Assurance for sample size determination in binomial reliability demonstration testing

K.J. Wilson, M. Farrow

School of Mathematics, Statistics & Physics, Newcastle University, UK

Abstract Frequently, manufacturers are required to demonstrate that products meet reliability targets. A typical way of doing this is to use reliability demonstration tests (RDTs), in which a number of products are put on test and the test is passed if a critical number or fewer fail. There are various methods for determining the sample size for such tests. Traditionally this was based on the size of a hypothesis test following the RDT. More recently, Bayesian approaches have been proposed based on the idea of risk criteria. However, these approaches do not lead to a single, optimal sample size and conflate the choice of sample size for the test and the analysis to be undertaken once the test has been conducted. In this paper we offer an alternative approach to sample size determination based on the idea of assurance. This approach can overcome each of these issues with risk criteria. Assurance chooses the sample size to answer the question: "What is the probability that the RDT will result in a successful outcome?" We demonstrate the use of assurance for sample size calculations in binomial RDTs and discuss the specification of prior distributions for the design and analysis of the test.

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 such 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 is observed. If the number of failures is below some pre-defined threshold then the test is said to be passed and the reliability of the product is demonstrated (Elsayed, 2012).

The design of an RDT is therefore given by the number of items put on test and the number of failures which are allowed for the test to be passed. An additional consideration which affects the test plan is the type of hardware product which is of interest. In this paper we consider products which are required to function on demand and can either work or fail, such as a parachute.

Calculating sample sizes for RDTs has been considered at least as far back as the 1960s. A traditional approach was to choose the sample size to give a specified size of a hypothesis test (see, for example, Section 3 of Meeker and Escobar (2004)). Another widely investigated approach is based on the idea of risk criteria, which evaluate the risk to the producer of the product and consumer of the product associated with incorrect conclusions from the RDT (Easterling, 1970; Marz and Waller, 1982; Hamada *et al*, 2008).

Assurance (also known as expected power and average power), has been proposed (Spiegelhalter *et al*, 1994; O'Hagan and Stephens, 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. There is no limitation with assurance that the analysis following the trial be a frequentist test, however.

In this paper we consider the use of assurance for sample size determination for RDTs. We develop an approach based on assurance to select both the number of items to test and the threshold number of failures 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. Following this introduction, we detail our approach to sample size determination using assurance for failure on demand data in Section 2. The following section illustrates the approach developed in the context of an example. The paper concludes with Section 4, which provides a summary of the approach and some areas for futher work.

2. Binomial reliability demonstration testing

2.1. *Traditional approaches*

Suppose that we have n devices which we are to put on test and that Y is the number which will fail the test. Also define π 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, a reliability demonstration test plan (Hamada *et al*, 2008) is given by the pair (n,c). The likelihood associated with the test is typically assumed to be binomial

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

Traditionally, a reliability demonstration test (RDT) would be analysed using a hypothesis test where $H_0 : \pi = \pi_T$, the quantity π_T is the target reliability and $H_1 : \pi > \pi_T$. For a $100(1-\alpha)\%$ critical level, we would reject H_0 when $\Pr(Y \leq c) \leq \alpha$. That is if $\sum_{y=0}^{c} \binom{n}{y} (1-\pi_T)^y \pi_T^{n-y} \leq \alpha$. If we set $c = 0$ then we could choose n to satisfy

$$n \geq \frac{\log(\alpha)}{\log(\pi_T)}. \tag{2}$$

Other approaches have been suggested, often based on the idea of risk criteria. In these approaches, test plans are chosen to keep the probabilities of making the wrong conclusions from the reliability demonstration test small. Classical risk criteria (Tobias and Trindade, 1995), average risk criteria (Easterling, 1970) and (Bayesian) posterior risk criteria (see, for example, Chapter 10 of Hamada *et al* (2008)) have been proposed.

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 $\boldsymbol{x}$ in the analysis to reduce n. They named the new tests which incorporated past observations reliability assurance tests (RAT).

2.2. *Assurance*

If we consider the posterior risk criteria, there is no single optimal (n,c) which results. The approach also 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.

Assurance has been proposed (Spiegelhalter *et al*, 1994; O'Hagan *et al*, 2005; Ren and Oakley, 2013) as the correct Bayesian approach to sample size calculations in medical statistics. In the context of reliability demonstration testing it can overcome each of the potential drawbacks detailed above.

Assurance chooses a sample size based on the answer to the question "what is the probability that the reliability demonstration test is going to result in a successful outcome?".

The probability of a successful test is given by

$$\Pr[\text{Successful test}] = \int_0^1 f(y \leq c \mid \pi) p(\pi) d\pi \tag{3}$$

$$= \int_0^1 \sum_{y=0}^{c} \binom{n}{y} (1-\pi)^y \pi^{n-y} p(\pi) d\pi, \tag{4}$$

where $p(\pi)$ is the prior distribution for π and $f(y \leq c \mid \pi)$ is the cumulative distribution function of y evaluated at c, which in this case is binomial. We would like this probability to be high. We may wish to choose a minimum acceptable value γ.

Let us suppose that we have data from previous tests of the form x_i for $i = 1, \ldots, I$ where $x_i \mid \pi_i \sim \text{bin}(n_i, 1-\pi_i)$, 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 various different locations. Then, let us suppose that our proba-

bilitiy of interest also comes from the same prior distribution $\pi \sim \text{beta}(a,b)$ and that we define a hyper-prior distribution over (a,b).

The probability of a successful test is now

$$\Pr[\text{Successful test} \mid \boldsymbol{x}] = \int_0^1 f(y \leq c \mid \pi) p(\pi \mid \boldsymbol{x}) d\pi \tag{5}$$

$$= \int_0^1 \sum_{y=0}^{c} \binom{n}{y} (1-\pi)^y \pi^{n-y} p(\pi \mid \boldsymbol{x}) d\pi. \tag{6}$$

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

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

(b)For $j = 1,\ldots,N$ draw $\pi^{(1)},\ldots,\pi^{(N)}$ as

$$\pi^{(j)} \sim \text{beta}(a^{(j)},b^{(j)}). \tag{7}$$

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

$$\Pr[\text{Successful test} \mid \boldsymbol{x}] \approx \frac{1}{N} \sum_{j=1}^{N} \left[\sum_{y=0}^{c} \binom{n}{y} (1-\pi^{(j)})^y (\pi^{(j)})^{n-y} \right], \tag{8}$$

We can use assurance in this way to choose the sample size n for a given value of c. The critical number of failures c is chosen based on the analysis to be carried out following the test. This flexible approach allows the use of 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 analyse the test data, the prior for the analysis can be different to that used here in the sample size determination.

2.3. *Cut-off choice*

2.3.1. *Frequentist approaches*

Consider the binomial test in Section 2.1. We would reject H_0 if $\sum_{y=0}^{c} \binom{n}{y} (1-\pi_T)^y (\pi_T)^{n-y} \leq 0.05$. Therefore we simply select the largest value of c for which this is true.

An alternative to the exact binomial test is to use a Normal distribution to approximate the binomial distribution. In this case the test statistic is $Z = [c/n - \pi_T]/[\sqrt{\pi_T \times (1-\pi_T)/n}]$ and we would choose c to be the largest value for which $Z < Z_{0.05}$, where $Z_{0.05}$ is the 5% critical value of the standard Normal distribution.

2.3.2. *Bayesian approaches*

Suppose that, in the analysis of the test result, we have a prior distribution for π of $p_A(\pi)$. Note that this may be different from $p(\pi)$. Then we may choose a rule which says that if the posterior probability that $\pi \leq \pi_T$ is small, then the test is passed. That is, the test is passed if $\Pr(\pi \leq \pi_T \mid Y = y) \leq 0.05$, for example.

In this case we would choose c to be the largest value for which $\Pr(\pi \leq \pi_T \mid Y = c) \leq 0.05$. This posterior probability can be calculated as

$$\Pr(\pi \leq \pi_T \mid Y = c) = \frac{\int_0^{\pi_T} f(c \mid \pi) p_A(\pi) d\pi}{\int_0^1 f(c \mid \pi) p_A(\pi) d\pi}, \tag{9}$$

$$= \frac{\int_0^{\pi_T} \binom{n}{c} (1-\pi)^c \pi^{n-c} p_A(\pi) d\pi}{\int_0^1 \binom{n}{c} (1-\pi)^c \pi^{n-c} p_A(\pi) d\pi}, \tag{10}$$

where $f(c \mid \pi)$ is the probability mass function of Y evaluated at c.

This approach produces a reliability demonstration test plan, as the data have been used to calculate the

sample size n only and not to analyse the results of the test. A reliability assurance test plan can also be produced in this way by utilising the data $\boldsymbol{X} = \boldsymbol{x}$ here in the choice of c. This can be achieved using MCMC in the same way as when calculating n from known c in the previous section.

2.4. *Choice of priors*

In order to use assurance to decide on the optimal test plan (n,c) we need a design prior distribution for π, $p(\pi)$, to be used in the assurance calculation and, if a Bayesian analysis is to be undertaken following the test, a second prior distribution on π, $p_A(\pi)$, to be used in the analysis.

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. The beta distribution parameters (a,b) are not suitable. However, we can reparameterise the beta distribution so that $\pi \sim \text{beta}(mp, m(1-p))$, where $p = a/(a+b)$ is the mean and $m = a+b$ is the size of the "prior sample" which the mean is based on. We can then define suitable hyper-prior distributions on (p,m), such as

$$p \sim \text{beta}(a_p, b_p), \tag{11}$$

$$m \sim \text{gamma}(a_m, b_m). \tag{12}$$

In the analysis following the test, the prior distribution of the producer may be deemed a controversial choice. There could be several different groups of people who will be affected by the decisions made following the success or failure of the test. Therefore, it may be more reasonable in the analysis of the test data to use a prior distribution which is relatively conservative. We call such a prior a sceptical prior.

The sceptical prior should be designed so that there is only a small probability that $\pi > \pi_T$. So, for example, if a beta(α,β) distribution is assumed for $p_A(\pi)$, then we might choose (α,β) to satisfy $1 - \dfrac{B(\pi_T;\alpha,\beta)}{B(\alpha,\beta)} = \delta$, where $\delta = 0.05$.

3. Example

We consider an example from Marz *et al* (1996). We would like to demonstrate the reliability of an emergency generator in a nuclear power plant. We will compare assurance based on two analysis methods: the exact binomial test and the Bayesian approach using a sceptical analysis prior. In the binomial test we reject H_0 if $p < 0.05$ and in the Bayesian approach we conclude that the test is passed if $\Pr(\pi \leq \pi_T \mid Y = y) \leq 0.05$.

We choose a beta distribution for the sceptical analysis prior with $\alpha = 13.6$ and $\beta = 2$. This gives $\Pr(\pi > \pi_T) = 0.05$. For the design prior $p(\pi)$ we choose to give (m,p) gamma and beta priors respectively with parameter values chosen to be $(78,2)$ and $(200,1)$. This equates to a prior mean with an expected value of 0.975 based on a prior sample with an expected value of 200. We consider a target reliability of 0.975.

Based on these specifications, we can plot the sample size n against the assurance for both of the approaches. This is given in Figure 1.

The binomial test case is the solid line and the sceptical prior case is the dashed line. We see that the binomial test gives the highest assurance for any particular sample size and the sceptical prior the lowest assurance in this case. The lines are decreasing as c remains constant but n increases, and then jump each time c increases. To achieve an assurance of 35% in this case would require $n = (418, 567)$ under the respective methods. We cannot achieve an assurance of greater than about 50% in this case.

However, we have data on the behaviour on demand of generators of the same type at other plants.

The data represent tests of 63 generators in nuclear power plants in the USA. 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 porportion of failures in the tests are given in Figure 2.

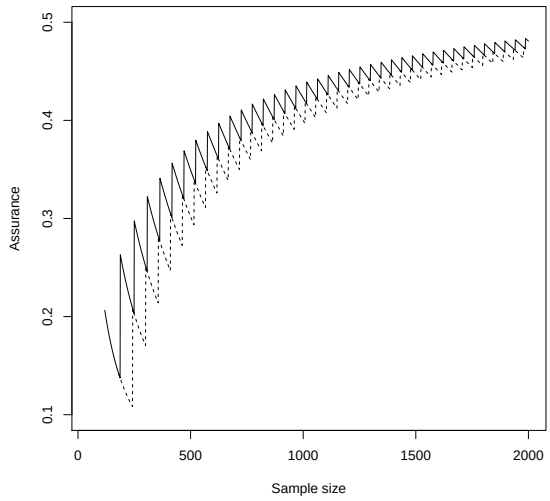

Figure1. The sample size *n* against the assurance for the binomial test (solid line) and the sceptical prior (dashed line).

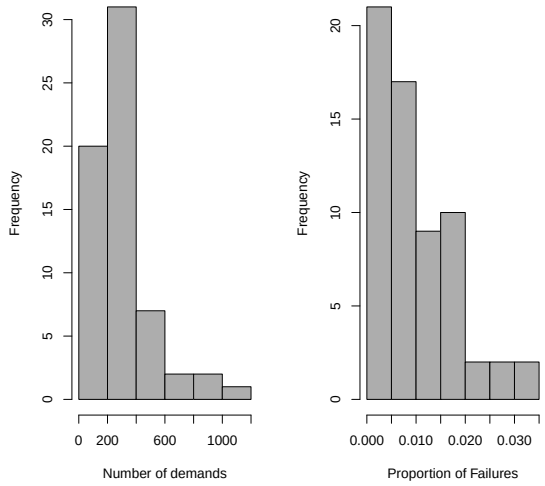

Figure2. The number of demands and proportion of failures of 63 emergency generators in nuclear power plants.

We see that we have some fairly large test data sets and that the failure proportions recorded in the test are very low, typically below 3%.

Using the same assurance prior as above, we generate 10,000 samples of (p,m) from the posterior predictive distribution in rjags , having discarded 1000 samples as burn in.

We are able to calculate the assurance for both of the methods as above using this posterior distribution for π. The results for different values of the sample size are given in Figure 3.

We see, by incorporating the data into the analysis, that we are able to provide a reliability assurance test which gives assurances of up to approximately 90%. To ensure assurance of 35% now only requires sample sizes of $n = (119, 243)$ for the analysis methods, both of which are much reduced. Using the sample sizes from the prior calculations would give assurances of $(0.715, 0.725)$ respectively.

4. Summary

In this paper we have considered the problem of sample size determination in reliability demonstration tests for hardware products leading to failure on demand data. The approach we have taken is to use the Bayesian concept of assurance, which chooses the sample size to answer the question “what is the probability that the RDT

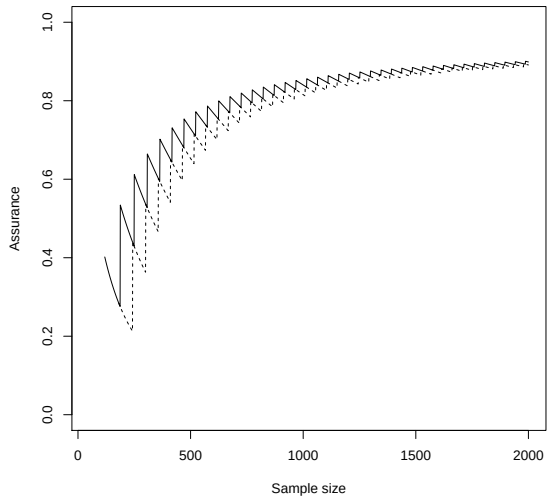

Figure3. The sample size *n* against the assurance for the binomial test (solid line) and the sceptical prior (dashed line).

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, unlike typical risk criteria based approaches, 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.

References

Marz HF, Kvam PH, Abramson LR (1996) Empirical Bayes' estimation of the reliability of nuclear-power-plant emergency diesel generators, *Technometrics*, 38, 11-24.

Hamada MS, Wilson AG, Reese CS, Martz HF (2008) *Bayesian Reliability*, Springer.

Meeker WQ, Escobar LA (2004) Reliability: the other dimension of quality, *Quality technology and quantitative management*, 1, 1-25.

Tobias PA, Trindade DC (1995) *Applied Reliability*, Van Nostrand Reinhold.

Easterling RJ (1970) On the use of prior distributions in acceptance sampling, *Annals of reliability and maintainability*, 9, 31-35.

Spiegelhalter DJ, Freedman LS, Parmar MKB (1994) Bayesian approaches to randomized trials, *Journal of the Royal Statistical Society: Series A*, 157, 357-416.

O'Hagan A, Stephens JW, Campbell MJ (2005) Assurance in clinical trial design, *Pharmaceutical Statistics*, 4, 187-201.

Ren S, Oakley JE (2013) Assurance calculations for planning clinical trials with time-to-event outcomes, *Statistics in Medicine*, 33, 31-45.

Plummer M (2016) rjags: Bayesian Graphical Models using MCMC, *R package version 4-6*, `https://CRAN.R-project.org/package=rjags`

O'Hagan A, Stevens JW (2001) Bayesian assessment of sample size for clinical trials of cost-effectiveness, *Medical Decision Making*, 21, 219-230.

Elsayed EA (2012) Overview of reliability testing, *IEEE Transactions on Reliability*, 61, 282–291.

Martz HF, Waller RA (1982) *Bayesian Reliability Analysis*, Wiley, New York.