will compare assurance based on three analysis methods: the exact binomial
test, the Bayesian approach using a skeptical analysis prior and
the Bayesian approach using a mixture analysis prior. In the
binomial test we use the exact test from Section 2.3.1, choosing
c to reject $H_0$ if $p < 0.05$ and in the Bayesian approaches we
choose $c$ so that the RDT is passed if $Pr_A(\pi \leq \pi_T | Y = y) \leq 0.05$. That is, for each sample size $n$, we find the maximum
allowed number of failures $c$ to be the largest value of $y$ for which
each of these conditions is true.

A beta distribution is used for the skeptical analysis prior with
$\alpha = 7.8$ and $\beta = 2.0$. These values are chosen to give a
small a priori probability of meeting the reliability target. They are
chosen so that $Pr_A(\pi > 0.96) = 0.05$ and the prior mean of $\pi$ is 0.8. The skeptical prior density is shown in Figure 1. A
possible disadvantage of this prior is that the mode, at $\pi = 0.87$,
is somewhat to the left of the target value so that values of $\pi$ just
a little less than 0.96 are relatively less favored by the prior. As
an alternative, we also consider a mixture prior, the density of
which is also shown in Figure 1. This prior was chosen to give
an almost uniform density over the range $0.80 < \pi < 0.96$ and
with the density decreasing, roughly linearly, as $\pi$ decreases
from 0.8 to 0.6. This is achieved by using a large number of
components, in this case $K = 40$. The components are chosen
so that they have modes 0.60, 0.61, ..., 0.99. The weights on the
components increase linearly between $\pi = 0.6$ and $\pi = 0.8$. Component $k$ is a beta($a_k, b_k$) distribution which has its mode
at $\pi = (a_k - 1)/(a_k + b_k - 2)$. We set $a_k + b_k = d = 302$ for all $k$. A constant value of $d$, combined with equal weights and equal
spacing gives the approximate uniformity where required. The
choice of a large value of $d$ makes the changes in the density
at $\pi = 0.6$, $\pi = 0.8$ and $\pi = 0.99$ relatively sharp. Thus,$a_k = (d - 2)(0.59 + 0.01k) + 1 = 178 + 3k$ and $b_k = 302 - a_k$. Our mixture prior gives $Pr(\pi > 0.96) = 0.11$.

The posterior distribution for the mixture given $y$ successes in $n$ trials is of the same form as the prior with the prior weights
$p_k^{(0)}$, updated to

$$p_k^{(1)} = \frac{\hat{p}_k}{\sum_{j=1}^{K} \hat{p}_k},$$

where

$$\hat{p}_k = p_k^{(0)} \frac{\Gamma(a_k + b_k)}{\Gamma(a_k + b_k + y)} \frac{\Gamma(a_k + n - y)}{\Gamma(a_k + b_k + n)},$$

and the updated beta distribution parameters are $A_k = a_k + n - y$ and $B_k = b_k + y$ for $k = 1, \ldots, K$. In our case $K = 40$ and
$a_k + b_k = 302$.

In the design prior $p(\pi)$ we give $m$, $p$ gamma$(200, 1)$ and
beta$(78, 2)$ priors, respectively. This results from an elicitation
process which starts by choosing the prior mean for $p$ of 0.975.
Then, by thinking about a hypothetical future sample and our
likely degree of uncertainty, we choose quantiles for $m$ and hence
find values for $a_m$ and $b_m$. Finally, by considering the correlation
in our beliefs about two hypothetical future samples, we are able
to find a value for $a_p + b_p$, which is combined with our assessed
prior mean for $p$, which is $a_p/(a_p + b_p)$, to find $a_p$ and $b_p$. Further
details are given in Section D of the supplementary materials.

Based on these specifications, we can plot the sample size $n$
against the assurance for each of the three approaches. This
is given in the left-hand side of Figure 2. The binomial test
is shown by the solid line, the skeptical prior by the dashed
line and the mixture prior by the dotted line. We see that the
mixture prior and the binomial test give the highest assurance
for any particular sample size and the skeptical prior the lowest
assurance. As we have a discrete distribution (binomial), then
we cannot achieve exactly a 5% posterior probability that the
product does not meet its reliability target with any choice of $c$. Therefore, $c$ is the largest number of failures for which
the posterior probability is below 0.05. As the sample size increases
and $c$ remains constant the assurance drops, which causes the
drops in the plots. Each time $c$ increases by 1 the assurance
jumps to a higher value than the previous jump. This causes the
increasing peaks in the plots.

To achieve an assurance of 50% in this case would require $n = (227, 279, 228)$ under the three respective methods, with
maximum allowed numbers of failures of $c = (4, 6, 4)$, respec-
tively. We cannot achieve an assurance of greater than 80% in
this case, as this is the prior probability that the target reliability
will be met under the design prior. Note that the assurance is
slowly converging toward 80%. For example, with a sample size
of $n = 10,000$ we achieve an assurance of 76.8% and with a
sample size of $n = 100,000$ we achieve an assurance of 79.4%.

However, we have data on the behavior on demand of gener-
ators of the same type at other plants. The data represent tests of
63 generators in nuclear power plants. In each case, the number
of failures on demand of the generator out of the total number
of demands was recorded. The number of demands and the
proportion of failures in the tests are given in Figure 3. We see
that we have some fairly large test datasets and that the failure
proportions recorded in the tests are very low, typically below
3%.

Using the same design prior as above, we generate 10,000
samples of $(p, m)$ from the posterior distribution using rjags
(Plummer 2016), having discarded 1000 samples as burn in.
The design posterior distribution of $\pi$, the probability that an
emergency generator will work on demand, is given in Figure 4
(solid line) alongside the design prior distribution for $\pi$ (dashed
line). We see that the posterior distribution of the probability
of success of a generator is tightly concentrated around very
high probabilities, with posterior mean and median of 0.990
and 0.992, respectively. The prior distribution is more diffuse,
although much of the density is still concentrated around high
probabilities.

We are able to calculate the assurance for each of the three
methods as above using this posterior distribution for $\pi$ in
place of the design prior. The results for different values of the

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1}
\caption{The skeptical prior and the mixture prior for the probability of a generator working on demand. The target reliability is shown by the vertical dotted line.}
\end{figure}
Figure 2. The sample size $n$ against the assurance for the binomial test (solid line), the skeptical prior (dashed line), and the mixture prior (dotted line) based on the design prior distribution (left) and design posterior distribution (right).

Figure 3. The number of demands and proportion of failures of 63 emergency generators in nuclear power plants.

Figure 4. The prior distribution (dashed line) and the posterior distribution (solid line) for $\pi$, the probability of success of an emergency generator on demand.

sample size are given in the right-hand side of Figure 2. We see, by incorporating the data into the analysis, that we are able to provide an RDT which gives assurances of up to almost 100%. To give assurance of 50% now only requires sample sizes of $n = (74, 141, 123)$ for the three analysis methods, all of which are much reduced. The corresponding maximum numbers of allowed failures would be $c = (0, 1, 1)$. Using the sample sizes from the prior calculations would give assurances of $(0.863, 0.882, 0.862)$, respectively.

The total computing time to determine the test plans based on the three analysis approaches for this example, including the inference on the historical data to obtain the design posterior distribution, is 3 min and 15 sec. The calculations were performed on a laptop running Windows 10 with an Intel(R) Core(TM) i5-8350U CPU @ 1.70GHz 1.90GHz processor with 16GB installed RAM and a 64-bit operating system. For a single test plan the computing time required, excluding the inference, is under 1 min.

We can compare the test plans using assurance found above with other methods from the literature. In Table 2, we compare assurance under the two analysis priors with a power calculation based on an exact binomial test, the classical risk criteria approach and the posterior risk criteria approach, which are all described in Section 2.1. For the power calculation, we chose a significance level $\alpha = 0.05$, and a power of 80% to detect a reliability of $\pi = 0.990$, which is the maximum likelihood estimate from the historical data. For the risk criteria-based
approaches we used acceptable and rejectable reliability levels of $\pi_0 = 0.99$ and $\pi_1 = 0.96$ and thresholds for the producer's and consumer's risks of 0.05. In all cases we choose the minimum acceptable sample size $n$ and find the corresponding cutoff $c$, and use these to calculate the assurance, classical producer's risk and classical consumer's risk. To calculate the assurance the design posterior is used. For the posterior risk criteria we base our calculations on both the design posterior and the design prior.

We see that the two frequentist approaches result in sample sizes that are larger than the other approaches, as they use neither the historical data nor the knowledge of the producer. The posterior risk criteria approach based on the design posterior gives a very small sample size of $n = 4$, which is due to the fact that a reliability of below 0.96 is given very small posterior probability. The assurance-based approaches keep the consumer's risk low at the expense of the producer's risk and the posterior probability. The assurance-based approaches keep the consumer's risk low at the expense of the producer's risk and the posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$. We see that each posterior distribution is relatively symmetrical on the log-scale. The majority of the density for the log reliable life is above the target value. We use the simulations from the design posterior distributions for $\beta$ and $\rho$ to find the assurance for any choice of sample size $n$. To do so, we also need to define the analysis prior. In this example, we consider a Bayesian analysis prior. The priors are chosen to give $\Pr_A(\tau > 0.5) = \log(6000)$.
with an Intel(R) Core(TM) i5-8350U CPU@1.70GHz 1.90GHz processor with 16GB installed RAM and a 64-bit operating system. Excluding the inference, the total computing time is 4 min.

The producer may have a maximum duration that they are able or willing to allow for the RDT or there may be additional costs for the producer associated with a longer duration. Therefore, they may wish to introduce a censoring time at which items which have not failed are removed from testing. Table 3 details the effect of different censoring times on the required sample size for this example. We see that censoring times greater than 16,000 hr would not increase the sample size. Decreasing the censoring time from this increases the sample size and censoring times less than 7000 hr would increase the sample size beyond the maximum the producer is willing to consider.

Testing half the vessels under each pressure may not lead to the optimal design. If we consider the pair of sample sizes \((n_{27}, n_{29})\) for the number of vessels to test at 27 and 29
Table 3. The effect of the censoring time $t_c$ (in hr) on the sample size required $n$ for the Weibull RDT.

| $t_c$  | $n$  | Assurance |
|-------|------|-----------|
| 16,000| 32   | 0.808     |
| 14,000| 34   | 0.805     |
| 12,000| 35   | 0.807     |
| 10,000| 37   | 0.806     |
| 8000  | 45   | 0.803     |
| 7000  | >60  | NA        |

Figure 8. The assurance based on the design posterior distribution for different combinations of $(n_{(27)}, n_{(29)})$.

Table 4. Each RDT design giving an assurance of greater than 80% sorted by total sample size.

| $n_{(27)}$ | $n_{(29)}$ | Total | Assurance |
|------------|------------|-------|-----------|
| 20         | 2          | 22    | 0.802     |
| 20         | 3          | 23    | 0.808     |
| 20         | 4          | 24    | 0.813     |
| 19         | 5          | 24    | 0.804     |
| 20         | 6          | 25    | 0.819     |
| 19         | 7          | 25    | 0.810     |
| 18         | 6          | 26    | 0.825     |
| 19         | 7          | 26    | 0.816     |
| 18         | 8          | 26    | 0.806     |

MPa, respectively, we can evaluate the assurance using the curve fitting technique in two-dimensions for all combinations of $n_{(27)}, n_{(29)} \in 1, \ldots, 20$, and the result, together with the simulated empirical proportions, is given as the surface in Figure 8.

We provide the combinations of $(n_{(27)}, n_{(29)})$ which give an assurance of at least 80% with the smallest combined sample size and their estimated assurance in Table 4. We see that we can achieve an assurance of greater than 80% by putting just 22 items on test, 20 at 27 MPa and 2 at 29 MPa. This is fewer than the 32 observations when using equal sample sizes between the two pressures. However, we may wish to put 24 or 25 items on test to increase the assurance further by 1.1% and 1.7%, respectively.

5. Summary and Further Work

In this article, we have considered the problem of sample size determination in RDTs for hardware products leading to failure on demand and time to failure data. The approach we have taken is to use the concept of assurance, which chooses the sample size to answer the question, “What is the probability that the RDT will lead to a successful outcome?” It can be used in conjunction with either frequentist or Bayesian analyses of the data following the test and separates the specification of prior beliefs in the design of the test from the prior to be used in the analysis following the test. Historical data can be incorporated into the sample size calculation without influencing the analysis of the RDT.

The methods have been fully developed in this article, including advice on how to specify both the design and analysis prior in each case. The code released with the article will allow practitioners who wish to use assurance to design their RDTs to do so. However, an important next step in this work will be to develop free open source software to make this process easier for practitioners.

The inference in the Weibull case has been performed using a combination of MCMC and a numerical scheme incorporating curve fitting. An adaptation worth investigation is to develop an augmented MCMC scheme to perform the inference. We used a Bayesian analysis with a skeptical analysis prior for the time to failure application. Other options would be to perform the analysis based on a mixture analysis prior as in the failure on demand application.

The assurance approach developed for the Weibull distribution is equally applicable to other lifetime distributions such as lognormal and gamma. This will form the basis of a future article. Our approach assumes that the stress levels are prespecified and we choose how many items to allocate to each. Another extension would be to also include the levels of the stress factor in the design problem, similar to the problems considered in the optimal design literature in accelerated life testing. In this case, a full decision-theoretic solution would be required.

The posterior risk criteria approach to RDT design is based on a single, shared prior distribution between the producer and the consumer. The approach could be adapted to incorporate separate priors for the design and analysis of the RDT.

Supplementary Materials

Appendices: Further details and explanation as follows. (pdf)

A Bayesian acceptance sampling in quality assurance
B Posterior risk criteria
C Proof of Proposition 1
D Choosing a prior distribution: binomial case
E Sensitivity to the skeptical analysis prior
F Parameter posterior distributions in the Weibull example

R code: This contains two R files:
1. Techno code.R—R code to run to reproduce the results in Example 4.1.
2. Techno.R—Two functions and the rjags model required for Example 4.1.
References

Collett, D. (1994), Modelling Survival Data in Medical Research, Boca Raton, FL: Chapman and Hall. [528]

Cook, A. R., Gibson, G. J., and Gilligan, C. A. (2008), "Optimal Observation Times in Experimental Epidemic Processes," Biometrics, 64, 860–868. [528]

Easterling, R. (1970), "On the Use of Prior Distributions in Acceptance Sampling," Annals of Reliability and Maintainability, 9, 31–35. [523,525]

Elsayed, E. (2012), "Overview of Reliability Testing," Journal of the American Statistical Association, 107, 402–408. [523]

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005), "Statistical Methods for Eliciting Probability Distributions," Journal of the American Statistical Association, 100, 680–701. [526]

Hamada, M. S., Wilson, A. G., Reese, C., and Martz, H. F. (2008), Bayesian Reliability, New York: Springer. [523,525,526,532]

Hamada, M. S., Wilson, A. G., Weaver, B. P., Griffiths, R. W., and Martz, H. F. (2014), "Bayesian Binomial Assurance Tests for System Reliability Using Component Data," Journal of Quality Technology, 46, 24–32. [523,525]

Jensen, W. (2015), "Binomial Reliability Demonstration Tests With Dependent Data," Quality Engineering, 27, 253–266. [523]

Lee, I., Hong, Y., Tseng, S., and Dasgupta, T. (2018), "Sequential Bayesian Design for Accelerated Life Tests," Technometrics, 60, 472–483. [524]

Lindley, D. V., and Singpurwala, N. D. (1993), "Adversarial Life Testing," Journal of the Royal Statistical Society, Series B, 55, 837–847. [523,524]

Lu, L., Li, M., and Anderson-Cook, C. M. (2016), "Multiple Objective Optimization in Reliability Demonstration Tests," Journal of Quality Technology, 48, 326–342. [525]

Martz, H. F., Kvaal, P. H., and Abramson, L. R. (1996), "Empirical Bayes' Estimation of the Reliability of Nuclear-Power-Plant Emergency Diesel Generators," Technometrics, 38, 11–24. [529]

Martz, H. F., and Waller, R. A. (1982), Bayesian Reliability Analysis, New York: Wiley. [523]

McKane, S., Escobar, L., and Meeker, W. (2005), "Sample Size and Number of Failure Requirements for Demonstration Tests With Log-Locatioin-Scale Distributions and Failure Censoring," Technometrics, 47, 182–190. [523]

Mease, D., and Nair, V. (2006), "Extreme (X-) Testing With Binary Data and Applications to Reliability Demonstration," Technometrics, 48, 399–410. [523]

Meeker, W., and Escobar, L. (2004), "Reliability: The Other Dimension of Quality," Quality Technology and Quantitative Management, 1, 1–25. [523,525]

Miller, F., Zohar, S., Stallard, N., Madan, J., Posch, M., Hee, S., Pearce, M., Vägerö, M., and Day, S. (2018), "Approaches to Sample Size Calculation for Clinical Trials in Rare Diseases," Pharmaceutical Statistics, 17, 214–230. [524]

Muirhead, R., and Şoaita, A. (2013), "On an Approach to Bayesian Sample Sizing in Clinical Trials," in Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton, eds. G. Jones and X. Shen, Beachwood, OH: Institute of Mathematical Statistics, pp. 126–137. [524]

Müller, P. (1999), "Simulation-Based Optimal Design," in Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting (Vol. 6), eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 459–474. [528]

Müller, P., and Palmer, J. (1996), "Optimal Design via Curve Fitting of Monte Carlo Experiments," Journal of the American Statistical Association, 90, 1322–1330. [528]

O'Hagan, A., and Stevens, J. (2001), "Bayesian Assessment of Sample Size for Clinical Trials of Cost-Effectiveness," Medical Decision Making, 21, 219–230. [524]

O'Hagan, A., Stevens, J., and Campbell, M. (2005), "Assurance in Clinical Trial Design," Pharmaceutical Statistics, 4, 187–201. [524,525]

Plummer, M. (2016), "rjags: Bayesian Graphical Models Using MCMC," R Package Version 4-6. [530]

Ren, S., and Oakley, J. (2014), "Assurance Calculations for Planning Clinical Trials With Time-to-Event Outcomes," Statistics in Medicine, 33, 31–45. [524,525]

Rios Insua, D., Ruggeri, F., Soyer, R., and Rasines, D. G. (2018), "Adversarial Issues in Reliability," European Journal of Operational Research, 266, 1113–1119. [523,524]

Rios Insua, D., Ruggeri, F., Soyer, R., and Wilson, S. (2020), "Advances in Bayesian Decision Making in Reliability," European Journal of Operational Research, 282, 1–18. [523]

Rufo, M., Martin, J., and Perez, C. (2014), "Adversarial Life Testing: A Bayesian Negotiation Model," Reliability Engineering & System Safety, 131, 118–125. [523]

Spiegelhalter, D., Abrams, K., and Myles, J. (2004), Bayesian Approaches to Clinical Trials and Health-Care Evaluations, New York: Wiley. [527]

Spiegelhalter, D., Freedman, L., and Parmar, M. (1994), "Bayesian Approaches to Randomized Trials," Journal of the Royal Statistical Society, Series A, 157, 357–416. [524,525]

Tobias, P., and Trindade, D. (1995), Applied Reliability, Boca Raton, FL: Chapman and Hall. [525]

Valley, R., Smith, C., Gale, J., and Woodward, P. (2015), "Advantages of a Wholly Bayesian Approach to Assessing Efficacy in Early Drug Development: A Case Study," Pharmaceutical Statistics, 14, 205–215. [524]

Winkler, R. (1967), "The Assessment of Prior Distributions in Bayesian Analysis," Journal of the American Statistical Association, 62, 776–880. [526]