Simulation Analysis of a Bayesian Test Plan for Sequential Data from a Homogeneous Poisson Process

by

Robert Christian Foster

A creative component submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Statistics

Program of Study Committee:
Alyson G. Wilson, Major Professor
Ulrike Genschel
William Meeker

Iowa State University

Ames, Iowa

2010

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Section 1. Introduction and Model Specification

1.1 Data

It is common in reliability analysis to wish to compute failure rates for various components with a goal of evaluating future components. However, suppose data exists in the following context: rather than failure times for individual units, measured instead are the number of failures and total running times for a system of units, in which each unit in each system was either replaced or repaired upon failure. As an example of such data, consider the number of system failures from the Farley 1 U.S. nuclear power plant, as shown in Table 1.1.

System	s_i (failures)	t_i (thousand hours)	$\hat{\lambda}$ (MLE)
1	5	94.32	5.3×10^{-2}
2	1	15.720	6.4×10^{-2}
3	5	62.880	8.0×10^{-2}
4	14	125.760	11.1×10^{-2}
5	3	5.240	57.3×10^{-2}
6	19	31.440	60.4×10^{-2}
7	1	1.048	95.4×10^{-2}
8	1	1.048	95.4×10^{-2}
9	4	2.096	191.0×10^{-2}
10	22	40.480	209.9×10^{-2}

Table 1.1 Pump failure count data from Farley 1 U.S. nuclear power plant, from Gaver and O'Muircheartaigh (1987).

The goal of this analysis is to develop a test plan for a new type of unit that is believed to be similar to the units from the original system. The purpose of this paper is to not only show how this may be done, but also to take the analysis one step further and analyze through simulation the resulting test plan itself, with specific inquiry into prior sensitivity analysis, choice of parameters of the test plan, and model misspecification.

1.2 Literature Review

The ideas in this paper have been investigated in the past, but no in-depth analysis of the accuracy has been attempted. Guthrie and Johns (1959) and Hald (1960) introduced the Bayesian acceptance sampling procedures used in this paper. Hamada et. al. (2008) offers an in-depth overview of applied methods in Bayesian reliability. In particular, Chapter 10, and especially Section 3, describes in more detail many of the methods used in this paper. Sun and Berger (1994) contains much of the statistical theory behind the methods used in this paper and other issues in reliability, with a focus on a Weibull distribution for times between failures. However, many of the conclusions reached still apply to homogeneous Poisson processes. Gelfand and Smith (1990) describes the now standard Markov Chain monte methods used to fit the Bayesian models in this paper. Brush (1986) gives a more thorough explanation of the difference between classical and Bayesian producer's risks, with numerical examples to illustrate. Gaver and O'Muircheartaigh (1987) contains the data cited in the introduction paper. Furthermore, it shows a different, type of analysis that could be used for data from a homogeneous Poisson process. Berger (1984) contains an overview of Bayesian methods for prior sensitivity analysis. Both application and theory are explored, and many examples are given. Guida, Calabra, and Pulcini (1989) extends the homogenous poisson process to a nonhomogeneous Poisson process and performs some simulation analysis in order to compare Bayesian and maximum likelihood fitting techniques.

1.3 Hierarchical Model

Remember that this analysis does not concern individual units, but rather systems of units that were replaced upon failure. Let i index each system. At all times assume that the data comes from $m > 1$ systems, each with individual failure rate λ_i. Assume each system follows a homogenous Poisson process. For a homogenous Poisson process, this is

$$s_i \sim \text{Poisson}(\lambda_i t_i)$$

where the s_i are assumed independent given the λ_i.

Since it is assumed that the failure rates for each system come from the same source, it is natural to model the λ_i hierarchically. For this paper, model the failure rates as

$$\lambda_i \sim \text{Gamma}(\alpha, \beta)$$

This allows the sharing of information between systems in order to better model the common source, which will prove important in the estimation of the parameters of the test plan.

1.4 Test Plan and Risks

Testing will proceed by observing a new system. Suppose that the number of individual units in the system and the amount of time needed for testing are not limiting factors. The test will be conducted as follows: define c as the maximum allowed number of failures and T as the desired operating time. Begin by operating the first unit until it fails. It will then be replaced by the second unit, which is operated until it fails, and so on. If the total operating time $t_1 + t_2 + ...$ reaches T before c failures are observed, then the test is passed. Otherwise, it is considered a failure.

There must be criteria for choosing T and c, which will be based on two separate levels of the failure rate λ: the acceptable failure rate λ_A and the rejectable failure rate λ_R, both of which are chosen a priori, and which represent lower (λ_R) and upper (λ_A) boundaries on the failure rate.

The necessary links between the testing parameters T and c and the acceptable and rejectable failure rates are the posterior producer's and consumer's risk. Define the posterior producer's risk as the probability that the true failure rate is actually less than the rejectable failure rate given that the test was failed, or

$$\mathbf{P}(\lambda \leq \lambda_R | \textit{Test is Failed}, \mathbf{s,t}) = \int_0^{\lambda_R} p(\lambda | s > c, \mathbf{s,\ t}) d\lambda$$

$$= \int_0^{\lambda_R} \frac{f(s > c|\lambda) p(\lambda|\mathbf{s,t})}{\int_0^{\infty} f(s > c|\lambda) p(\lambda|\mathbf{s,t}) d\lambda} d\lambda$$

$$= \frac{\int_0^{\lambda_R} \left[1 - \sum_{s=0}^{c} \frac{(\lambda T)^s \exp(-\lambda T)}{s!}\right] p(\lambda|\mathbf{s,t}) d\lambda}{\int_0^{\infty} \left[1 - \sum_{s=0}^{c} \frac{(\lambda T)^s \exp(-\lambda T)}{s!}\right] p(\lambda|\mathbf{s,t}) d\lambda}$$

where λ is the failure rate, s is the total number of failures, t is the sum amount of time between failures, and $p(\lambda|\mathbf{s,t})$ is the posterior distribution for λ. Similarly, define the posterior consumer's risk as the probability that the true failure rate is actually greater than the acceptable failure rate given that the test was passed, or

$$\mathbf{P}(\lambda \geq \lambda_A | \textit{Test is Passed}, \mathbf{s,t}) = \int_{\lambda_A}^{\infty} p(\lambda | s \leq c, \mathbf{s,\ t}) d\lambda$$

$$= \int_{\lambda_A}^{\infty} \frac{f(s \leq c|\lambda) p(\lambda|\mathbf{s,t})}{\int_0^{\infty} f(s \leq c|\lambda) p(\lambda|\mathbf{s,t}) d\lambda} d\lambda$$

$$= \frac{\int_{\lambda_A}^{\infty} \left[\sum_{s=0}^{c} \frac{(\lambda T)^s \exp(-\lambda T)}{s!}\right] p(\lambda|\mathbf{s,t}) d\lambda}{\int_0^{\infty} \left[\sum_{s=0}^{c} \frac{(\lambda T)^s \exp(-\lambda T)}{s!}\right] p(\lambda|\mathbf{s,t}) d\lambda}$$

Using these, T and c can be solved for, which allows not only for the original purpose of testing, but for the evaluation of the test plan itself.

Section 2. Estimation

2.1 Bayesian Parameter Estimation

This model will be fit using Bayesian techniques. Recall the hierarchical model

$$S_i \sim \text{Poisson}(\lambda_i t_i)$$

$$\lambda_i \sim \text{Gamma}(\alpha, \beta)$$

For the likelihood function, assuming observations $s_1, s_2, ..., s_m$ are conditionally independent, the following is obtained:

$$\begin{aligned} f(s|\lambda_i, t_i) &= \prod_{i=1}^{m} \frac{(\lambda_i t_i)^{s_i}}{s_i!} e^{-\lambda_i t_i} \\ &= (\lambda_i t_i)^{\sum_{i=1}^{m} s_i} e^{-(10\lambda_i t_i)} \prod_{i=1}^{m} \frac{1}{s_i!} \end{aligned}$$

For priors on α and β, set $\alpha \sim \text{InvGamma}(a_1, b_1)$ and $\beta \sim \text{InvGamma}(a_2, b_2)$. Then the joint posterior follows:

$$\begin{aligned} \pi(\lambda_1, \ldots, \lambda_m, \alpha, \beta|s, t) &\propto f(s|\lambda_1, \ldots, \lambda_m, \alpha, \beta)\pi(\lambda_1, \ldots, \lambda_m, \alpha, \beta) \\ &\propto \frac{\beta^{m\alpha}}{\Gamma^m(\alpha)} \left(\prod_{i=1}^{m} \exp\left(-\lambda_i(\beta + t_i)\right) \lambda_i^{s_i+\alpha-1} \right) \times \\ &\quad \left(\frac{1}{\alpha}\right)^{a_1+1} \exp\left(\frac{-b_2}{\alpha}\right) \left(\frac{1}{\beta}\right)^{a_2+1} \exp\left(\frac{-b_2}{\beta}\right) \end{aligned}$$

This unnormalized joint posterior distribution does not have a form that is recognizable as any common distribution. Since numerical methods are feasible, standard MCMC techniques can be used. The full conditionals are given below:

$$\begin{aligned}
\lambda_i|\alpha,\beta &\sim Gamma(\alpha+s_i,\beta+t_i) \\
\alpha,\beta|\lambda_1,\ldots,\lambda_m &\propto \frac{\beta^{m\alpha}}{\Gamma^m(\alpha)}\left(\prod_{i=1}^{m}\exp\left(-\lambda_i(\beta)\right)\lambda_i^{\alpha}\right)\times \\
&\quad \left(\frac{1}{\alpha}\right)^{a_1+1}\exp\left(\frac{-b_2}{\alpha}\right)\left(\frac{1}{\beta}\right)^{a_2+1}\exp\left(\frac{-b_2}{\beta}\right)
\end{aligned}$$

2.2 Estimation of Producer's and Consumer's Risks

Assuming a sample $\lambda^{(j)}$, where $j = 1...N$ indexes the sample, is available from the posterior predictive distribution for λ, the producer's risk can then be estimated as:

$$\mathbf{P}(\lambda \leq \lambda_R|\textit{Test is Failed},\mathbf{s},\mathbf{t}) \approx \frac{\sum_{j=1}^{N}\left[1-\sum_{s=0}^{c}\frac{(\lambda^{(j)}T)^s\exp\left(-\lambda^{(j)}T\right)}{s!}\right]I(\lambda^{(j)}\leq\lambda_R)}{\sum_{j=1}^{N}\left[1-\sum_{s=0}^{c}\frac{(\lambda^{(j)}T)^s\exp\left(-\lambda^{(j)}T\right)}{s!}\right]},$$

and the posterior consumer's risk as

$$\mathbf{P}(\lambda \geq \lambda_A|\textit{Test is Passed},\mathbf{s},\mathbf{t}) \approx \frac{\sum_{j=1}^{N}\left[\sum_{s=0}^{c}\frac{(\lambda^{(j)}T)^s\exp\left(-\lambda^{(j)}T\right)}{s!}\right]I(\lambda^{(j)}\geq\lambda_A)}{\sum_{j=1}^{N}\left[\sum_{s=0}^{c}\frac{(\lambda^{(j)}T)^s\exp\left(-\lambda^{(j)}T\right)}{s!}\right]},$$

as shown in Hamada et. al. (2008).

Then by defining maximum allowed probabilities γ and δ for the posterior producer's risk and the posterior consumer's risk, respectively, the following system of nonlinear equations is obtained.

$$\mathbf{P}(\lambda \leq \lambda_R|\textit{Test is Failed},\mathbf{s},\mathbf{t}) \leq \gamma$$

$$\mathbf{P}(\lambda \geq \lambda_A|\textit{Test is Passed},\mathbf{s},\mathbf{t}) \leq \delta$$

These can be solved for T and c by using the following algorithm:

1. Set c = 0.

2. Use numerical methods to calculate the value T such that posterior producer's risk, $\mathbf{P}(\lambda \leq \lambda_R | \textit{Test is Failed}, \mathbf{s,t})$, is equal to γ

3. If the posterior consumer's risk, $\mathbf{P}(\lambda \geq \lambda_A | \textit{Test is Passed}, \mathbf{s,t})$, is less than δ, end. Otherwise, set c = c + 1 and return to step 2.

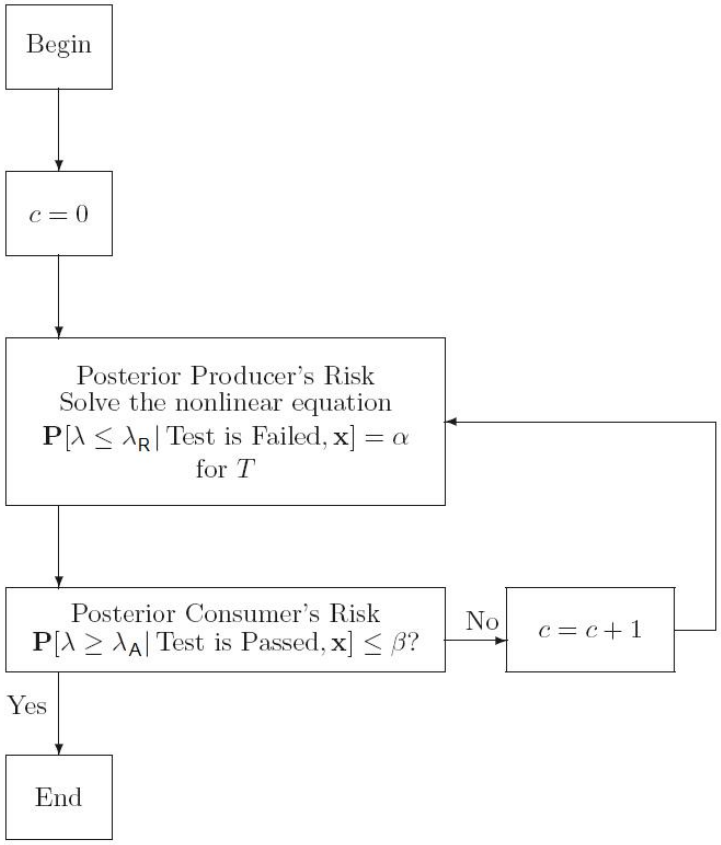

Figure 2.1 Graphical Illustration of Test Plan, from Hamada et. al. (2008).

Section 3. Evaluation of the test plan

3.1 Comparison Between Test Plans

In this paper it will often occur that there are two test plans that are based on models that are similar in terms of underlying distributions and methods that are used to estimate them, but not quite equal. Repeated simulation was used consistently across all test plans to acquire the posterior predictive distributions of variables of interest in order to compare the two test plans. In particular, comparisons will be made between test plans based on models with strong informative priors, models with incorrect data generating mechanisms, and between test plans using the same models, but different acceptable and rejectable error rates.

3.2 Simulation of Data

The following steps can be taken to simulate data from the model:

1. Fix the parameters of the hierarchical gamma distribution α and β.
2. Choose n as the number of systems to observed, so $i = 1, ..., n$. Simulate $\lambda_1, ..., \lambda_n$ from Gamma(α, β).

At this point, it must be decided how to obtain the number of failures, s_i, and the total operating time, t_i, as given one, it is possible to simulate the other. It was decided for this paper to fix s_i as constant for each system, as it would seem potentially reasonable that a real test would be conducted by running each system until a predetermined number of failures is observed, and because quality (or lack of quality, if so desired) of estimation for each λ_i can be ensured by increasing or decreasing s_i.

3. Fix s as the number of failures to be observed in each system, so $s_i = s \; \forall \; i$.

Recall that for a homogenous Poisson process with fixed time t_i, times between failures are modeled as

$$s_i \sim \text{Poisson}(\lambda_i t_i)$$

It can also be shown that for a homogenous Poisson process with a fixed number of failures s, times between individual failures can be modeled with an exponential distribution, and these can be summed to generate the total amount of time between failures. Specifically, let $j = 1, ..., s$ index the period between failure $j-1$ and failure j, denoting $j = 0$ as the starting time

4. For each λ_i, simulate s_i times between failures from the distribution

$$f(t_{i_j}) = \lambda_i \exp\{-\lambda_i t_{i_j}\}$$

   Then the total operating time for the system until s_i failures is $t_i = \sum_1^s t_{i_j}$.

5. Obtain estimates $\hat{\lambda_1}, ..., \hat{\lambda_n}, \hat{\alpha}, \hat{\beta}$ by MCMC methods and obtain a sample from the posterior predictive distribution for λ_i.

6. Set the acceptable producer's and consumer's risks and choose λ_A and λ_R to obtain a test plan T and c by the iterative algorithm described in Figure 2.1.

3.3 Measures of Accuracy

Once a test plan is generated, it can be evaluated. Since α and β were fixed in step 1 of the simulation procedure, the true underlying gamma distribution of the λ_i is known, hence simulation of new values from the same distribution of the data is possible. Repeat the following:

1. Simulate λ_k from $Gamma(\alpha, \beta)$

2. Simulate $t_{k_1}, ..., t_{k_{c+1}}$ values from $f(t_{i_j}) = \lambda_k \exp\{-\lambda_k t_{i_j}\}$ and calculate $\sum t_{k_j} = t_k$

3. If $t_k < T$, then conclude that the operating time did not exceed T before the number of failures exceeded c, so consider the system to have failed. If $t_k > T$, then consider the system to have passed.

These steps are repeated in order to generate a sufficient number of passes and failures for each test plan.

At step 3, it must be determined whether or not the simulated system was "classified" correctly. Consider the producer's and consumer's risks, $\mathbf{P}(\lambda \leq \lambda_R | \textit{Test is Failed}, \mathbf{s,t})$ and $\mathbf{P}(\lambda \geq \lambda_A | \textit{Test is Passed}, \mathbf{s,t})$, respectively. In the case of the producer's risk, it is considered "correct" to pass a system when the failure rate is less than or equal to λ_R. Extending on this idea, it should be "correct" to fail a system when the failure rate is *greater* than λ_R. This idea can be made into tabular form, mirroring the classic Type I and Type II error table.

	Test Passed	Test Failed
$\lambda_k \leq \lambda_R$	Correct	Incorrect
$\lambda_k > \lambda_R$	Incorrect	Correct

Table 3.1 Classification for Producer's Risk

Since *both* types of errors are potentially of interest, define the secondary producer's risk as $\mathbf{P}(\lambda > \lambda_R | \textit{Test is Passed}, \mathbf{s,t})$. Furthermore, define the producer's classification rate as $\mathbf{P}((\lambda \leq \lambda_R \cap \textit{ Test is Passed}) \cup (\lambda > \lambda_R \cap \textit{ Test is Failed}))$ so that a classification rate of 1 is ideal.

A similar table can be constructed for the consumer's risk.

	Test Passed	Test Failed
$\lambda_k < \lambda_A$	Correct	Incorrect
$\lambda_k \geq \lambda_A$	Incorrect	Correct

Table 3.2 Classification for Consumer's Risk

Define the secondary consumer's risk as $\mathbf{P}(\lambda < \lambda_A | \textit{Test is Failed}, \mathbf{s,t})$ and the consumer's classification rate as $\mathbf{P}((\lambda < \lambda_A \cap \textit{ Test is Passed}) \cup (\lambda \geq \lambda_R \cap \textit{ Test is Failed}))$.

Lastly, define the combined classification rate as the intersection of the producer's and consumer's classification rate – this is simply the probability of a "correct" in either table.

Note that the definitions of the secondary producer's and consumer's risks make intuitive sense, as replacing λ_A with λ_R in the secondary consumer's risk gives the producer's risk, and replacing λ_R with λ_A in the secondary producer's risk gives the consumer's risk.

In all investigations of the test plan, the producer's, consumer's, and combined classification rates will be central to the evaluation of the performance test plan. This is because while estimation of the parameters of the test plan are certainly of interest, estimation of these parameters should not be the main result. Rather, since the test plan is to be used as a tool for determining performance of new system, the *performance* of the test plan should be emphasized. Even in situations that the parameters T and c are poorly estimated, if the classification rates are not unreasonable then the test plan can still be useful.

Furthermore, the standard producer's and consumer's risk will be evaluated to ensure that they are, at a minimum, less than or equal to their chosen levels γ and δ.

Lastly, in a Bayesian analysis, the distance between the posterior distributions and the prior distribution may also be potentially of interest. In this paper, the Kolmogorov-Smirnov test statistic between the prior and the empirical posterior distribution of a parameter was, when appropriate, used to evaluate the distance between distributions.

3.4 Unsolvable Systems

It must also be noted that occasionally, sets of data were encountered where solving for T and c was not numerically possible. These sets were marked by observing the T value, which would be extremely close to the upper boundary of potential values set in the numerical maximization routine. These particular data sets were ignored for the reasons that they formed only a very small proportion of all data sets (generally less than five instances for every 5000 simulated data sets) and that unsolvable data sets would generally not be considered in real application.

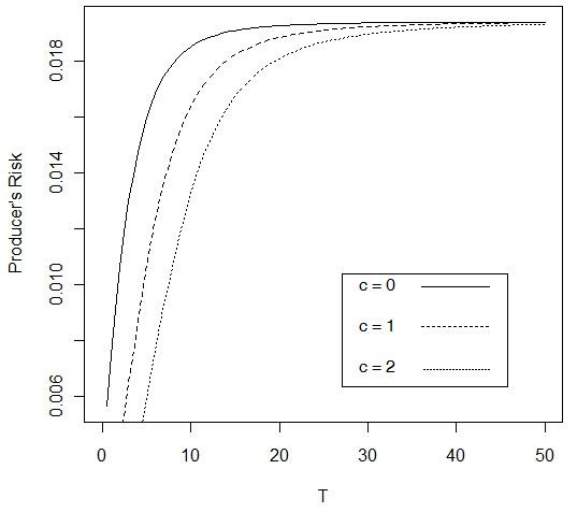

Figure 3.1 Producer's Risk vs. T for c = 0, 1, 2 in an Unsolvable Data Set

Furthermore, as shown in Figure 3.1, simply increasing the allowable number of failures does not resolve the problem. There is a limiting producer's risk, the proportion of draws from the posterior predictive distribution for λ that are less than λ_r. The test plan tends towards this limiting value in all cases as the required testing time T increases and the test becomes increasingly difficult and eventually nearly impossible to pass. Furthermore, changing the value of c only changes the rate at which the test plan approaches this limit by allowing more failures. Note that limiting producer's risk will be different for each system, as each system will have a different draw from the posterior predictive distribution for λ based on the posterior distribution for α and β.

Section 4. Analysis of Baseline Situation

4.1 Baseline Parameters

For the analysis, a baseline scenario was needed both for a general analysis of the test plan and as a guide with which to compare the change in certain variables when introducing deviations into the plan or model. After considering many different choices, it was decided to use a Gamma(4, 2.6) distribution for the failure rates λ_i for the reason that in preliminary simulations, it appeared that the specific shape of the distribution was less important in the outcome of the simulation analysis than the total probability bounded by the producer's and consumer's risks, and because this particular distribution worked well numerically.

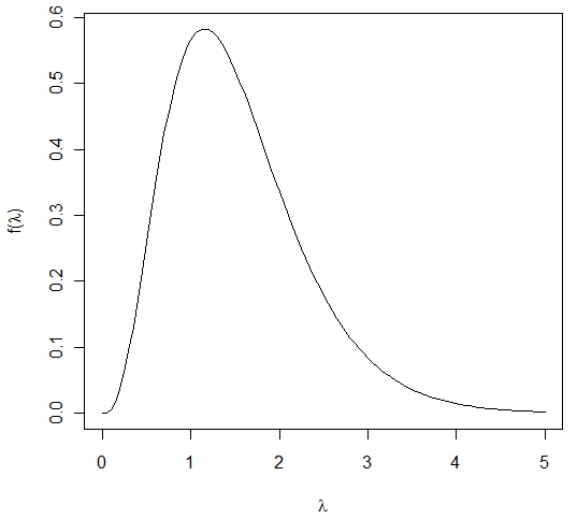

Figure 4.1 Distribution of λ_i

The acceptable and rejectable error rates were set at $\lambda_R = 1$ and $\lambda_A = 2$, as approximately

50% of the density of the distribution is contained within these two values.

The acceptable producer's and consumer's risks were set at $\gamma = \delta = 0.05$.

For priors on α and β, independent noninformative Inverse Gamma(0.001, 0.001) distributions were used.

5000 sets of data were randomly generated from the model. Posterior distributions were computed using a combination of Gibbs sampling and the Metropolis-Hastings algorithm. A period of 3000 burn-in was used and each 10000 draws from the posterior distribution were generated. Due to the large number of simulations, each posterior chain could not be directly observed for lack of time; however, diagnostics were used to ensure an overall general sufficiency in the number of draws. 10000 draws from the posterior predictive distribution for α and β were generated, one for each (α, β) pair.

For each data set, a corresponding test plan was determined. For each test plan, 500 error rates and corresponding times between failure were generated by the model, run through the test plan, and classified as described above.

4.2 Results from Baseline Analysis

The first variables potentially of interest are the testing parameters themselves, T and c. Since the test plan was repeatedly generated from data randomly generated from the model, the distribution of both testing parameters is available. These distributions will mostly be of interest in comparison to scenarios where poor estimation of the parameters is a concern, such as a prior with strong incorrect information or an incorrect model specification.

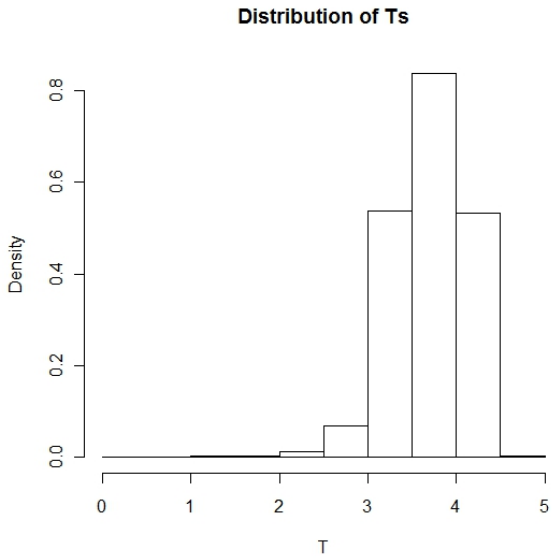

Figure 4.2 Distribution of Ts in Baseline Scenario

The distribution of the Ts is skewed slightly to the left with a mean of 3.702 and a 95% credible interval of (2.842, 4.355). The cs are better summarized in tabular form:

Number of Failures	Frequency	Relative Frequency
0	2	0.0004
1	7	0.0014
2	32	0.0064
3	144	0.0288
4	1885	0.3770
5	2930	0.5860

Table 4.1 Frequency Table of cs From Simulation

So there is a better than 50% chance of generating a test plan that requires 5 failures before the total time reaches T.

Another variable of interest is the pass rate, as all other variables will be affected by this. In the simulations, it appeared that a test plan that passed nearly everything had a low T and

c, and the consumer's risk was dominated by $P(\lambda \geq \lambda_A)$. Also, since the producer's risk is a probability conditioning on the test failing, the estimate of the producer's risk was be incredibly poor due to the low number of failures. Similarly, test plan that failed nearly everything had a poor estimate of the consumer's risk due to the low number of passes. Thus, the pass rate must be reasonable in order to insure good estimation of other variables of interest. This is also of practical use – there is little motivation to conduct a test when one already knows the result with high certainty.

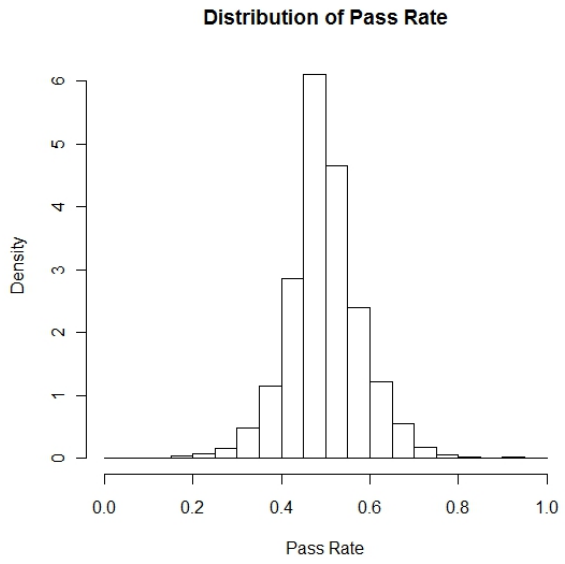

Figure 4.3 Distribution of Pass Rate in Baseline Scenario

For the baseline scenario, the distribution of the pass rate is almost perfectly symmetric with a mean of 0.499 and a 95% credible interval of (0.332, 0.680). Hence, the estimates of variables dependent on the pass rate will not be invalid due to sample size.

Consider now the producer's and consumer's risks.

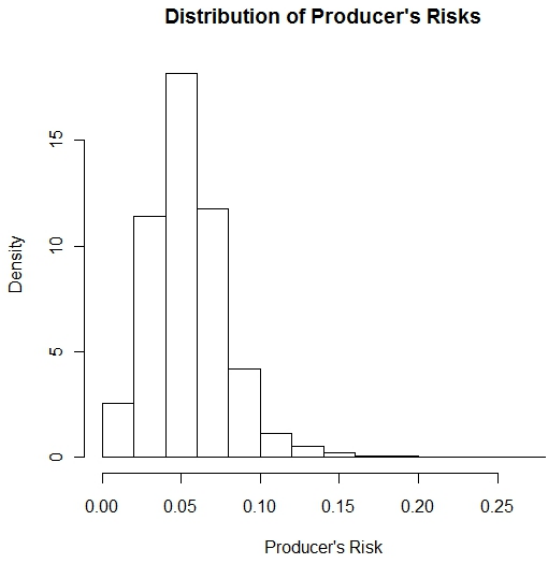

Figure 4.4 Distribution of Producer's Risk in Baseline Scenario

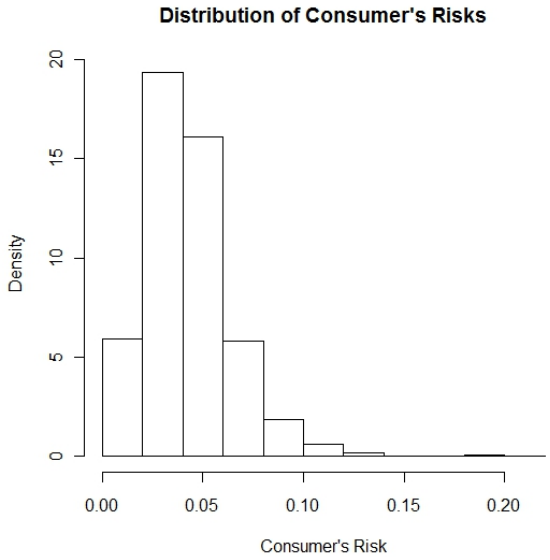

Figure 4.5 Distribution of Consumer's Risk in Baseline Scenario

Both distributions skew towards the right. The producer's risk has a mean 0.054 with a 95% credible interval of (0.015, 0.109), while the consumer's risk has a mean of 0.043 with a 95% credible interval of (0.009, 0.095), so the producer's and consumer's risks are being held

to their predetermined levels with a reasonable amount of variance.

However, consider the secondary producer's and consumer's risks, which were unconstrained in the formulation of the test plan.

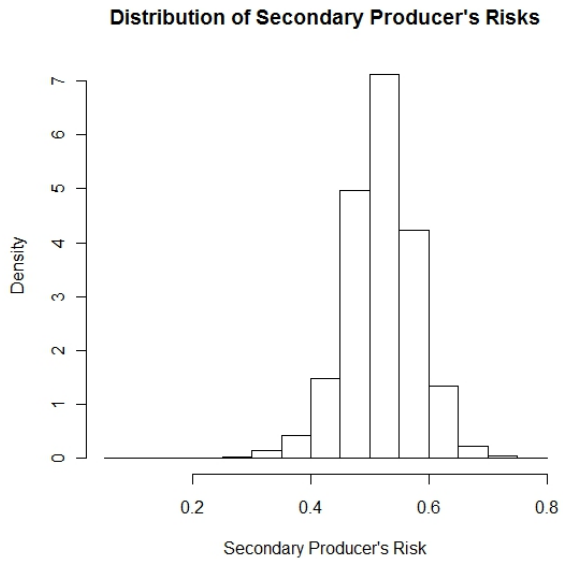

Figure 4.6 Distribution of Secondary Producer's Risk in Baseline Scenario

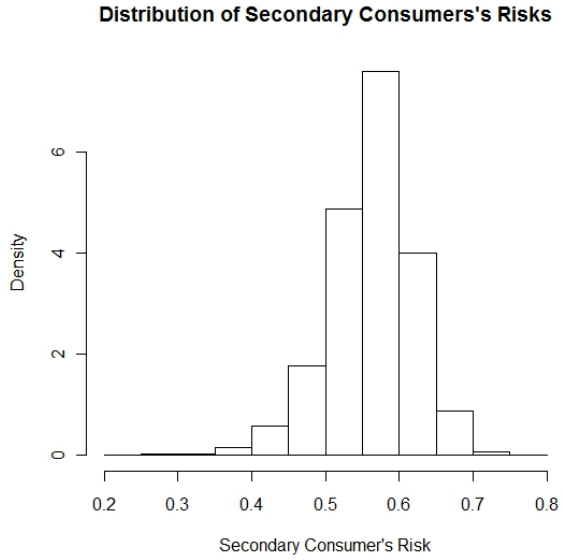

Figure 4.7 Distribution of Secondary Consumer's Risk in Baseline Scenario

Both distributions are, again, almost ideally symmetric, as neither risk is being held to certain values, as was the case with the producer's and consumer's risk. The secondary producer's risk has a mean of 0.520 with a 95% credible interval of (0.392, 0.635), while the secondary consumer's risk has a mean of 0.563 with a 95% credible interval of (0.433, 0.668). So often more than 50% and occasionally more than 60% of the time, the test plan is *passing* a system with a higher failure rate than λ_R and $failing$ a system with a lower failure rate than λ_A, which is certainly something to keep in mind when using the test plan.

Now observe the classification rates. The producer's and consumer's classification rates were quite similar, both distributions skewing slightly to the left, and both averaging close to 70%, with 95% credible intervals of (0.568, 0.802) and (0.560, 0.800), respectively.

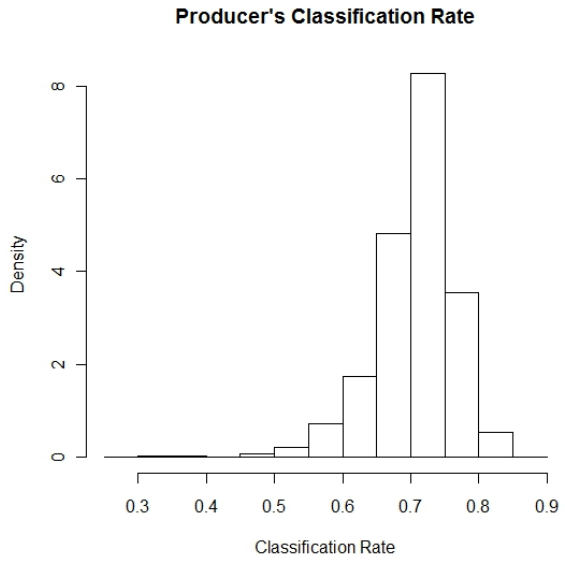

Figure 4.8 Distribution of Producer's Classification Rate in Baseline Scenario

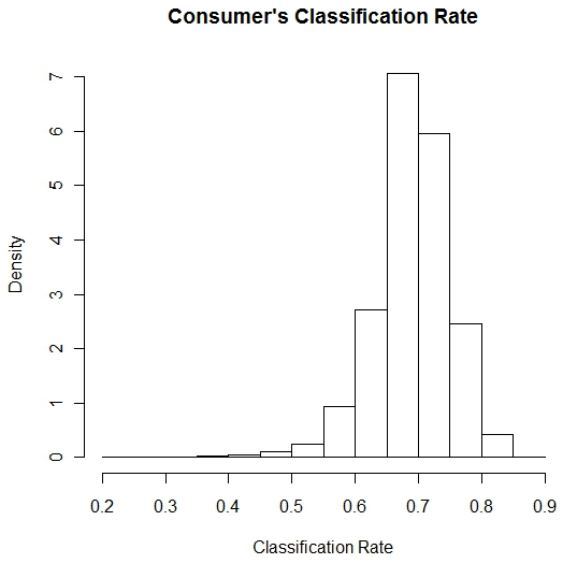

Figure 4.9 Distribution of Consumer's Classification Rate in Baseline Scenario

The mean combined classification rate was 0.451 with a 95% credible interval of (0.398, 0.498), so the test plan was correctly classifying systems with regards to whichever criterion you choose approximately forty to fifty percent of the time.

In this paper all further alterations of the test plan will be simple deviations from these baseline parameters, holding all others constant. These basic statistics are the ones that will be used to compare resulting test plans.

Section 5. Alteration of Acceptable and Rejectable Error Rates

The most direct influence the experimenter has on the variables of interest is through the manipulation of the acceptable and rejectable error rates λ_A and λ_R. It is easy to see why this is so – holding all other things equal, raising the acceptable error rate means that *in every scenario,* it is less likely that $\lambda \geq \lambda_A$, so the producer's risk is likely to be lower. Similarly, lowering the rejectable error rate λ_R lowers the consumer's risk. An example illustrates this.

Consider deviating from the baseline parameters only through the manipulation of λ_A and λ_R. In particular, use $\lambda_A = 0.85$ and $\lambda_R = 2.25$. Observe the distribution of the pass rate:

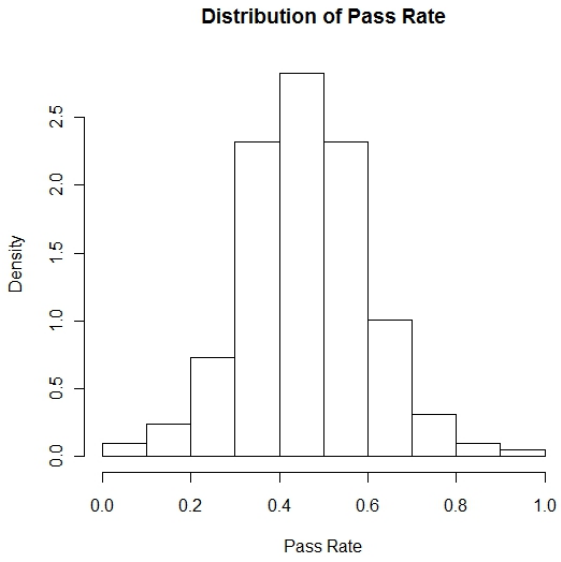

Figure 5.1 Distribution of Pass Rate With Widened Risk Levels

The distribution maintains the same symmetric shape as seen in the baseline scenario. Similarly, the mean pass rate is now 0.461, comparable to the baseline pass rate. However, it is immediately obvious that widening the risk levels widened the distribution of the pass rate,

making it much more likely to produce a test that passes nearly everything or nearly nothing. This is shown in the much wider 95% credible interval for the pass rate of (0.172, 0.758).

As hoped, the producer's and consumer's risks were held at 0.055 and 0.037, respectively. However, the secondary producer's and consumer's risks increased drastically.

Variable	Estimate	Lower CI Bound	Upper CI Bound
Secondary Producer's Risk	0.653	0.477	0.778
Secondary Consumer's Risk	0.718	0.596	0.810

Table 5.1 Estimates and 95% Credible Interval Widths for Risks

Similarly, the classification rates *decreased*

Variable	Estimate	Lower CI Bound	Upper CI Bound
Producer's Classification Rate	0.657	0.408	0.822
Consumer's Classification Rate	0.587	0.336	0.780
Combined Classification Rate	0.296	0.232	0.344

Table 5.2 Estimates and 95% Credible Interval Widths for Classification Rates

While the method of fitting the test plan *does* hold the producer's and consumer's risks to their predetermined levels or less, it appears that increasing the distance between λ_A and λ_R decreases any general measure of performance of the test plan. The question now arises to as why this occurs.

Going back to Figures 3.1 and 3.3, the classification tables for the test with respect to λ_A and λ_R, notice again that the *primary* consumer's risk for one table is the *secondary* risk for the other. Hence, when $\lambda_A = \lambda_R$, the two tables are the same, and so the primary and secondary risks are equivalent, and by using the algorithms described earlier to hold the producer's and consumer's risks to certain levels, the secondary producer's and consumer's risks must be held to those levels as well. Meanwhile, if both sets of risks are being decreased, the probability of

a "correct" classification must be being increased, and hence the classification rates will see a noticeable increase as well.

However, this is not without cost. Nine test plans were fit using the baseline conditions except choosing $\lambda_A = \lambda_R = 1$.

T	c	Combined Classification Rate	Producer's Risk	Consumer's Risk
64.119	60	0.938	0.060	0.070
89.224	82	0.932	0.069	0.066
66.772	59	0.944	0.065	0.020
74.499	72	0.960	0.034	0.059
75.111	69	0.942	0.069	0.024
84.001	80	0.958	0.035	0.062
52.238	49	0.942	0.061	0.050
82.470	76	0.914	0.087	0.084

Table 5.3 Test Plans for Equal Acceptable and Rejectable Error Rates

An ideal test plan should have risks close to zero and classification rates close to one, which would indicate that by all measures of classification, the probability of a correct classification is high. By this definition, these test plans are ideal. Both the producer's and consumer's risk are close to their nominal 5% levels, and the classification rates are close to 95%. However, not only are these plans computationally difficult to compute, they may not be particularly practical – failing a test plan would require running between 59 and 80 units to failure, and if the number of units available for testing *is* a concern due to constraints of time or money, the increased sample size may cause the test to not even be feasible to run. Furthermore, the idea of acceptable and rejectable risks was certainly developed with a practical use in mind, and a point may be reached where the distance between the two error rates can not be decreased without compromising the ability of the experimenter to give specific reasons in the context of

real-world problem why the acceptable and rejectable error rates should be set at such high or low values.

Section 6. Prior Sensitivity Analysis

In any Bayesian analysis, it is worthwhile to inquire as to the effect the prior distribution has on the estimation of the parameters. In this case, the question arises as to the effect of prior information on the test plan. Recall that in the baseline scenario, both parameters α and β of the gamma distribution were assigned noninformative Inverse Gamma(0.001, 0.001) priors.

6.1 Correct Prior

Consider first the situation in which an experimenter has strong, correct information as to the values α and β, and can choose a prior appropriately. In this instance, the deviation from the baseline scenario will be in assigning priors as $\alpha \sim$ Inverse Gamma(11, 40) and $\beta \sim$ Inverse Gamma(11, 26). To see the effect of the additional information, observe the distribution of pass rates.

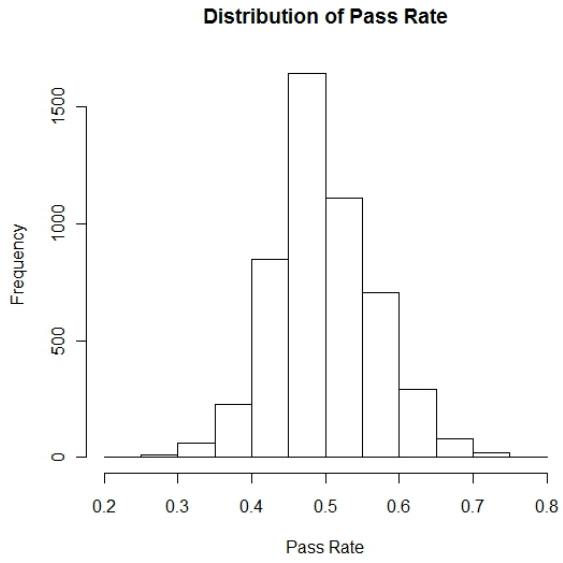

Figure 6.1 Distribution of Pass Rate With Informative Prior

The mean pass rate was 0.499 – similar to the baseline pass rate up to three decimals – but a 95% credible interval on the pass rate is (0.370, 0.642) – a several point decrease in the width of the interval. Similar results were observed for nearly every variable.

Variable	Baseline CI Width	Informative Prior CI Width
Pass Rate	0.348	0.272
Producer's Risk	0.094	0.077
Consumer's Risk	0.086	0.071
Producer's Classification Rate	0.240	0.196
Consumer's Classification Rate	0.234	0.194
Combined Classification Rate	0.100	0.090

Table 6.1 95% Credible Interval Widths for Predictive Distributions of Variables of Interest

The means of the variables using correct prior information remained close to the baseline means. However, the widths of the credible intervals are uniformly *smaller* when correct prior

information is used. Hence, it appears it is likely worthwhile to pursue such information, as estimates of variables of interest will have a smaller variance.

6.2 Incorrect Prior

Consider a situation in which a researcher has *incorrect* information regarding the values α and β. This scenario will be represented as a deviation from the baseline with $\alpha \sim$ Inverse Gamma(11, 50) and $\beta \sim$ Inverse Gamma(11, 10), so that α is believed to be greater than its actual value and β is believed to be lower, and the expected value of the error rate is believed to be 3.25 times greater than it actually is.

As one might expect, the distributions of T and c were different than the respective distributions in the baseline scenario.

Number of Failures	Frequency	Relative Frequency
2	3	0.0006
3	77	0.0154
4	2470	0.4940
5	2450	0.4900

Table 6.2 Frequency Table of cs With Incorrect Prior

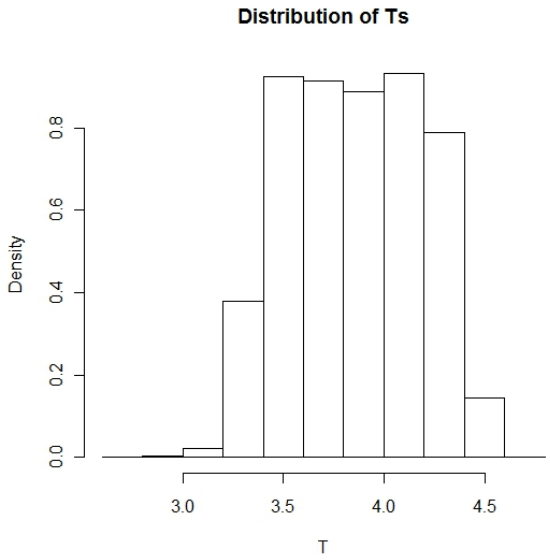

Figure 6.2 Distribution of Ts With Incorrect Prior

The distribution of cs now splits almost evenly between 4 and 5 failures, and likewise the distribution of Ts loses any sense of its ideal symmetric and unimodal shape. The mean is now 3.858, which can mean a significant increase if the T represents thousands of hours, as in Table 1.1.

However, the test plan should not be judged by the estimation of the parameters but by the *performance* of the plan. Interestingly, similar results were observed to the case where a correct informative prior is used. The pass rate decreased, yet the producer's risk and consumer's risk remained at similar levels. The producer's classification rate decreased slightly, and this was matched by a slight increase in the consumer's classification rate. The combined classification rate increased. Furthermore, a reduction in interval widths similar to those obtained when using a correct prior was observed.

Variable	Baseline Estimate	Incorrect Prior Estimate	Interval Reduction
Pass Rate	0.499	0.467	-0.110
Producer's Risk	0.054	0.060	-0.015
Consumer's Risk	0.043	0.0359	-0.026
Producer's Classification Rate	0.691	0.732	-0.078
Consumer's Classification Rate	0.708	0.671	-0.048
Combined Classification Rate	0.451	0.452	-0.008

Table 6.3 Estimates of Variables of Interest in Baseline and Incorrect Prior Scenarios and Reduction of Credible Interval Width

Uniformly a reduction in width of 95% credible intervals was observed. Keep in mind that for several of these variables, the range of the interval is only approximately 10 percentage points, so seemingly small reductions in interval widths can be large relative to the size of the interval itself.

The Kolmogorov-Smirnov test statistic between the prior distribution and the empirical posterior distribution was used as a measure of similarity between the prior and posterior. For the correct informative prior, the mean K-S test statistics for α and β were 0.238 and 0.226, respectively. For the incorrect prior, the mean K-S test statistics were 0.537 and 0.849 (for the baseline scenario with a noninformative prior, the statistics were very close to 1). So though the correct informative prior was clearly more desirable in terms of posterior fit, it appears that the use of generally reasonable prior information can also improve the estimation of the test plan.

Section 7. Incorrect Specification of Distribution

Robustness of a model is always an issue of concern. For example, despite a researcher's best efforts, it might happen that the model is specified incorrectly, and with no knowledge of the robustness of the test plan it is completely unknown whether the resulting analysis carries any worth. For this paper, robustness of the plan was investigated when the model used for the data is still a homogenous Poisson process with a gamma distribution for the λ_i, but the underlying data generating mechanism is different.

7.1 Correlated Times Between Failures

In almost any statistical analysis, an assumption of independence is required at some level. Recall that in fitting these test plans, times between failure events are the components of the model assumed to be independent. In the baseline example, times between failures were modeled as $P(t_{i_j} = x) = \lambda_i e^{\lambda_i x}$. Consider the situation where they are instead the product of an autoregressive function, $t_{i_j} = 0.5t_{i_{j-1}} + 0.5t^*_{i_j}$, where $t^*_{i_j}$ is randomly simulated from an exponential(λ_i) distribution.

Surprisingly, the performance of the test plan with correlated times between failures was only very slightly worse than the performance when using the baseline parameters and correct model specification. Most statistics describing the test plan itself were similar, with the average T value being 3.693 with a 95% credible interval of (2.842, 4.355) and a similar distribution of cs. The pass rate averaged 0.494 with a 95% credible interval of (0.334, 0.662).

Variable	Baseline Estimate	Correlated Times Estimate
Producer's Risk	0.054	0.061
Consumer's Risk	0.043	0.05
Secondary Producer's Risk	0.52	0.52
Secondary Consumer's Risk	0.523	0.563
Producer's Classification Rate	0.708	0.706
Consumer's Classification Rate	0.691	0.680
Combined Classification Rate	0.451	0.444

Table 7.1 Estimates of Variables of Interest in Baseline and Correlated Times Between Failures Scenarios

Statistics describing the performance of the test plan were similar in both scenarios, with the correlated times between failure scenario seeing a slight decrease in performance uniformly across all variables. Hence, it appears the test plan is resistant to at least moderate correlation in times between failures.

7.2 Multiple Distributions

As an example of a distribution that is misspecified at a lower level in the hierarchy, suppose that the data, which is assumed to be m systems, is actually m_1 systems with $\lambda_i \sim$ Gamma(α_1, β_1) and m_2 systems with $\lambda_i \sim$ Gamma(α_2, β_2), where α_1 and β_1 are very different than α_2 and β_2.

This was modeled in simulations was by choosing two gamma distributions: the first, the standard baseline gamma distribution of Gamma(4,2.6), and the second, a Gamma(10, 1) distribution.

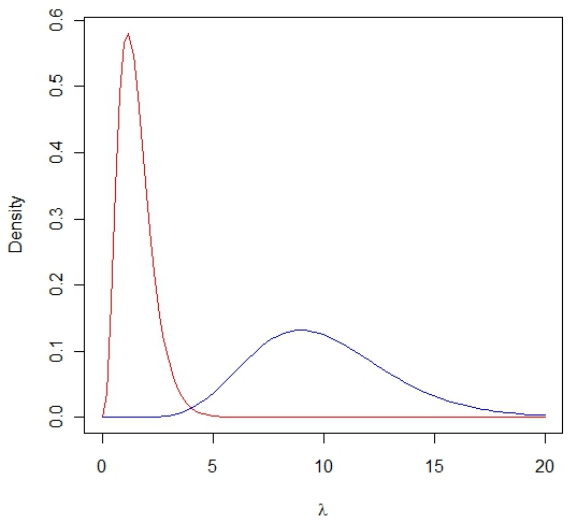

Figure 7.1 Distributions of λ

For any given λ_i, there was a 75% chance that it would be simulated from the second gamma distribution, and only a 25% chance that it would come from the first distribution.

As expected, there were multiple issues in both the fit and the performance of the model. The most glaring issue that a researcher would notice immediately is the new pass rate of 0.898 with a 95% credible interval of (0.822, 0.972). Note that this is not necessarily a function of the disparity between the two distributions – closer distributions were tried in the simulation process, and the results were even worse.

However, more distressing are the producer's and consumer's risk. The producer's risk was estimated at zero. Out of the 255,489 simulated tests that were failed, not once was $\lambda > \lambda_R$. While is the lowest possible producer's risk, there is a cost. The consumer's risk was estimated at 0.363 with a 95% credible interval of (0.294, 0.429). Similarly, the secondary consumer's risk was estimated at zero, and the secondary producer's risk was estimated at 0.779 with a (0.735, 0.820) 95% credible interval

The classification rates followed a similar pattern. The consumer's classification rate was 0.674 with a 95% credible interval of (0.592, 0.754), but the producer's and combined classification rates were 0.300 with a 95% credible interval of (0.222, 0.384).

7.3 Increasing Times Between Failures

In this analysis, it is also the case that the model assumes times between failures are identically and independently distributed. In the real world, this is not always a valid assumption – units that break may tend to function for longer or shorter periods of time after being repaired, for various reasons.

Again, recall that times between failures were modeled as $P(t_{i_j} = x) = \lambda_i e^{\lambda_i * x}$. Consider a situation in which as testing continues, the failure rate is increasing, and so times between failures are decreasing. In particular, consider $P(t_{i_j} = x) = ke^{k*x}$, where $k = \lambda_i^2 * \exp(\frac{t_j * 26}{log(2)})$ and t_j is the total testing time up until failure j. These particular numbers were chosen to ensure that a system with an initial failure rate of 1.538462 will have a failure rate twice as large when 26 units of testing time are reached and will continue to increase exponentially after that point. Twenty six was chosen as it is close to the average amount of testing time required for a system with failure rate 1.538462 to reach 30 failures, and 1.538462 was chosen as the average λ_i value.

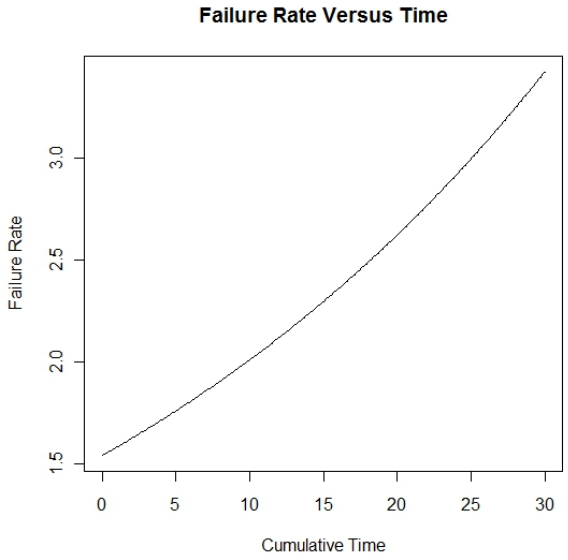

Figure 7.2 Increase in Failure Rate with Time for Starting $\lambda = 1.538462$

The pass rate of the test now becomes 0.26 with a 95% credible interval of (0.156, 0.342),

and with regards to the performance of the test plan, a pattern begins to emerge similar to what was seen in the multiple distributions scenario, not in the particular numbers, but in the apparent ebb and flow of the error rates. This time the producer's risk *increased* to have a mean of 0.098 with a 95% credible interval of (0.049, 0.170), while the consumer's risk *decreased* to a mean of 0.003 with a 95% credible interval of (0.000, 0.015). Conversely, secondary producer's risk *decreased* to a mean of 0.258 with a 95% credible interval of (0.133, 0.370) while the secondary consumer's risk *increased* to a mean of 0.678 with a 95% credible interval of (0.617, 0.737).

In terms of the classification rates, performance was clearly but not significantly affected. The producer's classification rate increased to a mean of 0.858 with a 95% credible interval of (0.814, 0.894), while the consumer's and combined classification rates decreased to means of 0.497 and 0.428 and 95% credible intervals of (0.3, 0.58) and (0.35, 0.486), respectively.

7.4 Incorrect Specification of Distributions

Lastly, consider the situation in which times between failures are Weibull distributed rather than exponentially distributed. This example deviates from the baseline by assigning the following specific form to the times between failures.

$$t_i = \sum_{j=1}^{s} t_{i_j}$$
$$P(t_{i_j} = x) = \frac{1.5}{\lambda_i} \left(\frac{x}{\lambda_i} \right)^{0.5} e^{-(x/\lambda_i)^{1.5}}$$

The results are presented in tabular form.

Variable	Baseline Estimate	Weibull Times Estimate
Producer's Risk	0.054	0.052
Consumer's Risk	0.043	0.005
Secondary Producer's Risk	0.52	0.379
Secondary Consumer's Risk	0.523	0.612
Producer's Classification Rate	0.708	0.813
Consumer's Classification Rate	0.691	0.615
Combined Classification Rate	0.451	0.465

Table 7.2 Estimates of Variables of Interest in Baseline and Weibull Times Between Failures Scenarios

Once again, there is a pattern of decreasing performance with regards to one error rate and increasing performance with regards to the other. However, in this situation the result is not too extreme – the producer's and consumer's risks stayed at their predetermined levels, and the combined classification rate actually *increased*. Interval widths were similar to other scenarios.

7.5 General Pattern

These examples represent only a few possible misidentifications of the model, yet the results are enlightening. When pushed towards breaking, the test plan responds by increasing performance in one area and decreasing performance with in another. This pattern is seen in almost every scenario tested. Furthermore, while it has been shown that a test plan with identical acceptable and rejectable error rates is ideal in terms of risks and classification rates, these results show a benefit to the dual error rate nature of fitting the test plan.

Section 8. Recommendations

Based on the simulation results, there are some recommendations for how to perform future analyses using this model and test plan. Care must be taken when selecting the acceptable and rejectable error rates λ_A and λ_R. In order to minimize risks and maximize classification rates these should be as close to each other as possible such that both a reasonable test plan is still produced and the the model is still interpretable. The secondary producer's and consumer's risks should be also be taken into account when forming the test plan. It may be that they are not of particular interest; yet even if they are not, the criterion for their minimization remains the same as the criterion for maximizing the classification rates: equal λ_A and λ_R.

Prior information should be used whenever available. As shown, even prior information that assigns a large amount of mass to α and β values different than the true ones can have a beneficial effect on the error and classification rates, specifically in a decrease in the variation of these rates, even though estimation of T and c are different than with a noninformative prior. And in the case of correct prior information, even weak prior information, the benefit is a smaller chance of fitting a test plan that produces extreme values in terms the risks and classification rates.

As for robustness of the model in estimation, the dual error rate nature of fitting the model appears to provide a certain advantage in that there is an apparent ebb and flow to the way the test plan responds to misspecification. Generally, where performance decreases with respect to the acceptable or rejectable error rate, performance increases with respect to the other. Unfortunately, based on these simulation results there does not appear to be a clear way to predict which direction the performance will tip towards. However, the amount of change is not usually extreme, and as shown, the risks and classification rates could potentially still be

reasonable.

BIBLIOGRAPHY

Berger, J. O. (1984) , "Robust Bayesian Analysis: Sensitivity to the Prior," *Journal of Statistical Planning and Inference*, 25, 303-328.

Brush, G. G. (1986), "A Comparison of Classical and Bayes Producer's Risk," *Technometrics*, 28, 69-72.

Gaver, D. P., and O'Muircheartaigh, I. G. (1987), "Robust Empirical Bayes' Analysis of Event Rates," *Technometrics*, 29, 115.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities,", *Journal of the American Statistical Association*, 410, 398-409.

Guida, M., Calabria, R., and Pulcini, G. (1989), "Bayes Inference for a Nonhomogeneous Poisson Process with Power Intensity Law," *IEEE Trans. Reliability.*, 38, 603-609.

Guthrie, D. and Johns, M. V. (1959), "Bayes Acceptance Sampling Procedures for Large Lots," *Ann. Math. Statist.*, 30, 896-925.

Hald, A. (1960), "The Compound Hypergeometric Distribution and a System of Single Sampling Inspection Plans Based on Prior Distributions and Costs," *Technometrics*, 2,

275-340.

Hamada, M. S., Wilson, A. G., Reese, C. S., and Martz, H.F. (2008), *Bayesian Reliability*, New York: Springer.

Sun, D., and Berger, J. O. (1994),"Bayesian Sequential Reliability for Weibull and Related Distributions," *Annals of the Institute of Statistical Mathematics*, 46, 221-249.