Failure record discounting in Bayesian analysis in Probabilistic Risk Assessment (PRA) – A space system application

by

Spyridon-Damianos Lekkakos

B.S. Aeronautical / Mechanical Engineering
Hellenic Air Force Academy, 1995

Submitted to the System Design & Management Program in
Partial Fulfillment of the Requirements for the Degree of

Master of Science in Engineering and Management

at the

Massachusetts Institute of Technology

May 2006

Signature of Author: …….
System Design & Management Program
May 14, 2005

Certified by: ……..
George E. Apostolakis
Professor, Engineering Systems / Nuclear Engineering
Thesis Supervisor

Accepted by: …….
Pat Hale
Director, SDM Fellows
Senior Lecturer, Engineering Systems Division

Failure record discounting in Bayesian analysis in Probabilistic Risk Assessment (PRA) – A space system application

by

Spyridon-Damianos Lekkakos

Submitted to the System Design & Management Program
on May 12, 2006 in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Engineering and Management

Abstract

In estimating a system-specific binomial probability of failure on demand in Probabilistic Risk Assessment (PRA), the corresponding number of observed failures may be not directly applicable due to design or procedure changes that have been implemented in the system as a result of past failures. A methodology has been developed by NASA to account for partial applicability of past failures in Bayesian analysis by discounting the failure records. A series of sensitivity analyses on a specific case study showed that failure record discounting may result in failure distributions that are both optimistic and narrow.

An alternative approach, which builds upon NASA's method, is proposed. This method combines an optimistic interpretation of the data, obtained with failure record discounting, with a pessimistic one, obtained with standard Bayesian updating without discounting, in a linear pooling fashion. The interpretation of the results in the proposed approach is done in such way that it displays the epistemic uncertainties that are inherent in the data and provides a better basis for the decision maker to make a decision based on his / her risk attitude. A comparison of the two methods is made based on the case study.

Thesis Supervisor: George E. Apostolakis
Title: Professor, Engineering Systems / Nuclear Engineering

(This page intentionally left blank)

Acknowledgments

First, I would like to thank my thesis advisor, Professor George E. Apostolakis, for his valuable support and guidance throughout the completion of this work.

I would also like to thank the SDM staff and my classmates, all of whom I have learned from, laughed with, and admired. Many of them have become and will remain personal friends. A special thanks to my colleague Christian LaFon, with whom I have spent innumerable hours working on the group projects and assignments during the SDM program.

Finally, I would like to express my gratitude to the Hellenic Air Force for financially supporting my academic pursuit.

This work is dedicated to my parents, George and Vaso, and my brother Stathis.

(This page intentionally left blank)

Table of Contents

List of Figures

(This page intentionally left blank)

List of Tables

(This page intentionally left blank)

1. Introduction

a) General Context

The risk analysis of high-reliability, complex systems, such as nuclear and space systems, can seldom be performed on the basis of large statistical databases. The reason is that, most of the times, there is a lack of operating data, while failures, if any, are often realized through equipment exhaustive testing and not during actual system operation. Because of the sparsity of relevant empirical data, all available sources of information (i.e., either from testing activities or from experience with similar items in a comparable application and environment), is essential to the assessment of a system's failure probability.

Moreover, throughout the life cycle of high-reliability systems, component design changes and test / inspection changes are implemented in response to observed failures or failure conditions, with the purpose of improving the system's reliability. Once a design or process has been changed, some fundamental risk analysis factors, such as the number of past failures and, consequently, past estimates of the failure rates, may not be directly applicable for use in system safety analysis.

Failure record discounting is an activity initiated by the need to reflect in a system's Probabilistic Risk Assessments (PRA) the "partial applicability" of failure records that have been observed either on similar / heritage systems, or on the same system before the implementation of design or test / procedure changes. The output of failure record discounting is thus a probability distribution that represents expert judgment about applicability of the record to be considered in PRA, instead of the actual occurred number of failures.

It is clear that there is uncertainty inherent in failure record discounting, which should be modeled, quantified, and displayed in the PRA results. Since engineering judgment is an important input to the process, the subjective (Bayesian) interpretation of probability -by which a person's subjective probability of an event describes his / her

degree of belief in the event- is the only natural means to deal with failure record discounting and its effect on the overall results.

There is not yet an accepted methodology for modeling failure record discounting. The problem has two dimensions: First, a system-relevant method for failure record discounting should be established, which must account for the total uncertainty associated with the non-direct applicability of the failures; second, an appropriate Bayesian updating method should be applied to deal with the discounted data. It is important to mention that failure record discounting could impact risk estimation across all PRA elements; therefore, it should be done "carefully".

b) Space System PRAs

Space system PRAs are still in an early development phase, while methodological improvements on existing modeling approaches are required to adequately model the dynamic nature of space flights (Apostolakis, 2004). Moreover, due to the frequent configuration changes (new or improved systems are regularly introduced), there is seldom an abundance of statistics to describe past performance of a specific system (Pate'-Cornell, 2001).

In general, space systems have two basic characteristics:

- There are very few failures that have occurred in flight, but some failures have occurred during testing, or have been discovered during the ground processing flow
- Component design improvements and test / inspection changes are regularly introduced to address failure conditions, causing significant changes in the in-flight reliability of the hardware

There are two levels of uncertainties inherent in the models used in space system PRAs. First, there is epistemic uncertainty about the model parameter values; a situation that is common in almost any PRA. This uncertainty can effectively be addressed within the

Bayesian framework. Second, and probably most important, there is epistemic uncertainty about the model assumptions, the resolution of which requires extensive fundamental research (Apostolakis, 2004) and will only partially be considered in the present work.

If we take into account that space system PRAs are in an early development phase, thus, there is a vague knowledge about the values of the several failure parameters; and, in addition, engineering judgment is a major input to the failure record discounting process; we will recognize that these conditions provide space for bias and optimism in the elicitation and use of subjective information. This is a problem realized in several early PRAs by different industries (Mosleh, Bier & Apostolakis, 1988), when actuarial data -when accumulated- gave risk values beyond the predicted ones. Therefore, the application of failure record discounting to space PRA requires consideration of several aspects, but most importantly, of the inherent epistemic uncertainties, which should be modeled, evaluated, and appropriately presented to enable risk-informed decision making.

c) Reference Case

Our reference case is a system from the National Aeronautics and Space Administration (NASA) Space Shuttle, which has undergone several design changes over the course of the relevant space program. At the system-level, there are two failure modes, "uncontained failure" and "benign failure", the failure records of which were discounted to reflect design improvements and test / inspection changes. The parameter estimation for the probability of system failure per operation (on-demand) was performed at the system-level without further modeling at the component-level.

Over the course of the Space Shuttle program the system of our reference case has evolved with three configurations A, B, and C, which have an estimated 80% commonality in their component lists. Thus, in assessing the probability of failure on demand of the latest configuration C, the failure records from the two heritage

configurations A and B were included in the relevant database and were also subjected to discounting.

NASA has developed a failure record discounting and a Bayesian updating method to account for the implementation of design or test / procedure changes in the system. The aforementioned methods are presented in detail in Chapter 2.

d) Proposition

The objective of this thesis is to evaluate the feasibility of an approach developed by NASA to consider failure record discounting in Bayesian analysis within PRA framework. The case study used as a basis for this work is a Space Shuttle system, on which the aforementioned approach was applied for the first time. Our purpose is to make use of available frameworks, methods, and tools that are widely acceptable by the PRA community in order to:

- Evaluate current methodology applied on our reference case
- Recommend alternative approaches when necessary
- Identify general cautions, pitfalls, and guidance to be considered in failure record discounting
- Make an assessment of the uncertainties associated with failure record discounting and demonstrate them in our final results

e) Structure of this Thesis

We begin by presenting the approaches developed by NASA for failure record discounting and Bayesian updating with discounted data, and the results of their application to our reference case (Chapter 2).

We then proceed with the literature review to present how some relevant to our problem issues have been addressed by the PRA community in other applications (Chapter 3). We continue with the reevaluation of our reference case based on our findings from Chapter 3 and our observations (Chapter 4), and we close with our conclusions (Chapter 5).

(This page intentionally left blank)

2. Existing Methodology

The methodology developed by NASA in (NASA, 2005) consists of two steps: (1) failure record discounting and (2) Bayesian updating with discounted data.

The following part (par. 2.1 to 2.3) is based on (NASA, 2005), where both the methodology and the results are thoroughly described.

2.1 Failure Record Discounting

Failure record discounting begins with the assessment of redesign effectiveness and test / inspection[1] effectiveness, which are the basic factors to be considered. The redesign and test / inspection effectiveness represent the portion of failures prevented by the improved design or the new test / inspection, respectively. The evaluation of the effectiveness is based on engineering judgment, as well as on other factors such as the failure history since the change, the nature of the failure (spontaneous or progressive), etc.

An effectiveness range is used instead of a point estimate to address the subjectivity for a given record and the uncertainty involved in estimating the discount value. NASA has developed guidelines for the initial assessment of fix effectiveness which are presented in Table 1-1.

After the redesign effectiveness, and test / inspection effectiveness have been evaluated, a discount factor (γ) is assigned to the failure record. This discount factor is nominally evaluated as

$$\gamma = (1 - \text{redesign or test / inspection effectiveness}) \qquad (2\text{-}1)$$

[1] Refers to checklists of tests / inspections that must be performed before a mission. Tests and inspections are effective only for those failures that initiate during ground processing and could persist and/or grow to the mission if not detected. They are not effective for failures that could occur during the mission.

Based on the nominal discount factor, the final discount factor is set by engineering judgment. The discount factor (γ) actually represents the applicability of a given failure after the implementation of the fix initiated by the failure.

Rational	Effectiveness
Fix is proven in the field with no known failures since implementation; field time is 6 missions or greater AND hot time is greater than or equivalent to design life (240 starts)	90 – 100%
Engineering is highly confident that the problem is understood and a hardware fix eliminates the problem based on test, analysis, or substantiation; field time is less than 6 missions AND hot time is greater than or equivalent to operational life (60 starts)	80 – 90%
Engineering is moderately confident that the problem is understood and a hardware fix eliminates the problem based on test, analysis, or substantiation; field time is less than 6 missions AND hot time is less than operational life OR the number of rig tests are statistically significant	50 – 80%
Deviation Approval Request (DAR) Limit AND revised procedures, inspections, OR process controls actually affecting the way hardware is processed or handled	70 – 95%
New or revised procedures, inspections OR process controls actually affecting the way hardware is processed or handled	50 – 75%
New or revised procedures which are "goodness fixes" such as caution notes, warnings, etc.	30 – 65%

Table 2-1. Failure record discounting guideline

2.2 Weighted Average Likelihood Bayesian Updating Method

The Weighted Average Likelihood method treats the discount factor (γ) as the subjective probability that the discounted record applies to the present component / system reliability.

More specifically, if n is the total number of failures that have been observed on the system, which must be discounted to reflect the initiated fixes, there is uncertainty about the new number of failures x that should be used in Bayesian parameter estimation. The Weighted Average Likelihood method accounts for this uncertainty by computing the likelihoods that zero, one, two, etc. of the discounted records actually apply to the present circumstances, using a Binomial distribution assumption. Thus, if we assume that an average discount factor (γ) is used for several records, then the likelihoods for each possible outcome for x ($0, 1, \ldots, n$) are estimated from a Binomial distribution assumption, where the average discount factor (γ) is the parameter of the Binomial.

For example, we want to estimate the probability p of a failure on demand[1] accounting for the observed failures not having direct applicability. Assuming that for a given number of n observed failures, in N total number of observed demands, there is a probability γ that an observed failure is applicable; the likelihood function L for x out of n failures being applicable is defined as:

$$L_x = P_x \frac{N!}{x!(N-x)!} p^x (1-p)^{N-x} \qquad (2\text{-}2)$$

where

$$P_x = \frac{n!}{x!(n-x)!} \gamma^x (1-\gamma)^{n-x} \qquad (2\text{-}3)$$

represents the partial failure record applicability, assuming that each observed failure has the same probability γ of being applicable.

Thus, the Bayesian updating is done for each possible outcome, the mean and variance are estimated for each outcome, and an overall weighted average mean and variance are computed, using the weights estimated from a binomial distribution with the discount factor as the parameter.

[1] The method is similar when estimating the probability of failure in a given time period. In this case, Poisson likelihood is used instead of binomial, the total number of demands is replaced by total observed time period, and the binomial parameter p is replaced by Poisson parameter λ. We do not elaborate on the Poisson formulations because it is not applicable in the reference case.

The likelihood L can be extended in various ways. If each failure has a different applicability probability, then P_x in Equation (2-3) can be replaced by the probability that some given failures are applicable and others are not applicable. The likelihood can also be extended to the case where the applicability probability γ has an associated uncertainty distribution. If the distribution is discrete then the likelihood given by Equation (2-2) has an additional summation over the probability for different values of γ. If the uncertainty distribution for γ is continuous then the summation of γ probabilities is replaced by an integral.

Application to the Beta-Binomial Model

Assuming a Beta-Binomial Bayesian model[1] with a Beta prior for p denoted by $Beta(\alpha_o, \beta_0)$ and a probability density function (pdf) of the general form,

$$f(p;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \qquad (2\text{-}4)$$

then according to the Weighted Average Likelihood method, the posterior distribution $\pi_1(p/D)$ of p given the data D, is estimated as follows:

$$\pi_1(p/D) = \sum_{x=0}^{n} P_x Beta(\alpha_0 + x, \beta_0 + N - x) \qquad (2\text{-}5)$$

Thus, the posterior is a weighted average of the individual Beta distributions $Beta(\alpha + x_o, \beta_0 + N - x)$ with weights P_x.

[1] The standard Bayesian approach to estimating p in a Beta-Binomial model and the moment matching method used for estimating the posterior shown in Equation (2-5) are presented in more detail in Appendix C.

2.3 Application of Existing Methodology to the Reference Case

a) Discounting

Over the course of the Space Shuttle program, the system of our reference case has been evolved with three configurations, each of which was accompanied by its own failure record database. Discounting of a record was used when the failure that generated the record resulted in a design change, or a test / inspection procedure change, to reduce the failure frequency due to that failure cause. All previous records from the same failure cause were also discounted by the same amount.

Before the discounting process, an initial screening was carried out by the Agency's analysts to eliminate events that were considered not applicable, such as failures that were test facility induced, failures in experimental configuration, partial failures (or precursors), and failures that were initiated by the reference system but realized in other elements and the opposite (interface failures).

By the time of the reference case PRA, there had been a total of 458 system test failures during an accumulation of 2980 demands and 1,000,226 seconds of operation. After the initial screening, the total number of system failures was reduced to 76 events, while the number of demands and the operating time were adjusted accordingly.

Each failure was then categorized as "uncontained" or "benign". Uncontained failures are failures that lead to catastrophic loss of the system, which is assumed to be Loss of Crew or Vehicle (LOCV). Benign or safe failures are defined as failures that lead to a safe shutdown of the system, without catastrophic consequences for the mission. The applicable failure records for all the three configurations A, B, and C are included in Appendix A.

The fixes (design or test / procedure changes) were then identified from the descriptions included in the problem records, and each fix was assigned an effectiveness factor based on the framework described in Section 2.1. The final discounting factors assigned by NASA to the applicable failures are included in Appendix B.

As far as the number of demands is concerned, due to the various testing procedures and different flight profiles between and within the three system configurations, not all system flight duration times were the same. Therefore, in order to create a common basis for evaluation, the total number of system duration or exposure time was converted into equivalent Configuration C single missions of approximately 520 seconds. The aggregated operational data and the discounted failure records per system version, as given in Appendices A and B, are shown in Tables 2-2 and 2-3, respectively.

Configuration	Operational Time (sec)	# Equiv. Conf. C Missions (520 sec)	# Catastrophic Failures	# Benign Failures
A	158,579	305	0	4
B	685,880	1319	8	16
C	249,600	480	4	1

Table 2-2. Operational data per system configuration

	Catastrophic			Benign		
Conf.	**# Failures**	**# Discounted Failures**	**Applicability Factor (γ)**	**# Failures**	**# Discounted Failures**	**Applicability Factor (γ)**
A	0	0	n/a	4	3	n/a
B	8	1.15	14.37%	16	1.6	10%
C	4	0.2	5%	1	0.05	5%

Table 2-3. Failure record discounting data per system configuration

b) Bayesian Updating

In order to account separately for the three different configurations, a multi-step Bayesian approach[1] was applied for estimating the distribution of the probability of failure for the latest version of the system. The steps followed for the application of the approach to our reference case are shown in the flowchart of Fig. 2-1.

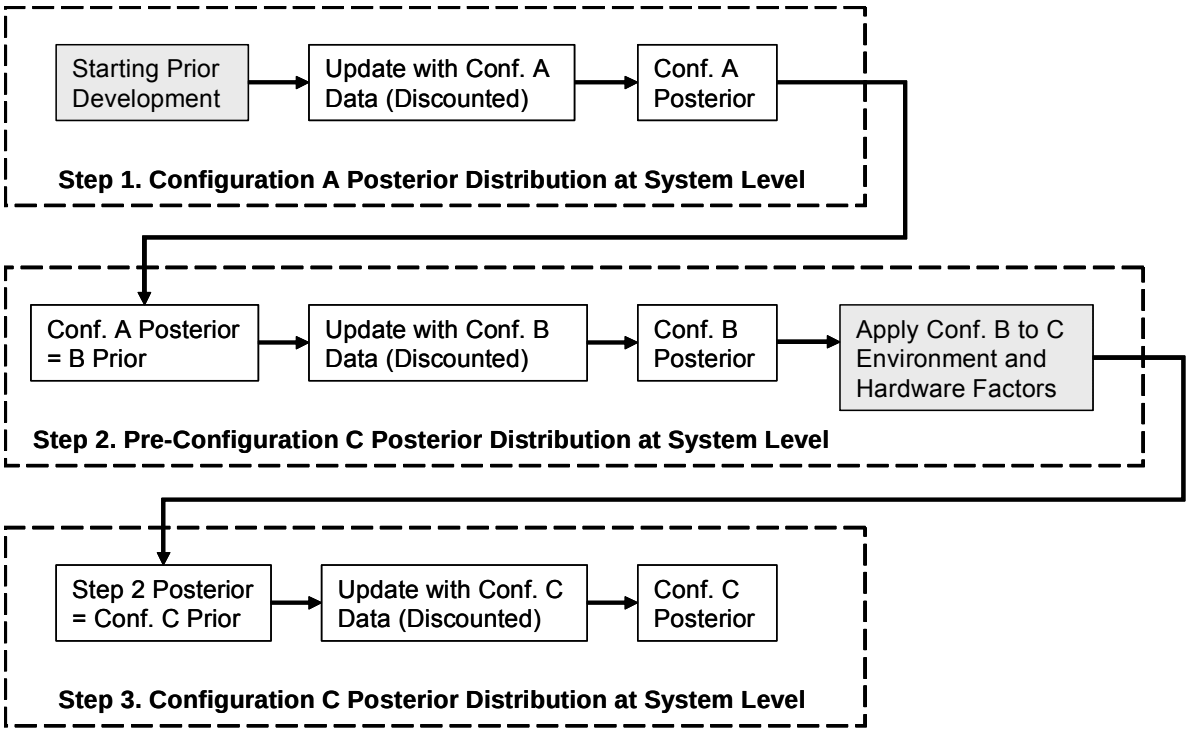

Figure 2-1. Multi-step Bayesian approach flowchart

All the above steps are a straightforward application of the Weighted Average Bayesian Updating method presented in Section 2.2, except Step 1 and "Environmental and Hardware Adjustment" in Step 2, which are described in more detail below.

[1] In our case, a multi-step Bayesian approach means that the posterior distribution of one configuration becomes the prior for the successive.

Step 1. Configuration A Posterior Distribution

The first task in Step 1 is the development of a starting prior. The use of a non-informative prior[1] was considered in order to express a vague state-of-knowledge and let data dominate the resulting posterior. Jeffreys' U-shaped *Beta(0.5, 0.5)* was finally chosen among a Uniform *Beta(1,1)* and a non-dominating *Beta(0.01, 0.01)* as starting prior.

The official Configuration A reliability demonstration and flight data had zero catastrophic failures. This equates to zero failures out of 305 equivalent missions. *Beta(0.5, 0.5)* was then updated with the binomial evidence to give a *Beta(0.5, 305.5)* catastrophic failure posterior for Configuration A.

As far as benign failure posterior estimation is concerned, a different approach was followed based on the fact that the official Configuration A reliability demonstration and flight data had four benign failures. Without performing extensive research into the database, a top-level conservative engineering judgment was used, and it was decided that two of the failures should be discounted by 50% for use in Bayesian updating. Thus, a Beta distribution was used directly, setting α = number of failures = 3 and β = number of demands = 305.

Environmental and Hardware Adjustment

Environmental and hardware adjustment was an intermediate step in which system-wide improvements of Configuration C over Configuration B were considered. Due to the nature of the multi-step Bayesian updating, the Configuration B posterior needed to be adjusted for the Configuration C environmental and hardware improvements before it could be used as Configuration C prior.

The operating environment adjustment represents the improvement in terms of environmental conditions (pressures, speeds, vibrations, temperatures, etc.) on system

[1] Non-informative priors are distributions which are mathematically constructed to represent a vague state of knowledge (Siu & Kelly, 1998).

performance from Configuration B to C. A reliability assessment of Configuration C showed that, at extreme operating conditions (104% power level), the ratio of the failure rates of 14 key components, between Configuration B and C, is considerably large (average 1.689). In other words, the average component failure rate in Configuration B is 1.689 times the one in Configuration C, in a sample of 14 key components. The overall environmental factor was thus approximated by a Uniform distribution with lower bound 1.520 and upper bound 1.858 (average ± 10%).

In addition to the operating environment improvements, Configuration C was also improved over Configuration B due to system-wide redesign and new hardware implementations. This resulted in the elimination of some critical failure modes, improved overall safety margin, expanded operational capabilities, and reduced maintenance. Based on the aforementioned improvements and engineering judgment, a hardware robustness factor was approximated by a Uniform distribution with lower bound 1.1 and upper bound 1.3.

C) Results

The obtained results for the system's uncontained and benign failure probabilities are given in Tables 2-4 and 2-5, respectively. The standard Bayesian approach to estimating p in a Beta-Binomial model and the moment matching approximations used for getting the moments of the posterior distributions, matching the beta posterior to a lognormal distribution, and adjusting for environmental and hardware changes are presented in Appendix C. An average applicability factor (γ) for each of the Configurations B and C was used for the computation of the weights and the estimation of the posteriors by Eqs. (2-3) and (2-5), respectively. The number of discounted failures and the average failure applicability factors that were used in each configuration are shown in Table 2-3.

Uncontained Failure (Catastrophic)						
	Distribution	Mean	Variance	5th	50th	95th
A Posterior	Beta(0.5, 305.5)	1.63E-03	5.32E-06	6.44E-06	7.45E-04	6.27E-03
Evidence	1319 missions / 8 failures (1.15 discounted)					
B Posterior	Beta(1.012, 995.6)	1.02E-03	1.02E-06	5.37E-05	7.08E-04	3.03E-03
Environment Adj.	Beta(1.006, 1656.3)	6.07E-04	3.66E-07			
Hardware Adj.	Beta(1.002, 1965.2)	5.09E-04	2.59E-07			
Evidence	480 missions / 4 failures (0.2 discounted)					
C Posterior	Lognormal	4.91E-04	2.34E-07	9.03E-05	3.50E-04	1.36E-03

Table 2-4. Catastrophic failure results with existing methodology

Benign Failure						
	Distribution	Mean	Variance	5th	50th	95th
A Posterior	Beta(3, 305)	9.74E-03	3.12E-05	2.67E-03	8.70E-03	2.04E-02
Evidence	1319 missions / 16 failures (1.6 discounted)					
B Posterior	Beta(3.44, 1212.4)	1.60E-03	1.57E-06	8.64E-04	2.56E-03	5.70E-03
Environment Adj.	Beta(3.392, 2003.5)	1.69E-03	8.40E-07			
Hardware Adj.	Beta(3.359, 2364.9)	1.42E-03	5.98E-07			
Evidence	480 missions / 1 failure (0.05 discounted)					
C Posterior	Lognormal	1.20E-03	4.25E-07	4.54E-04	1.05E-03	2.43E-03

Table 2-5. Benign failure results with existing methodology

3. Literature Review

The problem of failure record discounting is neither explicitly nor sufficiently addressed by existing methodologies in the literature. In the absence of any particular method, we focus our research on getting insights from different approaches that may help us best address the problem of failure record discounting, as well as resolve some of the issues realized in the existing methodology and its application to the reference case. Specifically, our literature research is focused on the following areas:

- Use of the Bayesian approach in PRA
- Expert Opinion Elicitation
- Bayesian updating methods
- Relevant Cases

Our major findings about these topics are presented in the following sections.

3.1 The Bayesian Approach

The first step in safety studies of technical systems is the development of the relevant "model of the world"; that is, the mathematical model that is constructed for the interpretation of the physical situation of interest (Apostolakis, 2004). A mathematical model is necessary for the estimation of system's safety or risk parameters based on the statistical analysis of available data. The model should be adequately realistic such that the deduced results to have practical value, but, it should also be simple enough to be handled by available computational methods (Rausand, 2004).

In probabilistic models of the world, e.g., one that uses the binomial distribution for estimating the probability of observing r failures in n demands, the resulting probability is conditional on our knowledge about the numerical values of the parameters

p, as well as on our accepting the model assumption of an underlying Bernoulli process (Apostolakis, 1994).

The Bayesian approach makes use of the subjective interpretation of probability, by which a person's subjective probability of an event describes his / her degree of belief in the event. The risk analysis of low-probability events must employ the logic of the subjective (or Bayesian) approach since rarely will enough actual data exist to use the frequentist definition, by which the probability of an event has been defined as its long-run relative frequency.

The Bayesian approach is the most appropriate when dealing with imprecise / uncertain data because it exhibits the following strengths:

- It allows the incorporation of state-of-knowledge uncertainties and the interpretation in the results of their magnitude and effects, it is therefore the one better suited for decision analysis applications, such as those driving PRA (Siu & Kelly, 1998; Paté-Cornell, 2002)
- It can incorporate a wide variety of available information (Siu & Kelly, 1998)
- Even in the absence of large amounts of data it can still be used, without jeopardizing its internal consistency and axiomatic structure (Bier & Mosleh, 1988)
- It exhibits computational advantages since propagation of uncertainties through complex models is relatively simple (Siu & Kelly, 1998)

However, the use of imprecise / uncertain data requires modeling decisions and there are uncertainties inherent in such modeling, which must be quantified and displayed in the results (Siu & Kelly, 1998). Furthermore, in the case of sparse or imprecise data, sensitivity studies are required to provide useful insights about how some of the assumptions affect the analysis (Siu & Apostolakis, 1986).

Apostolakis (Apostolakis, 1990, 1994) distinguishes between epistemic and aleatory uncertainties in risk analysis, which is widely accepted by the PRA community

(e.g., Siu & Kelly, 1998; Paté-Cornell, 2002, O'Hagan & Oakley, 2004). Aleatory (or stochastic) uncertainties arise from inherent randomness in the process (i.e., coin flipping) and are described by the model of the world, whereas epistemic (or state-of-knowledge) uncertainties are due to imperfect knowledge about the validity of model assumptions and the numerical values of its parameters.

This distinction is useful because, while aleatory uncertainty cannot be removed from the model output, epistemic uncertainty is in principle reducible when more or better information is obtained (O'Hagan & Oakley, 2004). The epistemic uncertainties are modeled by epistemic probability models, usually in the form of probability density functions (pdf), which represent our state-of-knowledge regarding the validity of the model assumptions and the numerical values of the parameters (Apostolakis, 1994).

Regarding the epistemic uncertainty related with the set of model assumptions, this is defined by the appropriateness, completeness, and exhaustiveness of the assumptions (Apostolakis, 1994). As far as parameter uncertainties are concerned, these reflect the lack of significant operating experience of the system modeled, data fuzziness, use of generic sources of information, and engineering judgment (Kaplan, 1992; Martz et al, 1996; Siu & Kelly, 1998).

The treatment of uncertainty in risk computations is critical since it affects the basis on which decision making takes place. Paté-Cornell (Paté-Cornell, 2002) states: "when the evidence regarding the fundamental phenomena is incomplete, the use of a sophisticated Bayesian approach that allows for displaying in the results the magnitude and the effects of epistemic uncertainties is necessary".

3.2 Expert Opinion Elicitation

Failure record discounting is, by definition, an activity that involves significant inputs of engineering judgment. The purpose of our review of the literature on expert opinion

elicitation is to identify techniques, guidelines, and cautions that should be considered when eliciting engineering judgment to improve the quality of the results.

3.2.1 Eliciting Engineering Judgment

Engineering judgment and expert opinions are required because risk assessments must deal with rare events (Apostolakis, 1990). Expert opinion elicitation techniques are techniques that involve interviewing experts, and asking them to assess unknown quantities, or probabilities of possible future events. Since this knowledge entails a level of subjective confidence, the methods used for elicitation, quantification and aggregation of expert opinions is an important issue.

The literature stresses that it is important, when conducting an expert opinion **elicitation** process, to do this in a structured, clear and transparent way in order to enhance rational consensus. The notions of bias, precision, and dependence are crucial, and often have a considerable effect on the results; thus they should be considered all the way through structuring, testing, and applying an expert elicitation process. Important issues to consider include the selection of experts, flexibility and process design, interactions among the participants, sensitivity analyses of the different "predictions", and the coordinating role of the risk assessment team (Clemen & Winkler, 1999).

The **aggregation** of expert opinions and the mechanisms that are used are also essential to the quality of the results. These mechanisms can be iterative (e.g., Delphi method), interactive (i.e., meeting of experts with the objective to identify and structure the hypotheses, and to link evidence and the various hypotheses), or analytical (e.g., a Bayesian integration of expert opinions based on the confidence of the decision maker in each source). In general, there is not a unique process for combining probability distributions in risk analysis, but it may involve both mathematical and behavioral aspects, depending on the requirements of a given application (Clemen & Winkler, 1999).

Regarding the **quantification** of expert opinion, probability assessments tend to be more representative of the analyst's state of knowledge when formal methods are used (Apostolakis, 1990). Moreover, if expert opinions are to be combined with data, then the Bayesian formalism is needed in quantifying them (Siu & Kelly, 1998). However, even when formal methods are employed, specific concerns about the completeness of an analysis should be raised when the expert elicitation process produces very low probabilities and frequencies, or overly peaked distributions, artificially indicating a strong state of knowledge (Siu & Apostolakis, 1986; Apostolakis, 1990).

Kaplan (Kaplan, 1992) presents an alternative approach to eliciting expert knowledge, called "expert information" approach, in which experts are asked for their information and knowledge, rather than for their opinions. The motivation behind this indirect approach is based on the assumption that, while the experts presumably have much knowledge in their particular domains, they are usually not trained or experienced in the use of probability as a language with which to express a state of confidence or state of knowledge.

3.2.2 Cautions in the Use of Engineering Judgment in PRA

It has been widely documented that assessment of subjective probabilities are subject to cognitive biases. Mosleh, Bier and Apostolakis (Mosleh, Bier & Apostolakis, 1988) refer to two biases that are particularly important in PRA: (1) the possibility of systematic overestimation or underestimation, and (2) overconfidence; that is people's tendency to give overly narrow confidence intervals that are not representative of their state-of-knowledge.

Considering for such biases in advance will enable the analyst improve the quality of the expert opinions when they are first elicited (Bier & Mosleh, 1988). There are several techniques that have been suggested for that purpose (Mosleh, Bier & Apostolakis, 1988):

- Calibration training, which involves providing feedback to the expert about his / her performance in past assessments
- Encouraging the expert to identify evidence that contradicts his / her initial opinion
- Problem decomposition; that is eliciting expert opinion on parts of the problem and then synthesizing the responses to formulate the forecast
- Aggregating the opinions of multiple experts, which tends to yield better results than relying on the judgment of a single expert; the fundamental principle that underlies the use of multiple experts is that a set of experts can provide more information than a single expert

Another approach for improving the quality of expert opinions is to adjust them after they have been elicited, in order to remove cognitive biases. Mosleh and Apostolakis (Mosleh & Apostolakis, 1984) have developed a mathematical Bayesian-based technique to correct for suspected biases once the elicitation process is already complete.

According to this technique, the likelihood function that an expert will provide an estimate x* for the probability of failure of a system, assuming that the true value of x is known, can be modeled by a lognormal distribution, i.e.,

$$L(x^*/x) = \frac{1}{\sqrt{2\pi}\sigma x^*}\exp\left\{-\frac{1}{2}\left[\frac{\ln x^* - (\ln x + \ln bx)}{\sigma}\right]^2\right\} \qquad (3\text{-}1)$$

where b is the bias factor and σ is the dispersion factor. The bias factor (b) measures the analyst's assessment of the tendency of the expert to underestimate ($b<1$) or overestimate ($b>1$) the true value of x. The dispersion factor (σ) is a direct measure of the expert's expertise. A small value of σ implies that the expert is likely to produce an estimate close to the true value. A large value of σ implies that the expert may provide a good estimate, but is quite likely to provide a bad one.

Previous attempts to use expert opinions for quantifying risk in space system PRAs have given risk estimates that reflected both underestimation and overconfidence. The most important contributors to that issue were: engineering overconfidence due to the lack of a formal elicitation method and training of the experts (Paté-Cornell & Dillon, 2001); problem composition, which was based on 'politically driven' optimistic assumptions (Paté-Cornell, 2002); and engineering tendency to rationalize risk signals e.g., precursor (near-miss) events[1] (Kaplan, 1992).

3.3 Bayesian Updating with Discounted Data

Several approaches have been developed for dealing with uncertain data. Siu and Kelly (Siu & Kelly, 1998) suggest that the basic framework for dealing with uncertain data has still to address the question of whether uncertain data should be dealt with within the likelihood function or the posterior distribution. The same authors also state: "additional fundamental work on the nature and treatment of uncertain data appears to be needed to resolve this question". In the following paragraphs, we will present an outline of the approaches that are considered more applicable to our problem.

3.3.1 Posterior Averaging Approach

The posterior averaging approach, described by Siu and Apostolakis (Siu & Apostolakis, 1984), Siu (Siu, 1990), and Martz et al (Martz et al, 1996), begins with the assignment of a subjective probability for every possible interpretation of the uncertain data, then it obtains a posterior distribution for the failure parameter φ that corresponds to each

[1] Near misses were not considered in our reference case as well. However, Kaplan uses a simple model to demonstrate that if precursor events were taken into consideration in the risk assessment of the Space Shuttle prior to the Challenger explosion, the resulting failure probability distribution would have been much more consistent with the evidence of the Challenger accident. Moreover, our literature review (Paté-Cornell & Dillon, 2001, Phimister, 2005) shows that consideration of near-miss events is important because it increases the *content* of operating evidence in a system's PRA, especially in the risk analysis of low-probability / high-consequence failures. Thus, estimates made without taking into account precursor events may not reflect all the system risk.

scenario, and finally, it weights these individual posteriors to obtain a combined average posterior distribution.

For example, in a failure-on-demand scenario, given that *r* failures have occurred in *n* demands, we assume that –for some reasons- the number of failures is uncertain, and is characterized by a discrete probability distribution *p(r)*. According to the posterior averaging approach, the posterior distribution for the failure parameter φ can be approximated as follows:

$$\pi_1(\phi / p(r), n) = \sum_{r=0}^{\infty} \pi_1(\phi / r, n) \cdot p(r) \qquad (3\text{-}2)$$

where $\pi_1(\varphi / r, n)$ is the posterior distribution that would be obtained from a conventional application of Bayes's theorem, given that *r* failures have occurred.

3.3.2 Weighted Likelihood Approach

In the weighted likelihood approach the data uncertainties are dealt within the likelihood function. According to this approach, in a failure-on-demand scenario, if the number of failures *r* is uncertain and characterized by a discrete probability distribution *p(r)*, the likelihood function is estimated as follows (Tan & Xi, 2003):

$$L(r / \phi, n) = \sum_{r=0}^{\infty} L(p(r) / \phi, n) \cdot p(r) \qquad (3\text{-}3)$$

Siu and Kelly (Siu & Kelly, 1998) give an example of an exponential likelihood in which the evidence (*E*) has the uncertain form: $E = \{a < t < b\}$. The likelihood function is then determined using the cumulative distribution function for the failure time:

$$L(E / \lambda) = \int_a^b \lambda e^{-\lambda t} dt = e^{-\lambda a} - e^{-\lambda b} \qquad (3\text{-}4)$$

where λ is the characteristic parameter of the exponential distribution.

3.3.3 Data Averaging Approach

The data averaging approach, described by Siu and Apostolakis (Siu & Apostolakis, 1986), and Siu and Kelly (Siu & Kelly, 1998), is an ad-hoc approach for handling the partial applicability of recorded failures, by discounting each failure according to the probability that it is applicable for future applications. In this approach, each failure is assigned a discounting fraction that represents the degree of applicability of the failure. The sum of the discounted failures is then used in the likelihood function, such as in the binomial likelihood or Poisson likelihood, to estimate the applicable failure probability.

The data averaging approach results in the following approximate posterior distribution:

$$\pi_1(\phi / p(r), n) = \pi_1(\phi / \bar{r}, n) \qquad (3\text{-}5)$$

where

$$\bar{r} = \sum_{r=0}^{\infty} r \cdot p(r) \qquad (3\text{-}6)$$

As discussed by Siu (Siu, 1990) and Siu and Kelly (1998), the "data averaging approach, while not strictly correct, yields results reasonably close to those obtained using the posterior averaging approach for a number of PRA problems". The problem with this Bayesian updating approach is that the resulting estimate may not be representative in cases where the spread of the probability distribution *p(r)* is wide, or when there are significant state-of-knowledge dependences (e.g., common cause failures) between data (Siu, 1990).

3.3.4 Likelihood in terms of Observation

The Likelihood in terms of Observation is a method developed by Tan and Xi (Tan & Xi, 2003), which, when applied in a beta-binomial model, gives a posterior of the general form

$$\pi_1(p/n,a,b) = \sum_{i=0}^{n} Q[i] \cdot Beta(p/a+i, b+n-i) \qquad (3\text{-}7)$$

where i is the number of applicable failures in n demands and the weights $Q[i]$ are calculated from the likelihood function with the use of the Classical error of a test's outcome.

Since this method involves modeling with Classical error (that is, accounting for the probability of observing a failure that does not exist, or not observing one that does exist), is probably more suitable for use in PRAs of testing equipment, health monitoring devices, or medical applications.

3.3.5 Overarching Method

This method, suggested by the Independent Peer Review Panel (IPRP) of our reference case (IPRP, 2005), is based on the assumption that, when failure record discounting is done, there could be a number of possible applicable failure records. However, instead of weighting each possible outcome, this method would simply bracket all possible outcomes by choosing the 5th percentile of the posterior distribution of the left-most outcome (the most optimistic outcome), and the 95th percentile of the right-most outcome (the most pessimistic outcome), and fit the theoretical distributions to these percentiles. Alternatively, instead of the 5th, the 50th percentile of the optimistic outcome can be chosen in a "modified" Overarching method.

The Overarching method is an ad hoc method, which simply attempts to cover all reasonable possible outcomes. It is the decision makers who decide which "average" distribution they should use based on the insights gained by plotting the posterior distributions resulting from the most optimistic and the most pessimistic interpretation of the data.

3.4 Relevant Cases

In this paragraph we refer to some cases found in the literature, on which several modeling approaches for the use of imprecise data are attempted. The posterior averaging and the weighted likelihood approaches are the Bayesian updating methods used in most cases, while the data averaging approach has proved to be a good approximation in the cases where it is evaluated. Also, Monte Carlo approaches are found convenient for use in more complicated cases (e.g. use of non-conjugate, multivariate models, dependence among parameters etc.). The problems addressed in each case are presented hereafter.

Siu and Apostolakis (Siu & Apostolakis, 1986) deal with the problem of assessing the detection time of fires in nuclear power plants. Due to the nature of fire detection (there is inherent uncertainty in the time elapsed between fire ignition and detection), the available evidence on the detection times of fires in nuclear power plants is rather sparse. A posterior averaging approach is developed for interpreting the uncertainty in available evidence in the probability distribution for detection time.

Siu (Siu, 1990) proposes a Monte Carlo method for the treatment of data uncertainties in the case of multi-parameter (multivariate) models, where dependences between the different parameters and between data are very likely to exist. The method is applied in three cases: analysis of auxiliary feedwater pump common-cause failure, analysis for Boiling Water Reactor safety relief valve failure on demand (discrete data), and nuclear power plant fire suppression rate analysis (continuous data). The evaluation of alternative approaches shows that, in the absence of significant state-of-knowledge dependences between data, the data-averaging approach yields reasonably accurate results much more quickly than the Monte Carlo approach.

Guarro et al (Guarro et al, 1995) deal with the problem of using in the Cassini mission PRA available evidence from heterogeneous sources, e.g., systems with similar design and missions. A Bayesian weighted likelihood approach is proposed, by which the evidence obtained from similar systems is assigned a weight that represents the

'degree of applicability', based on the level of similarity between the two systems. The Cassini Event likelihood is then assumed to be a weighted product of the two extreme applicability forms, i.e., zero applicability and total applicability, of the likelihood functions that make use of the non-Cassini evidence.

In another study, Guarro and Tomei (Guarro & Tomei, 2004) deal again with the problem of using data from heritage predecessors when estimating the probability of failure of new launch vehicles. A weighted likelihood approach is applied, which does not discount -this time- any failures in absolute terms, but permits the assignment of stronger relative weight to the record of more "similar" subsystem predecessors than the record of more dissimilar predecessors, in determining the successor's posterior probability of failure distribution.

Kaplan (Kaplan, 1992) deals with the reliability analysis of successive generations of helicopter equipment, where a new version has resulted from the implementation of several changes on a previous one. Engineering judgment (the 'expert information' method) is the only approach used for obtaining a consensus prior for the latest version, using as background the posterior of the previous version.

Martz et al (Martz et al, 1996) deals with the problem of uncertainty in the number of binomial demands and/or the number of failures, when assessing the probability of a failure on demand of a stand-by system. The posterior averaging Bayesian updating method is proposed for use in a beta-binomial conjugate model, which models the probability of failure of the high pressure coolant injection system in boiling water nuclear reactors. In this case, existing data were uncertain on whether they involved multiple or single injections, a distinction of special interest, since the system failing mechanisms appeared to be different for failure to reopen and for initial failure to open.

Martz and Hamada (Martz & Hamada, 2003) present a Bayesian Markov chain Monte Carlo method, which can be used to quantify the effects of uncertainty in the operating time (t) and/or the number of event occurrences (x) when estimating the event

occurrence rate in a Poisson model. The advantages of this method over the posterior averaging, when tested in the case of a high pressure coolant injection system in boiling water nuclear reactors, are that non-conjugate priors can be used and evaluated, while calculations are easier to be executed when both the operating time and the number of event occurrences are uncertain.

Pietzsch et al (Pietzsch et al, 2004) present an approach to early assessment of new medical technologies, which makes use of available information from different data sources. More specifically, prior distributions from several sources (similar items, animal testing, expert opinion, etc.) are assigned different weights, based on the perceived quality and relative preference of each source, so that the aggregated prior for the system of interest is obtained by linear pooling. Furthermore, when appropriate, discounting takes place directly on a source's prior distribution by increasing the variance and changing the moments.

Guikema and Paté-Cornell (Guikema & Paté-Cornell, 2005) work on the treatment of 'infancy problems' in the reliability analysis of space launch systems. More specifically, by using both Bayesian probability and frequentist statistics, they analyze the probability of failure of launch vehicles in their first five launches. An important finding in the paper is that, the reliability analysis of subsequent generations within the same vehicle family revealed that there is not any significant improvement in the reliability of more recent generations of launch vehicles over the previous generations.

(This page intentionally left blank)

4. Proposed Approach

The objective in this Chapter is to propose an appropriate Bayesian updating approach for calculating the probability of catastrophic failure[1] for the latest Configuration C of our system, by using the available operational information from the heritage configurations A and B, while considering the changes in design and test / inspection procedures that have been implemented in the system over the course of the Space Shuttle program.

In Sections 4.1 to 4.3, we present our insights based on the problem structure, sensitivity studies, and NASA's approach and results; in Section 4.4, we develop our proposed approach and present our results; in Section 4.5, we deal with the benign failure assessment case; and, in Section 4.6, we present an alternative approach to our problem.

4.1 Bayesian Parameter Estimation Model

a) General

Our approach is in accordance with the basic methodology for "Bayesian parameter estimation in probabilistic risk assessment", as presented in Siu and Kelly (Siu & Kelly, 1998). In its general form, Bayesian parameter estimation has four steps:

(1) Identification of the parameter(s) to be estimated

(2) Development of the prior distribution that appropriately quantifies the analyst's state of knowledge about the unknown parameter(s)

(3) Collection of evidence and construction of an appropriate likelihood function

(4) Derivation of the posterior distribution using Bayes' theorem

[1] Due to some very interesting differences in our results between catastrophic and benign failures, our analysis in the following sections is based only on catastrophic failure data. The benign failure analysis is included as a separate case study in Section 4.5

The Bayesian parameter we want to estimate in this section is the probability of failure (catastrophic or benign) on demand of a space system that operates in NASA's Space Shuttle. The basic characteristics that should be considered in our model are described briefly below:

- Over the course of the Space Shuttle, the system has evolved with three configurations A, B, and C; that is, there are two heritage predecessors (A and B).
- There is a very high commonality (80%) in the component lists among the three configurations.
- Design changes have been made both at the system level (to improve the hardware and operating environment conditions) and at the component level (as responses to failures).

Our database includes the aggregated operational data per system version (Table 2-2), aggregated discounted failure data (Table 2-3), brief description of the individual failures (Appendix A), and the relevant corrective actions and discounted factors initiated and assigned by the Agency (Appendix B).

In the following sections we describe our modeling decisions as far as the basic methodology for Bayesian parameter estimation is concerned, in accordance with the general characteristics of our problem and the need to account for failure record discounting.

b) Consideration of Heritage Configurations

There are several modeling approaches for taking into account the operational experience of previous configurations (heritage systems) in PRA. One is the use of a multi-step Bayesian updating method, as NASA did, according to which the posterior distribution function of each predecessor is the prior distribution function for the successor configuration.

A 'weighted prior' is a different approach that could be applied in our case, according to which, each predecessor is modeled separately; their posteriors are then weighted based on the perceived quality of the data and the level of similarity between versions; and then, they are used for the estimation of the last configuration's prior by linear pooling (Pietzsch et al, 2004; Guarro et al, 1995; Guarro & Tomei, 2004). The prior of the latest configuration can also be obtained by engineering judgment using as background the posteriors of the previous versions (Kaplan, 1992). This approach has the same basic concept as the overarching method presented in Section 3.3.5.

There are also some more rigorous, Bayesian-based approaches (two-stage Bayes, maximum likelihood, and maximum entropy), for using information from systems of the same family, to develop an informative prior distribution for the assessed system (Siu & Kelly, 1998). In addition to the computational difficulties of these methods, their sophistication is probably not needed in our problem, which is focused on the impact of failure record discounting in PRA results.

We believe that NASA's approach is reasonable since there are not any significant differences, especially between Configurations B and C, as far as the involved technology is concerned. Thus, a multi-step Bayesian approach is also used in our model. Alternatively, we test a 'weighted prior' approach for sensitivity analysis reasons explained in the following sections.

c) Construction of Appropriate Likelihood and Prior Distribution

The basic question to be answered for the construction of an appropriate likelihood function is the definition of the underlying process that generates the data to be used in the parameter estimation process. In our case, the use of the binomial distribution is reasonable since events are generated on a demand basis (missions that either succeed or fail) and they can be considered independent. The absence of aging, which is another requirement for the use of the binomial model, is not an important consideration since most time-sensitive units are replaced on a regular basis.

Our likelihood function is derived from three modeling assumptions: (a) the demand failure rates p characterize a Bernoulli process for each configuration; (b) the distribution for our model's parameter in each configuration is beta; (c) the data sets from each configuration are conditionally independent; and, (d) the numerical value of p may be different from one configuration to the next.

The use of a conjugate prior distribution, that is, a distribution that has the property that the posterior distribution also belongs to the same family as the prior, is an option, which, for the problem at hand, naturally leads to the use of a beta-binomial model.

In the absence of any catastrophic failures in configuration A, NASA uses a non-informative prior, Jeffreys' U-shaped Beta (0.5, 0.5), which is updated with the evidence observed in this configuration, in order to end up with a prior distribution for configuration B. Non-informative priors are distributions that are constructed to represent a vague state of knowledge (Siu & Kelly, 1998). The use of a non-informative prior distribution allows the posterior distribution to be formed almost exclusively by the data. The sensitivity of the results to this assumption will be discussed later in this Chapter.

4.2 Failure Record Discounting

4.2.1 Guiding Rules and Cautions

The fundamental quantity in the Bayesian framework is the *evidence*. Very often the evidence coincides with the statistical data. In our case, however, the evidence includes, in addition to the data, the fact that design or test / inspection changes have been made, the insights from tests and reliability analyses, etc., and this information cannot be ignored. This problem is not straightforward in the sense that the statistical data do not coincide with the evidence.

This deviation from the usual Bayesian application can be addressed with the contribution of experts, who are called upon to provide their subjective probabilities about the likelihood of some events, based on their total body of knowledge. In the literature review, we have found several treatments of expert opinions in similar cases. More specifically, they may

- Provide a point estimate of the model's parameter, which is then treated as Bayesian evidence in a way similar to the methodology described in Appendix E (Mosleh & Apostolakis, 1984; Siu & Kelly, 1998)
- Provide weights on distributions obtained from similar systems, to be used in the development of a generic prior for the system under assessment (Guarro et al, 1995; Guarro & Tomei, 2004)
- Directly provide an estimation of the prior using as background the posteriors of the previous versions (Kaplan, 1992)

In none of the reviewed cases in the literature has engineering judgment been used for discounting failure records; that is, reducing the number of failures that exist in actuarial data. Our concern is that, when engineering judgment is used for a direct assessment of failure applicability based on design or procedure improvements, there is room for **biases** in the elicitation and use of subjective information. These biases will tend to overestimate the impact of these improvements.

Special caution should be exercised when assessing failure record applicability factors in order to avoid generating overly narrow failure distributions implying a strong state of knowledge. This is the case when failure data are discounted while success data are not. The reason is that, in Bayesian analysis, the impact of the evidence {n=10, N=1000} on the prior is much more significant than that of the evidence {n=1, N=100}; thus, if the number of failures in the first evidence is discounted to n=2, this will result in an overly narrow posterior distribution. The implication for the failure record discounting approach is that it may need some refinement to account for some kind of

"success" discounting in the case of extensive failure record discounting. This case is separately investigated in Section 4.6 (Alternative Approaches).

4.2.2 Results of Existing Discounting Approach

As we have mentioned, failure record discounting is an activity that is based on engineering judgment; thus, the most formal way to evaluate the appropriateness of the relevant method is by executing sensitivity studies and displaying the effects of the discounting factors used. Our observations concern some points where we believe that optimistic estimations were obtained. Since these observations are not based on a rigorous knowledge of the system under assessment, we decided to present them in Appendix D, and focus our analysis in this Chapter on the results of the existing method.

As described in Chapter 2, there are two kinds of subjective intervention in NASA's approach, failure record discounting and environmental and hardware adjustment. In failure record discounting, the failure records of each configuration are discounted to account for the component level design or test / inspection procedure changes; and then, this new form of evidence is used in the Bayesian updating method for posterior distribution estimation. In environmental and hardware adjustment, the posteriors obtained for configuration B are directly adjusted to account for configuration C relevant, system-wide changes, which have improved the environmental and hardware operating conditions.

In Section 2.1, we saw that failure record discounting is done by assigning a probability γ to the different failure records, which represents the analyst's state of knowledge that a specific action (redesign or procedure change) will prevent the reappearance of a specific failure. The discounting was done extensively and, as a result, the 8 catastrophic failures observed in Configuration B were discounted to 1.15, while the 4 catastrophic failures observed in Configuration C were discounted to 0.2.

As far as environmental and hardware adjustments are concerned, these were done by dividing the configuration B posterior by an environmental factor having a *Uniform(1.520, 1.858)* distribution representing improvements in operating environment, and then, the resulted distribution by a hardware factor having another *Uniform(1.1, 1.3)* distribution representing hardware improvements.

The combined impact of these two interventions on the results is shown in Fig.4-1 where the catastrophic failure prior distributions for Configuration C, with and without discounting and adjustment are compared.

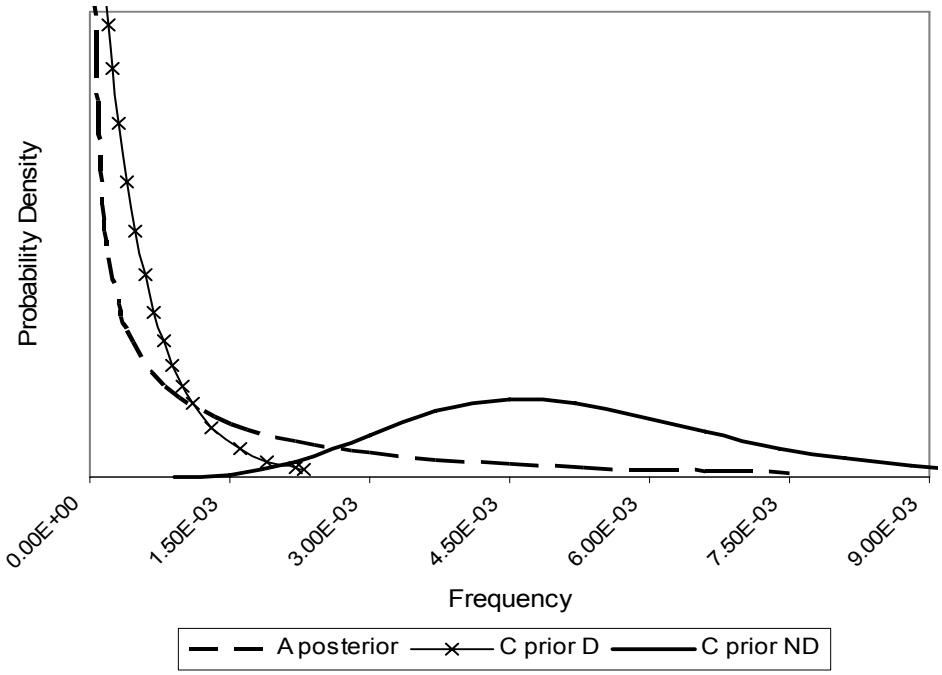

Figure 4-1. Configuration C catastrophic failure prior distribution with discounting (D) vs. without discounting (ND)

We can note that the prior obtained with discounting[1] is overly narrow and significantly shifted to the left in comparison with the one obtained without. If we consider that the *actual* evidence collected in configuration C is 4 failures in 480 demands (i.e., resulting

[1] Failure applicability factor γ=0.14375 in Configuration B failure records; environmental adjustment *Uniform(1.520, 1.858)*; hardware adjustment factor *Uniform(1.1, 1.3)*

in a point estimate of 8.33E-03), we realize that the likelihood fits marginally within the 90% failure bounds of the undiscounted prior, while it is more than 5 times greater than the 95th percentile of the discounted one (given in Table 2-4).

Moreover, the form of the Configuration C prior distribution obtained with discounting is a reversed-J shaped. This resulted from the update of the non-informative U-shaped starting prior distribution with over-discounted data; in other words, the number of discounted failures was not substantial enough to change the shape of the starting prior distribution.

Thus, based on the available operating experience in Configuration C, we can conclude that the assessment of this configuration's prior distribution has been optimistic and overconfident[1].

The bottom line is that the applied discounting factors resulted in risk estimations that are more than an order of magnitude lower than the likelihood. Under these conditions, the ultimate question is "how confident can we feel that the posterior distribution obtained for our system (Configuration C) is representative of its risk in order to base our decisions on this?" From a purely engineering perspective, our point is that risk reductions of that magnitude presume a major engineering achievement, which should be supported not only by engineering judgment but also by strong evidence in terms of operational experience.

4.2.3 Implications for the Expert Opinion Elicitation Process

The experience of NASA scientists on their areas of expertise is unquestionable; however, biased subjective estimations have been observed in many PRA applications, and are

[1] On the other hand, as we will see in Section 4.6, the prior distribution obtained for the probability of benign failure for configuration C with the same discounting approach is very consistent with the *actual* evidence. This observation highlights how complicate is the problem of failure partial applicability; however, it does not change our approach to the issue.

usually caused by the overconfidence that people often have, especially those who are experts in their technical areas.

The main reason for the observed over-discounting in our case, is that the guideline used for that purpose (Table 2-1) assigns very large discounting factors (and also small ranges), which probably are justifiable only when accompanied by strong operational evidence at the system level and not only from independent component testing. Furthermore, since space system PRAs are in an early development phase, experts in this field may lack the required PRA experience regarding the expression of their judgment in terms of probabilities.

We are not in the position to propose specific recommendations as far as expert opinion elicitation is concerned, within the context of this thesis. Our review on expert opinion elicitation that is included in Chapter 3 presents general guidelines, cautions, and techniques that can help improve the quality of engineering judgment. Our point is that the advantages of a sound methodology for expert opinion elicitation with respect to the approach of complicate PRA problems should not be overlooked.

4.3 Bayesian Updating

4.3.1 Observations on Weighted Average Likelihood Method

The Weighted Average Likelihood method used by NASA is essentially the same as the Posterior Averaging Approach given in (3-2), where $p(r)$ is replaced by P_x from (2-3), which is the Binomial probability that x out of the n observed failures are applicable with parameter the failure record applicability factor γ.

In general, there are two levels of epistemic uncertainty in the posterior averaging approach. First, there is uncertainty about the discounting factor γ, which defines different scenarios about the number of the observed failures that are applicable in the analysis. Second, there is uncertainty about the likelihood of each scenario, which is

represented in the assignment of different weights to the above scenarios to be used in the posterior averaging approach.

According to NASA's approach, the weights (P_x) used in the posterior averaging Bayesian updating approach are directly related to the average failure record applicability factor (γ) in an aleatory relation (binomial probability). However, both the failure applicability factor (γ) and the weighting parameter (P_x) are not aleatory probabilities; in other words, the probability that a specific failure or a failure scenario be applicable or not cannot be assessed in a random manner like the toss of a coin.

The above problem can be overcome in one of two possible ways:

- Use of expert opinion to define the weights (P_x) for each possible scenario.
- Use of the data averaging approach given in Section 3.3.3, which in our case, instead of a weighted posterior distribution, it uses directly the number of discounted failures as evidence in Bayesian updating.

The literature review (Siu & Kelly, 1998; Siu, 1990; Martz et al, 1996) shows that, in the absence of significant state-of-knowledge dependences between failures (i.e., common cause failures), the data averaging approach yields reasonably accurate results. The comparison of the results obtained by the weighted average likelihood method versus the data averaging one, with the use of the same applicability factors estimated by NASA, is shown in Fig.4-2.

The results of the data averaging method were obtained by using in (3-5) the average number of applicable failures as given by (3-6) with the use of the individual failure discounting factors from Appendix B. The average number of failures used in (3-5) for Configurations B and C were 1.15 and 0.2, respectively.

As the results in Table 4-1 and the graph in Fig.4-2 suggest, the data averaging approximation is quite reasonable in our case, while it has the advantages that the need

for generating weights to be used in the posterior averaging approach is eliminated[1], the calculations required are more straightforward, and the resulting distributions have a mode. Thus, the data averaging approach is used in our modeling approaches hereafter.

	Weighted Average Likelihood		**Data Averaging**	
	B posterior	**C Posterior**	**B posterior**	**C posterior**
Mean	1.015E-03	4.912E-04	1.015E-03	4.895E-04
St. Dev.	1.008E-03	4.838E-04	7.898E-04	3.729E-04
5th	5.370E-05	9.030E-05	1.385E-04	7.108E-05
50th	7.076E-04	3.499E-04	8.196E-04	3.987E-04
95th	3.028E-03	1.356E-03	2.561E-03	1.218E-03

Table 4-1. Weighted average likelihood vs. data averaging method results

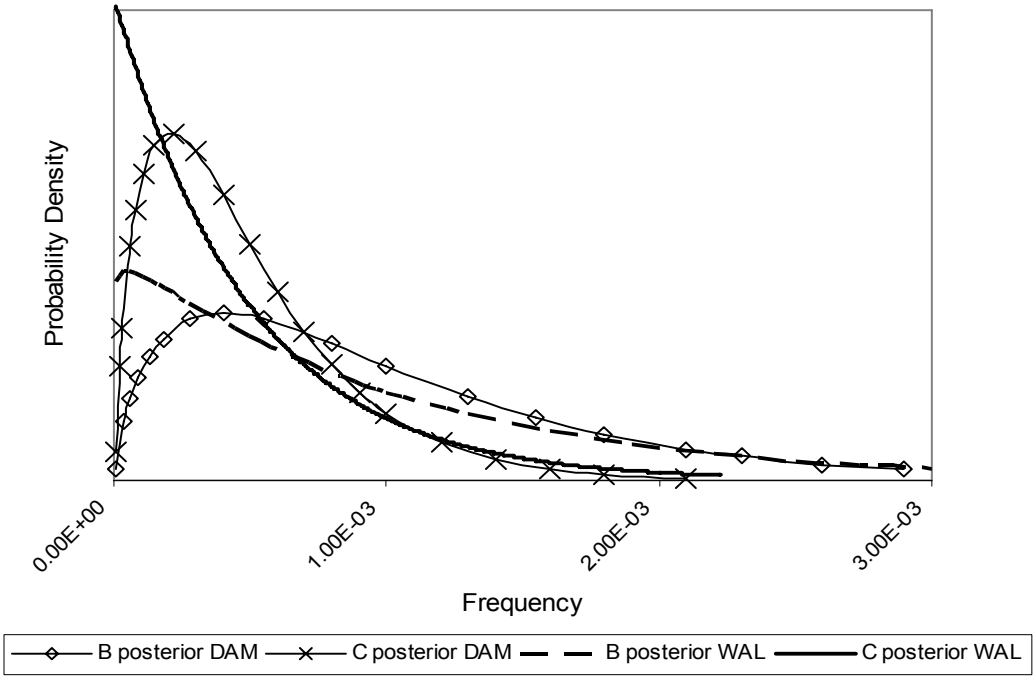

Figure 4-2. Weighted average likelihood (WAL) vs. data averaging method (DAM) results

[1] This happens because it is the data (failures) that are weighted according to (3-5), in which *p(r)* for each individual failure is equal to its applicability factors (γ). The resulting average number of failures is then used as evidence in (3-4) for computing the posterior distribution.

4.3.2 Observations on Multi-step Bayesian Updating Approach

We have seen in Section 4.2.2 that the assignment of liberal discounting factors has resulted in very optimistic results, which are significantly different from the ones obtained without discounting. In this section, we examine how our assumption to consider the evidence from the heritage configurations A and B in a multi-step Bayesian model, may have contributed to the above results.

The multi-step Bayesian updating method is a standard method used in Bayesian parameter estimation analysis, when the posterior distribution of the parameter of interest is updated with the new evidence to give an updated posterior. The selection of this approach in our application was based on the assumption that the different configurations have a very high level of similarity (80%); thus, the evidence of a heritage configuration is assumed to be directly applicable to the successor one.

We believe that, when discounting is done extensively, its interpretation in the results is such that the high level of similarity assumption may not be valid. In other words, while some adjustment on risk estimates is justifiable, a resulting estimate that is an order of magnitude reduced compared to the one obtained without discounting cannot refer to a *similar* system. Thus, basing our argument at the logical space and only, we believe that a multi-step Bayesian updating approach is not appropriate in the case of extensive discounting.

In order to realize the impact of a multi-step Bayesian updating approach on risk estimates we just follow the process: First, our system's generic prior (A posterior) distribution is updated with configuration B discounted (by 85%) failures; the resulting posterior (B posterior) is adjusted by 40% to account for the improvements in operating environment; the resulting distribution is then discounted by another 17% to account for system-level hardware improvements; and then, the resulting distribution, which has been sequentially discounted and adjusted several times, is configuration's C prior (which is also updated with data discounted by 95%).

In the literature review we saw that priors (and evidence) from heritage systems are not usually used directly in a multi-stage Bayesian updating manner, but instead they contribute to the estimation process of the latest version prior, by either a weighted average relationship on the basis of their level of 'similarity' (Guarro et al, 1995; Guarro & Tomei, 2004; Pietzsch et al, 2004), or by providing the background for eliciting expert opinions (Kaplan, 1992).

In a similar manner, we decided to test a "weighted prior" approach in order to get a better insight in the impact of multi-step Bayesian updating on Configuration C prior risk estimations, by using the same discounting and adjustment factors as NASA. The difference is that discounting and adjustment do not happen in a sequential order, but instead their effect is weighted through a weighted averaging approach. More specifically, a Configuration C prior distribution obtained with full failure record discounting (but without environmental and hardware adjustment), and a second one obtained with environmental and hardware adjustment (but without failure record discounting) were weighted to generate a combined prior distribution for Configuration C as follows:

$$\pi_0(\theta) = w_{fd} \text{ x } \pi_{fd}(\theta) + w_{ad} \text{ x } \pi_{ad}(\theta) \qquad (4\text{-}1)$$

where $[w_{fd}, \pi_{fd}(\theta)]$ and $[w_{ad}, \pi_{ad}(\theta)]$ are the weights and prior distributions for Configuration C for the case when only failure record discounting and only environmental and hardware adjustment is done, respectively.

The resulting prior with equal weights (0.5, 0.5) is shown in Fig.4-3, while a sensitivity study of the application of different weights is provided in Table 4-2. The results were obtained with moment matching (Appendix C).

(This page intentionally left blank)

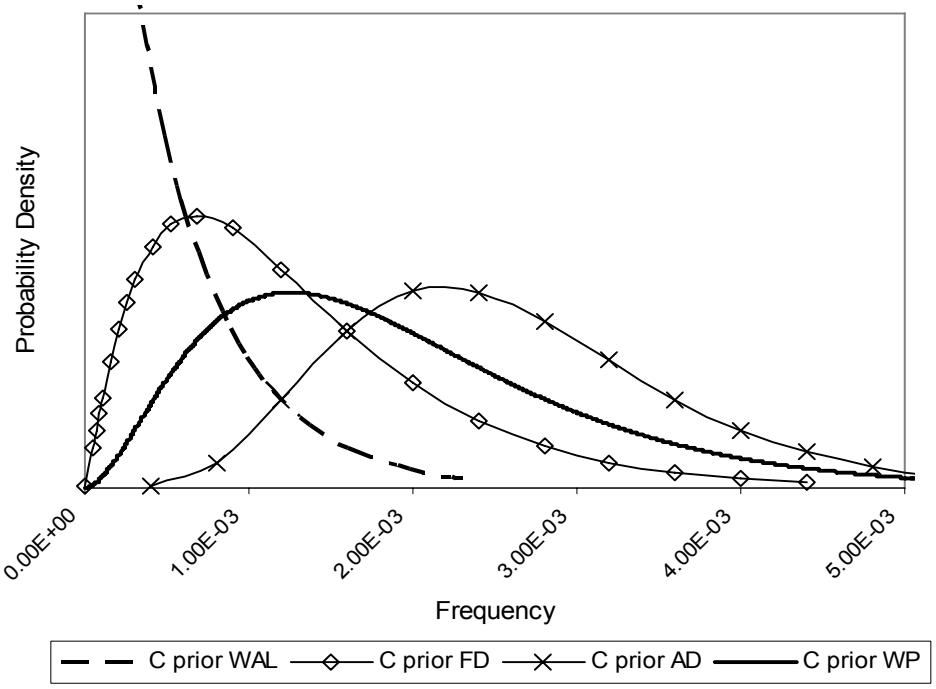

Figure 4-3. Configuration C prior distribution with weighted prior method (WP) vs. weighted average likelihood (WAL)

	Configuration C Prior Characteristic Values						
	Full Discounting	**No Discounting**	**"Weighted Prior" Approach**				
(w_{FD}, w_{AD})	n/a	n/a	(0.2, 0.8)	(0.4, 0.6)	(0.5, 0.5)	(0.6, 0.4)	(0.8, 0.2)
Mean	5.01E-04	5.23E-03	2.32E-03	2.07E-03	1.94E-03	1.81E-03	1.55E-03
St. Dev.	5.09E-04	1.79E-03	1.11E-03	1.15E-03	1.15E-03	1.13E-03	1.05E-03
5th	2.36E-05	2.67E-03	8.40E-04	5.97E-04	5.02E-04	4.22E-04	3.04E-04
50th	3.43E-04	5.03E-03	2.15E-03	1.86E-03	1.71E-03	1.58E-03	1.32E-03
95th	1.52E-03	8.48E-03	4.40E-03	4.25E-03	4.13E-03	3.98E-03	3.58E-03

Table 4-2. Sensitivity analysis of the "Weighted Prior" approach

We draw two conclusions from Fig.4-3 and the sensitivity study in Table 4-2:

- The modeling decision to use a multi-step Bayesian updating approach 'magnifies' the impact of our discounting approach on the resulting risk estimates, because the prior distribution obtained that way is by far more narrow than any of the distributions obtained with either failure record discounting or environmental and hardware adjustment.
- The 'discounting' impact of failure record discounting on the risk results is more significant than that of environmental and hardware adjustment in terms of its tendency to shift the Configuration C combined prior distribution to the left.

4.3.3 Sensitivity Analyses of Model Assumptions

a) Non-informative Prior

Bayesian updates can provide useful results in data sparse situations, but this is very much dependent on the accuracy of the prior distribution. If data are truly sparse the likelihood function may be very diffuse, and in these situations the prior distribution function (even a non-informative one) exerts considerable influence on the shape of the posterior distribution. This is the case when we have little data and analyst's prior beliefs are relatively vague (Phimister, 2005).

At the beginning of the analysis we used Jeffrey's non-informative Beta(0.5,0.5) prior distribution. To test our assumption we calculated the configuration C priors obtained from different starting non-informative priors, such as Uniform Beta(1,1) and a non-dominating Beta(0.01, 0.01). The resulting Configuration C prior distributions obtained with and without discounting and adjustment are presented in Table 4-3.

	Configuration C Prior Characteristic Values					
	No Discounting			**With Discounting**		
Starting Beta (α, β)	**(1,1)**	**(0.01,0.01)**	**(0.5,0.5)**	**(1,1)**	**(0.01,0.01)**	**(0.5,0.5)**
Mean	5.54E-03	4.93E-03	5.23E-03	6.52E-04	3.52E-04	5.01E-04
St. Dev.	1.84E-03	1.74E-03	1.79E-03	4.66E-04	3.38E-04	4.06E-04
5th	2.89E-03	2.46E-03	2.67E-03	1.13E-04	2.20E-05	6.02E-05
50th	5.33E-03	4.73E-03	5.03E-03	5.45E-04	2.52E-04	3.97E-04
95th	8.87E-03	8.09E-03	8.48E-03	1.56E-03	1.03E-03	1.30E-03

Table 4-3. Sensitivity analysis of different starting non-informative Beta

As we can notice in the sensitivity analysis results of Table 4-3, the impact of different starting non-informative Beta distributions is minimal in the case when no discounting is done. On the other hand, when discounting is done, the results obtained from different starting distributions are not so similar. The implication for our study is that when discounting is done extensively, even the use of a *non-informative* prior may carry some *information* in risk results, thus, additional justification is needed on the selection of an appropriate non-informative starting prior.

b) Failure Record Discounting Factor

In this sensitivity analysis, we test the effect of the assignment of different applicability factors (γ) on the risk estimates, without adjusting for environmental and hardware operating conditions. The results of our analysis are presented in Table 4-4.

	Configuration C Posterior Characteristic Values						
Applicability Factor (γ)	**1.0**	**0.8**	**0.6**	**0.5**	**0.4**	**0.3**	**0.2**
Mean	5.94E-03	4.80E-03	3.66E-03	3.09E-03	2.52E-03	1.95E-03	1.38E-03
St. Dev.	1.67E-03	1.51E-03	1.32E-03	1.21E-03	1.09E-03	9.61E-04	8.08E-04
5th	3.48E-03	2.62E-03	1.79E-03	1.40E-03	1.03E-03	6.77E-04	3.65E-04
50th	5.78E-03	4.64E-03	3.50E-03	2.93E-03	2.36E-03	1.79E-03	1.22E-03
95th	8.93E-03	7.51E-03	6.06E-03	5.31E-03	4.54E-03	3.75E-03	2.92E-03

Table 4-4. Sensitivity analysis of different failure record applicability factors (no environmental and hardware operating conditions adjustment)

The effect of the applicability factor (γ) on the risk results is getting higher as we move to lower values, i.e., from applicability factor 0.8 to 0.7 we have a decrease by 12% in the average failure parameter, while from 0.3 to 0.2 we have a reduction by 30%.

Another very interesting observation is that the impact of any applicability factor on risk results, compared to the results obtained without discounting ($\gamma=1$), is a reduction of the magnitude of the involved discounting; that is, about 20% for $\gamma=0.8$, 60% for $\gamma=0.4$, 80% for $\gamma=0.2$, and so on. This means that failure record discounting has a *direct* impact on risk results. It is reminded to the reader that NASA has estimated a failure applicability factor of 0.14 for configuration B and 0.05 for configuration C.

c) Operating Conditions Adjustment

In this analysis, we test the impact of different combined environmental and hardware adjustments on the risk results, without failure record discounting. The results of our analysis are presented in Table 4-5.

	Configuration C Prior Characteristic Values			
Combined Adjustment Uniform (a, b)	**(1.16, 1.26)**	**(1.38, 1.50)**	**(1.42, 1.96)**	**NASA (1.89, 2.17)**
Mean	4.32E-03	3.63E-03	3.10E-03	2.58E-03
St. Dev.	1.53E-03	1.30E-03	1.27E-03	1.01E-03
5th	2.14E-03	1.79E-03	1.34E-03	1.17E-03
50th	4.14E-03	3.48E-03	2.92E-03	2.45E-03
95th	7.12E-03	6.00E-03	5.43E-03	4.43E-03

Table 4-5. Sensitivity analysis of different combined uniform adjustments (no failure record discounting)

We can notice in Table 4-5 that the impact of the different combined environmental and hardware adjustment on the risk results is smoother than that of the discounted failure records.

4.4 Proposed Approach

4.4.1 Discussion

A truly rigorous Bayesian analysis requires not only normative expertise (e.g. good knowledge of probability and statistics), but also substantive expertise, i.e., a good understanding of the particular issue being investigated. We do not have the technical background to evaluate how much NASA's actions have improved system's safety or how reasonable are the discounting factors assigned to represent these improvements. However, we believe that, since a Bayesian parameter estimation approach is used, a set of basic principles must be followed.

In the Bayesian parameter estimation framework, data have a dominant role in the process of removing uncertainty or increasing our state of knowledge about the modeled parameter. This is the reason why non-informative priors are often used in order for the posterior distributions to be dominated by the actual evidence.

As we have mentioned, failure record discounting is an activity that should increase our uncertainty about the failure parameter. Furthermore, in the previous sections, we have seen that the application of excessive discounting factors does not only have a significant impact on the risk results but it can also violate our modeling assumptions.

It is our thesis that, in problems of this nature (sparse failures and little operating experience), failure record discounting is insightful if it is used as a means for providing an optimistic interpretation of the data, when the collected failures do not have direct applicability in PRA due to system improvements. This optimistic interpretation should somehow be combined and presented along with a distribution obtained with actuarial data (pessimistic) to provide a basis for risk-informed decision making. The aggregation of the two input distributions can be performed subjectively by executing a series of sensitivity studies using a weighted averaging approach with several weights.

More specifically, the resulting distribution has the following form:

$$\pi(\theta) = w_D \text{ x } \pi_D(\theta) + w_{ND} \text{ x } \pi_{ND}(\theta) \qquad (4\text{-}2)$$

where $[w_D, \pi_D(\theta)]$ and $[w_{ND}, \pi_{ND}(\theta)]$ are the weights and the distributions that represent discounting and no discounting, respectively.

Our approach is in accordance with the spirit of the "overarching method" presented in Section 3.4.5, which argues that it is the decision makers who decide which "average" distribution they should use based on the insights gained by plotting the posterior distributions resulting from the most optimistic and the most pessimistic interpretation of the data.

The question to be answered now is "what level of discounting is representative of an optimistic outcome?" In the absence of substantive knowledge of the system, we are not in a position to evaluate an appropriate level of discounting. However, we would like to demonstrate how the proposed approach works, and also provide some alternative results within the context of this thesis.

In the light of our analysis, observations, and literature review[1], we think that the application of an overall failure applicability factor of $\gamma=0.6$, and an adjustment representative of the improvement in operating environment approximated by a *Uniform(1.520, 1.858)*, can be representative of an optimistic outcome[2] that does not violate the model assumptions.

4.4.2 Results

Our results were obtained by the process shown at the flowchart of Fig.4-4. A multi-step Bayesian updating approach was used for obtaining Configuration C prior distribution with (D) and without discounting and adjustment (ND). The combined distribution for Configuration C prior was then obtained from (4-2) and a theoretical distribution (beta) was fitted on it with moment matching (Appendix C). The fitted Configuration C prior distribution was updated with Configuration C discounted and actual data to derive Configuration C discounted (D) and undiscounted (ND) posterior distributions, respectively. The combined distribution for Configuration C posterior was then derived from (4-2) and a beta distribution was fitted on it with moment matching.

The results obtained with the use of different weights (w_D, w_{ND}) for the optimistic (D) and the pessimistic (ND) outcomes are given in Table 4-6.

[1] An important finding in the reliability analysis of subsequent generations within the same vehicle family of several launch vehicles (Guikema & Paté-Cornell, 2005) reveals that there is not any statistically significant improvement in the reliability of more recent generations of launch vehicles over the previous generations.

[2] The sensitivity studies show that the combined impact of these factors on risk estimates is a reduction by about 65%, which indeed is a very optimistic outcome.

(This page intentionally left blank)

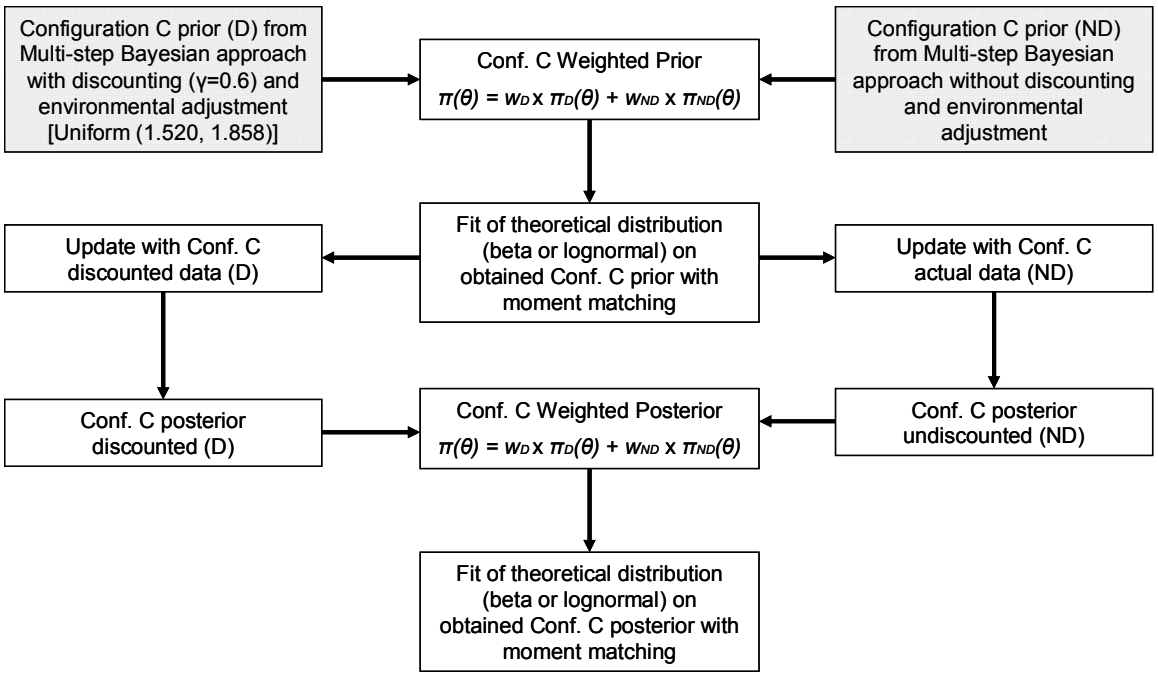

Figure 4-4. Flowchart for Configuration C prior and posterior distributions generation with proposed approach

	Configuration C Posterior Characteristic Values						
	NASA	**No Discounting**	**Proposed Approach**				
(w_D, w_{ND})	n/a	n/a	(0.8, 0.2)	(0.6, 0.4)	(0.5, 0.5)	(0.4, 0.6)	(0.2, 0.8)
Mean	4.91E-04	5.94E-03	3.61E-03	4.28E-03	4.55E-03	4.79E-03	5.24E-03
St. Dev.	4.83E-04	1.67E-03	1.64E-03	1.84E-03	1.86E-03	1.85E-03	1.74E-03
5th	9.03E-05	3.48E-03	1.39E-03	1.76E-03	1.97E-03	2.20E-03	2.74E-03
50th	3.50E-04	5.78E-03	3.36E-03	4.02E-03	4.30E-03	4.56E-03	5.05E-03
95th	1.36E-03	8.93E-03	6.66E-03	7.68E-03	7.97E-03	8.18E-03	8.40E-03

Table 4-6. Sensitivity analysis of Configuration C posterior distribution on the use of different weights in proposed approach

We can notice that the resulting distributions have mean values of about an order of magnitude greater than NASA's estimate. These distributions are also significantly wider. In comparison to the estimates obtained with no discounting, we can see that the resulting distributions have means that can be from quite close to 60% reduced[1]. The most important observation is that the resulting distributions are wider than the distribution obtained without discounting, an observation that supports our claim that failure record discounting is an activity that should increase our uncertainty about the failure parameter.

The interpretation of the results as shown in Table 4-6 has several advantages. First, it displays the uncertainties that are inherent in the case of failure record partial applicability[2]. Most importantly though, it provides a better basis for the decision maker to make a decision based on his / her risk attitude. For example, a risk adverse decision maker would base his / her decision on the Configuration C posterior distribution that assigns a larger weight on the undiscounted outcome.

The Configuration C prior and posterior distributions obtained from the use of our approach with the assignment of equal weights (w_D=0.5, w_{ND}=0.5) on the optimistic and the pessimistic elements have been plotted in Fig.4-5 and Fig.4-6, respectively. We used equal weights to express our vague state of knowledge about the representativeness of the different outcomes.

[1] Extreme case (w_D=1, w_{AD}=0) that results to mean 2.54E-03, not shown in the Table 4-5.

[2] Assuming a consensus of engineering judgment on the optimistic outcome

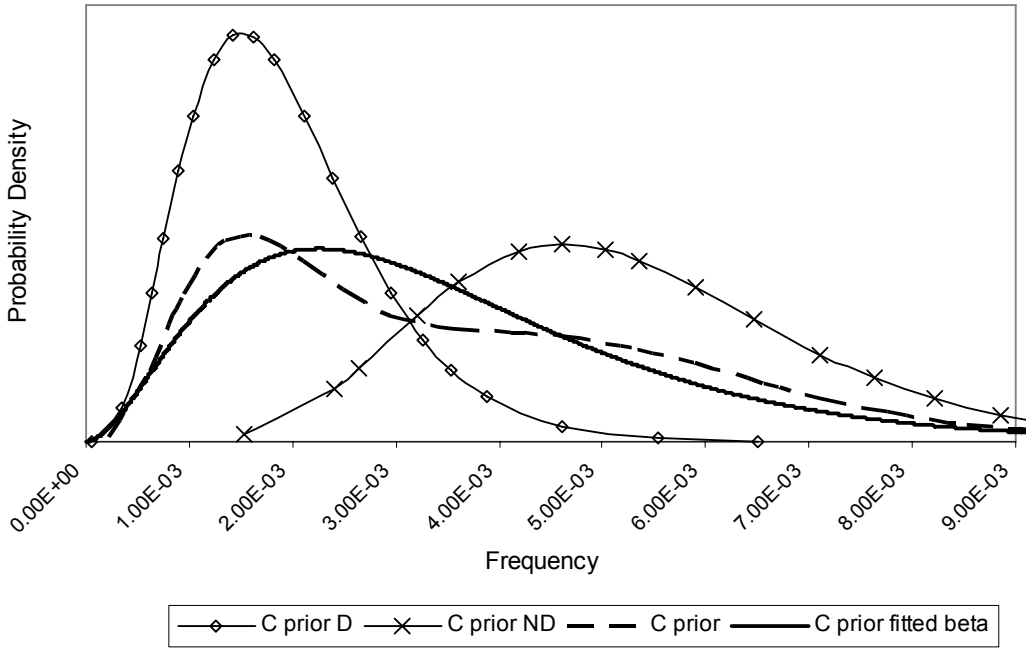

Figure 4-5. Configuration C prior distribution with proposed approach

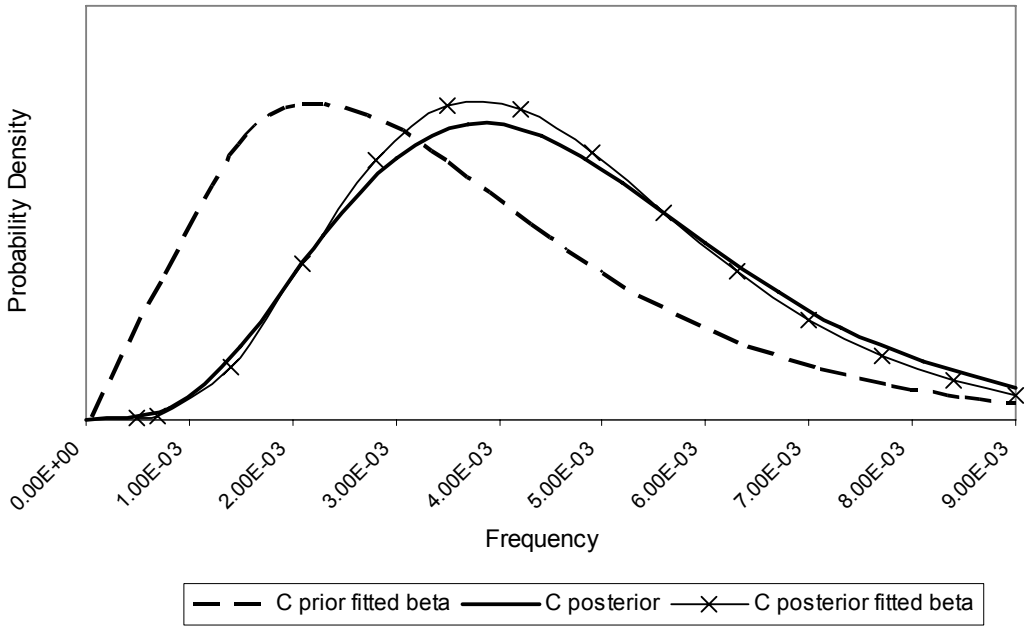

Figure 4-6. Configuration C posterior distribution with proposed approach

An insight from Fig.4-5 and Fig.4-6 is that a theoretical distribution (beta) is fitted rather well on the resulting weighted prior (and posterior) distribution for Configuration C. Second, the Configuration C posterior distribution is shifted significantly to the right in response to the observed failures in this configuration (*n=4, N=480*). We remind the reader that the Configuration C prior and posterior distributions obtained with NASA's approach are exactly the same.

4.5 Benign Failure Assessment

The benign failure case is different than that of the catastrophic failure in terms of the consistency of the collected evidence. More specifically, the number of failures in configurations A and B is very consistent (4 failures on 305 demands in Configuration A vs. 16 failures on 1319 demands in Configuration B), while the in Configuration C there is only 1 failure in 480 demands.

As a result, the Configuration C (actual) likelihood does not fit within the 90% failure bounds of the undiscounted prior distribution, while it fits well within the 90% failure bounds of the prior distribution obtained from the weighted likelihood approach with NASA's discounting level.

There are two possible reasons for that particularity. Either the technological improvements implemented in Configuration C were extremely successful in eliminating the failure conditions that could lead to benign failure, or, the collected evidence is not adequate for shaping a representative state-of-knowledge based on a Bayesian parameter estimation approach. Our analysis of the results obtained with the proposed approach will try to put some light on this issue.

The results obtained with the proposed approach with the use of different weights (w_D, w_{ND}) for the optimistic (D)[1] and the pessimistic (ND) outcomes are given in Table 4-7.

	Configuration C Posterior Characteristic Values						
	NASA	**No Discounting**	**Proposed Approach**				
(w_D, w_{ND})	n/a	n/a	(0.8, 0.2)	(0.6, 0.4)	(0.5, 0.5)	(0.4, 0.6)	(0.2, 0.8)
Mean	1.20E-03	9.96E-03	3.89E-03	4.39E-03	4.80E-03	5.30E-03	6.72E-03
St. Dev.	6.52E-04	2.16E-03	1.93E-03	2.12E-03	2.19E-03	2.24E-03	2.25E-03
5th	4.54E-04	6.69E-03	1.35E-03	1.57E-03	1.84E-03	2.22E-03	3.49E-03
50th	1.05E-03	9.81E-03	3.58E-03	4.06E-03	4.47E-03	4.99E-03	6.47E-03
95th	2.43E-03	1.38E-02	7.51E-03	8.35E-03	8.87E-03	9.45E-03	1.08E-02

Table 4-7. Sensitivity analysis of Configuration C benign failure posterior distribution on the use of different weights in proposed approach

We see in the results in Table 4-7 that the distributions obtained with the proposed approach have mean values that are equally spread in between the mean from NASA's approach and the one from not discounting, while they are all of the same order of magnitude. A more important observation though, is that the distribution obtained with NASA's approach is an *order of magnitude narrower* than those obtained with the proposed approach, implying a strong state of knowledge. Finally, the Configuration C likelihood fits within the 90% failure bounds of the obtained prior distributions when the weight of the optimistic prior is greater than 0.5[2].

The results obtained with NASA's approach are shown in Fig.4-7, while the Configuration C prior and posterior distributions obtained with the proposed approach

[1] The same failure applicability factor ($\gamma=0.6$) and environmental discounting factor [*uniform(1.520, 1.858)*] as in catastrophic failure case were applied.

[2] Not shown explicitly in Table 4-6

with the assignment of equal weights (w_D=0.5, w_{ND}=0.5) on the optimistic and the pessimistic elements are shown in Fig.4-8 and Fig.4-9, respectively.

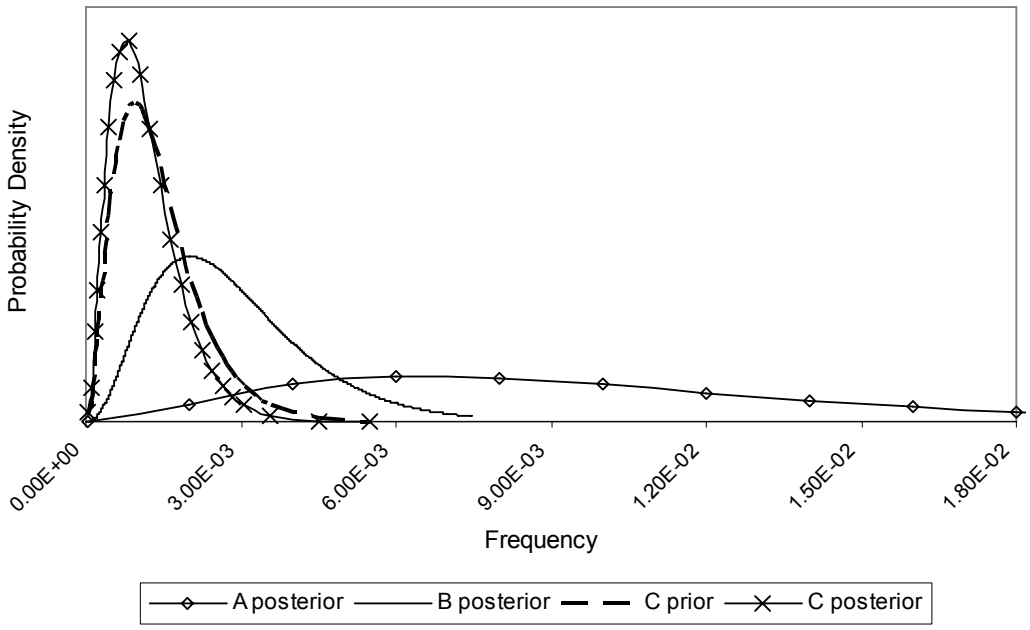

Figure 4-7. Configuration C benign failure results with weighted likelihood approach

We see in Fig.4-8 that a beta distribution does not fit well the resulting weighted prior distribution for Configuration C. On the other hand, the weighted posterior distribution for Configuration C (Fig.4-9) is described perfectly by a theoretical beta; however, this is because the optimistic and pessimistic elements of the Configuration C posterior are calculated from the Configuration C fitted beta prior.

A more important observation is that the Configuration C posterior distribution is shifted significantly to the left in response to the observed failures in this configuration (*n=1, N=480*). The Configuration C posterior distribution obtained with NASA's approach does not have the same behavior, although the failure observed in this configuration is discounted by 95% (Table 2-3). This means that Configuration C prior distribution already implied a strong state of knowledge.

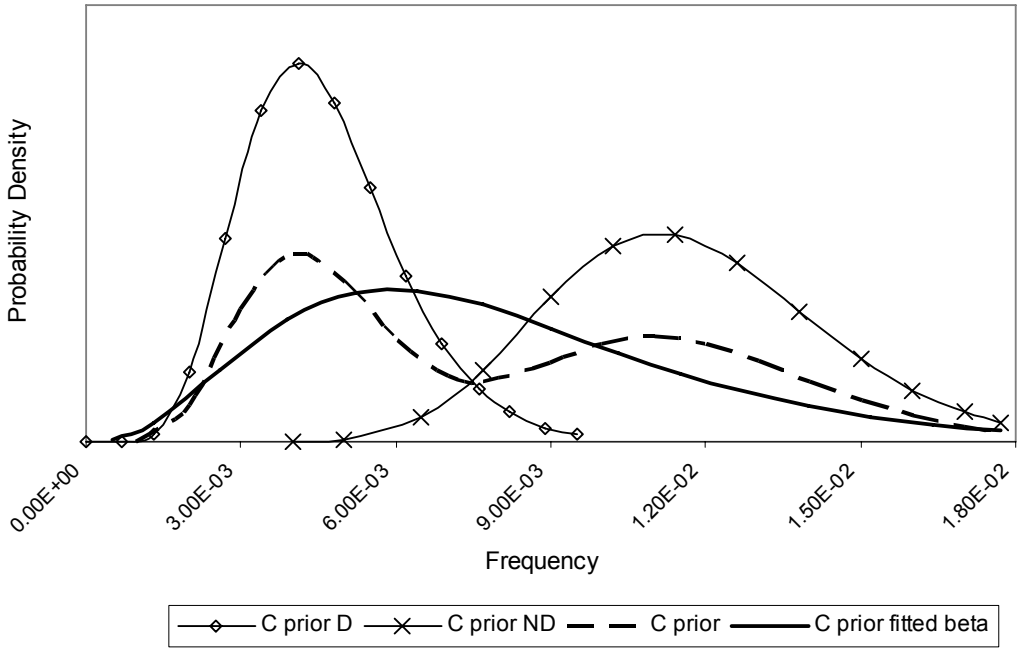

Figure 4-8. Configuration C benign failure prior distribution with proposed approach

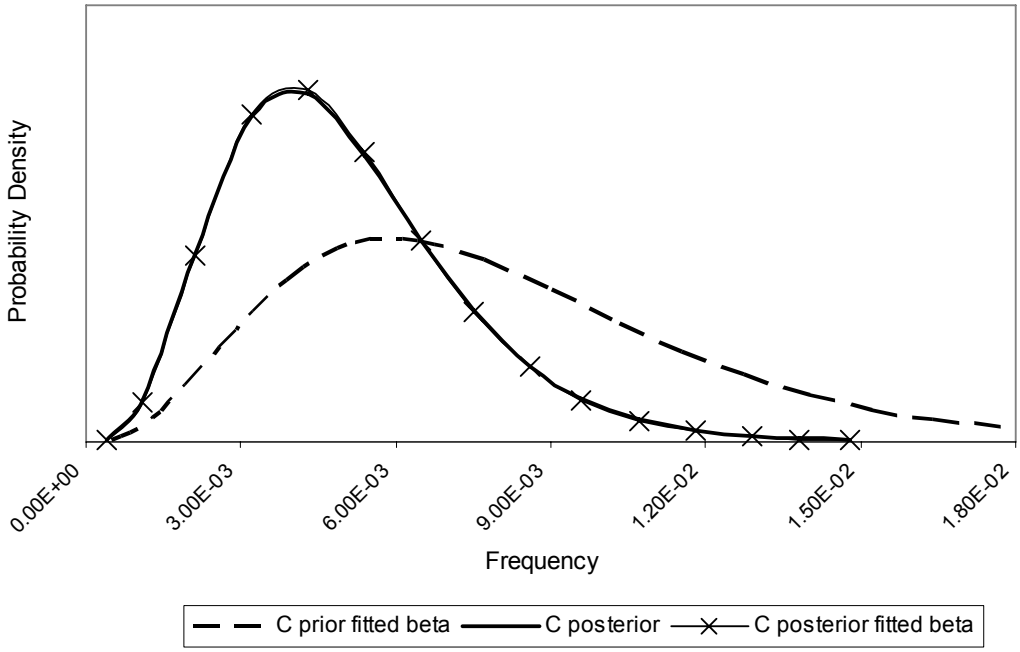

Figure 4-9. Configuration C benign failure posterior distribution with proposed approach

The fundamental question in this PRA analysis is "which of the two distributions shown in Fig.4-7 and Fig.4-9 is more representative of the current state of knowledge in order for a decision maker to base his / her decision on that distribution?"

As we see in Fig.4-8 and Fig.4-9, the posterior distribution obtained with the proposed approach accounts for the evidence in Configuration C by significantly shifting it to the left, while, at the same time, it maintains the 'influence' of the observed failures in previous configurations. We believe that, despite the small number of observed failures in Configuration C, the distribution in Fig.4-9 is representative of the current state of knowledge and has an appropriate behavior within Bayesian framework, while the distribution obtained by NASA's approach is both optimistic and narrow reflecting underestimation and overconfidence.

4.6 Alternative Approaches

We have seen in Section 4.2.2 that, when the problem of partial applicability of observed failures is modeled with a failure record discounting approach that discounts only the failure records without adjusting the "success records" as well, the derived distribution for the probability of failure is both optimistic and overly narrow, artificially implying a strong state of knowledge. This is because data have a dominant role in Bayesian analysis in shaping the posterior distribution of the parameter modeled.

An alternative modeling approach to the problem of partial applicability of observed failures is discounting both the failure and success records in order to account *neutrally* for the uncertainty that is inherent in the operating evidence. Engineering judgment is required for assessing the success applicability factors to be used. These factors could represent either the level of similarity in the system before and after the implementation of the changes that have been effected as a result of past failures, or even the *confidence* of the experts on their estimations on failure applicability factors.

The application of this approach to our reference case has given the results shown in Table 4-8. The failure applicability factors and model used are the same as in NASA's approach, with the difference that the data averaging Bayesian updating method is used instead of the weighted likelihood one.

Two cases are examined. In both cases, success applicability factors in each configuration are set equal to the relevant failure applicability factors assessed by NASA. This is done for demonstration reasons, but also, as an extreme interpretation of the uncertainty in failure record discounting. Since there are not any catastrophic failures in Configuration A, the success history from that configuration may have a disanalogous influence on the results; thus, in the second case the success history from Configuration A is also discounted by a factor equal to the combined success discounting in configurations B and C.

	Configuration C Posterior Characteristic Values				
	NASA	**No Discounting**	**Proposed Approach (equal weights)**	**Success Discounting**	**Success Discounting (& Conf. A)**
Mean	4.91E-04	5.94E-03	4.55E-03	1.81E-03	3.85E-03
St. Dev.	4.83E-04	1.67E-03	1.86E-03	1.38E-03	2.92E-03
5th	9.03E-05	3.48E-03	1.97E-03	2.64E-04	5.64E-04
50th	3.50E-04	5.78E-03	4.30E-03	1.48E-03	3.14E-03
95th	1.36E-03	8.93E-03	7.97E-03	4.50E-03	9.55E-03

Table 4-8. Configuration C catastrophic failure results with success discounting

The Configuration C posterior distribution derived from this approach (in both cases) is much wider than that obtained with NASA's method. In the second case, it is also wider than both the one obtained with the proposed approach and the one obtained from a standard Bayesian updating approach without discounting. On the other hand, the values of the mean and the median in both cases are lower than those obtained with the proposed

approach and without discounting. The meaning of this observation is that failure record discounting has a more significant impact on the resulting posterior distribution than success record discounting.

5. Summary and Conclusions

5.1 Summary

5.1.1 The problem

The issue of uncertain data in Probabilistic Risk Assessment (PRA) was investigated. The problem is that the operating information containing past failures of the system may not be directly applicable in Bayesian parameter estimation due to different environmental conditions and / or design improvements that have been implemented as a result of past failures.

This is a typical case in space systems where design or test / inspection procedure changes are applied on a regular basis. The motivation of this thesis was the Space Shuttle PRA Data Analysis Report (NASA, 2005), in which a methodology to account for partial applicability of past failures is presented. This methodology was applied for the first time to a system, which has evolved with three configurations A, B, and C over the course of the Space Shuttle program, while several changes have been implemented in response to observed failures or failure conditions. The application of NASA's approach to this system resulted to very small risk estimates for the latest configuration (C). This fact motivated our interest in reevaluating the above approach by using the specific case study as the basis for our work.

5.1.2 Current Approach

NASA's approach to the problem is primarily based on failure record discounting; that is, failures observed in previous or current versions of the system are discounted to reflect the implemented changes that followed. The failure applicability factors (γ), which are the means for failure record discounting, are based on the effectiveness of the design or procedure changes (assessed by relevant guidelines), and extensive use of engineering judgment.

The failure applicability factors (γ) are used in a Bayesian updating method developed by NASA, named Weighted Average Likelihood, to compute the posterior distribution of the system with respect to the discounted failure records. The basic assumption in this method is that the likelihood of a failure record being applicable in Bayesian updating can be approximated by a Binomial distribution with parameter the applicability factor (γ). That is, if n is the number of actual failures, the likelihood (P_x) of a failure record that counts $x=0,1,..,n$ failures of being applicable in Bayesian updating, is estimated by a Binomial distribution with parameter γ. The posterior distribution is then obtained by a posterior averaging approach, which averages the posterior distributions obtained for each of the applicable failure records ($x=0,1,..,n$), by using as weights their corresponding likelihoods (P_x).

This approach was applied for estimating the posterior distribution *within* each configuration. The model used for taking into account the operational experience of *previous* configurations was a multi-step Bayesian updating method, according to which the posterior of each predecessor is the prior for the successor configuration. Furthermore, the posterior distribution for Configuration B was adjusted twice, before being used as a prior for Configuration C, to account for system-wide improvements in operating environment and hardware. Also, a non-informative *Beta(0.5, 0.5)* was used as a starting prior distribution for Configuration A, while results were obtained for two different failure modes of the system, "catastrophic" and "benign" failure.

5.1.3 Insights from the Literature

The literature stresses the fact that the use of imprecise / uncertain data requires modeling decisions and there are uncertainties inherent in such modeling, which must be quantified and displayed in the results. Sensitivity studies are required to quantify these uncertainties and provide useful insights about how some of the assumptions affect the analysis. Furthermore, in the absence of significant operating information in the risk analysis of low-probability / high-consequence failures, the use of other kind of available

information, such as near miss events (precursors), is an important consideration than may affect results.

It has been widely documented that assessments of subjective probabilities are subject to cognitive biases, such as systematic (under-) overestimation and overconfidence. This is a problem realized in several early PRAs by different industries, including space systems. It is the role of the PRA analyst to consider such biases by selecting an appropriate technique for the elicitation, aggregation, and quantification of expert opinions, or even for their adjustment after being elicited, in order to remove cognitive biases.

Our literature review revealed several approaches that have been developed for Bayesian updating with uncertain data. The posterior averaging, weighted likelihood, data averaging, likelihood in terms of observation, and an ad-hoc overarching approach have been proposed. Siu and Kelly (Siu & Kelly, 1998) suggest that the basic framework for dealing with uncertain data must address the question of whether uncertain data should be dealt with within the likelihood function or the posterior distribution.

Finally, several cases have been found in the literature that deal with the problem of consideration of uncertain data in specific PRA applications. Two cases concern the use of operating information from heritage systems in space system PRAs and are more relevant to our problem. The posterior averaging method is used in most cases, while the data averaging approach has proved to be a good approximation in any of the cases that was evaluated.

5.1.4 Proposed Approach

Our objective was to propose an appropriate Bayesian updating approach for calculating the probability of failure on demand for the latest configuration C of our system, by using the available operational information from the heritage configurations A and B, while

considering the changes in design and test / inspection procedures that have been implemented in the system over the course of the Space Shuttle program.

Our approach is based on the view that failure record discounting provides insights if it is used as a means for providing an optimistic interpretation of the data in PRA, when the collected failures do not have direct applicability. The optimistic interpretation is then combined and presented along with a distribution obtained with actuarial data (pessimistic) to provide a basis for risk-informed decision making. The aggregation of the two input distributions is performed subjectively by executing a series of sensitivity studies using a weighted averaging approach with several weights. It is the decision makers who decide which "average" distribution they should use based on the insights gained by plotting the posterior distributions resulting from the most optimistic and the most pessimistic interpretation of the data.

A series of sensitivity studies on NASA's approach were executed to provide insight into several dimensions of the problem. A guiding rule in developing our approach is that special caution should be exercised when assessing failure record applicability factors in order to avoid generating overly narrow and optimistic failure distributions artificially indicating a strong state of knowledge. This is the case when failure data are discounted while success data are not. The results obtained with and without discounted confirmed our concern.

Furthermore, the Weighted Average Likelihood method developed by NASA was found to have considerable deviations from the way epistemic quantities, such as the applicability of a failure record (P_x), should be treated within Bayesian framework. Thus, the Data Averaging Bayesian updating was used in our approach instead.

Our results in the reference case were obtained by the use of a multi-step Bayesian updating approach for obtaining the optimistic and pessimistic elements of our approach, which were then weighted with several weights. A relatively soft subjective adjustment of NASA's failure record discounting factors was considered important in order to obtain the optimistic element of our approach without violating any of the model

assumptions. This adjustment was based on our analysis, observations, and literature review. The application of the proposed approach on the catastrophic failure model resulted in posterior distributions, which have mean values of about an order of magnitude greater than NASA's estimate, and are also significantly wider. Finally, the approach was also applied in the benign failure case with similar results.

5.2 Conclusions

The interpretation of the uncertainty that is inherent in the case of partial failure applicability depends on the assumptions and the models used to account for the uncertain data and apply them in Bayesian analysis.

The methodology developed by NASA is based on failure record discounting. Failure record discounting can have a very optimistic effect on the obtained posterior distributions. This is due to the dominant role of data in Bayesian analysis in shaping the posterior distribution. When failure records are discounted, without some kind of "success record" discounting, a significantly left-shifted and overly narrow posterior distribution should be expected, depending of course on the volume of the data.

Furthermore, when engineering judgment is used for a direct assessment of failure applicability based on design or procedure improvements, there is room for biases in the elicitation and use of subjective information. These biases will tend to overestimate the impact of these improvements. This is probably a concern in our case, where the existing method resulted to risk estimates an order of magnitude smaller than those obtained from a standard Bayesian updating method with actuarial data, without any major technological change being involved.

The sensitivity studies of the existing approach on the different modeling assumptions and decisions showed that:

- The data averaging approach yields reasonably accurate results without deviating from the way epistemic uncertainties should be treated in Bayesian framework.
- When extensive discounting is involved, the multi-step Bayesian updating approach, which is based on the assumption of high level of similarity among the system different configurations, magnifies the effect of discounting on the results.
- When discounting is done extensively, even the use of a non-informative prior exerts some influence on the shape of the posterior distribution.
- Failure record discounting has a direct impact on risk results that is analogous of the involved discounting.

The proposed approach deals with most of these issues by weighting an optimistic and a pessimistic interpretation of the data in a linear pooling function. By doing so, the engineering judgment, which is a possible source of bias, is constrained to the one element of the model. Thus, there is more space for developing a more liberal optimistic distribution as far as engineering consensus is achieved and model assumptions are not violated.

A basic advantage of the proposed methodology over the existing one is that it communicates better the inherent uncertainties in the data, by giving distributions that are wider than the distribution obtained without failure record discounting, an observation that supports our claim that failure record discounting is an activity that should increase our uncertainty about the failure parameter. Moreover, the interpretation of the results is done in such way that it displays the uncertainties that are inherent in the case of failure record partial applicability and provide a better basis for the decision maker to make a decision based on his / her risk attitude.

Finally, the application of an alternative method that discounts both the failure and the success records was considered. The shape of the resulting distribution depends

on the level of success discounting involved. However, when success discounting factors are similar to failure discounting factors, the resulting distribution is wider than the one obtained with the proposed approach, but more remarkably, it is also wider than the one obtained from a standard Bayesian updating approach without discounting. This observation supports our claim that, a modeling decision to discount the failure records could impact risk estimation across all PRA elements, therefore, it should be done carefully.

(This page intentionally left blank)

References

Apostolakis G., Kaplan S., Garrick B. & Duphily R., 1980, "Data Specialization for Plant Specific Risk Studies", *Nuclear Engineering and Design*, 56:321-329

Apostolakis G. & Kaplan S., 1981, "Pitfalls in Risk Calculations", *Reliability Engineering*, 2:135-145

Apostolakis G., 1990, "The Concept of Probability in Safety Assessments of Technological Systems", *Science*, 250:1359-1364

Apostolakis G., 1994, "A Commentary On Model Uncertainty", Report NUREG / CP-0138, *US Nuclear Regulatory Commission*, Washington, D.C.

Apostolakis G., 2004, "How Useful is Quantitative Risk Assessment?", *Risk Analysis*, 24:515-520

Apeland S., Aven T. & Nilsen T., 2002, "Quantifying Uncertainty under a Predictive, Epistemic Approach to Risk Analysis", *Reliability Engineering and System Safety*, 75:93-102

Bier V. M. & Mosleh A., 1988, "The Subjective Bayesian Approach to PRA: Theoretical and Practical Perspectives", *Reliability Engineering and System Safety*, 23:269-275

Clemen R. T. & Winkler R. L., 1999, "Combining Probability Distributions From Experts in Risk Analysis", *Risk Analysis*, 19:187-203

Guarro S., Bream B., Rudolph K. & Mulvihill R., 1995, "The Cassini Mission Risk Assessment Framework and Application Techniques", *Reliability Engineering and System Safety*, 49:293-302

Guarro S. & Tomei E., 2004, "Launch Vehicle Development Risk Projection Methodology", *Proceedings of the 7th International Conference on Probabilistic Safety Assessment and Management (PSAM 7 – ESREL '04)*, Berlin, Germany, June 14-18, Editors: C. Spitzer, U. Schmocker, and V. N. Dang, Springer, London, United Kingdom.

Guikema S. & Paté-Cornell M., 2005, "Probability of Infancy Problems for Space Launch Vehicles", *Reliability Engineering and System Safety*, 87:303-314

O'Hagan A. & Oakley J. E., 2004, "Probability is Perfect, but We Can't Elicit it Perfectly", *Reliability Engineering and System Safety*, 85:239-248

Independent Peer Review Panel (IPRP) on the PRA of the Space Shuttle, Final Report, Office of Safety and Mission Assurance, NASA Headquarters, April 4, 2005

Kaplan S., 1990, "On the Inclusion of Precursor and Near Miss Events in Quantitative Risk Assessments: A Bayesian Point of View and a Space Shuttle Example", *Reliability Engineering and System Safety*, 27:103-115

Kaplan S., 1992, "'Expert Information' versus 'Expert Opinions.' Another Approach to the Problem of Eliciting / Combining / Using Expert Knowledge in PRA", *Reliability Engineering and System Safety*, 35:61-72

Martz H. F., Kvam P. H. & Atwood C. L., 1996, "Uncertainty in Binomial Failures and Demands with Applications to reliability", *International Journal of Reliability, Quality and Safety Engineering*, 3:43-57

Martz H. F. & Hamada M. S., 2003, "Uncertainty in Counts and Operating Time in Estimating Poisson Occurrence Rates", *Reliability Engineering and System Safety*, 80:75-79

Mosleh A. & Apostolakis G., 1984, "Models for the Use of Expert Opinions", *In Low-Probability / High-Consequence Risk Analyses: Issues, Methods, and Case Studies*, edited by R. A. Waller and V. T. Covello, Plenum, New York

Mosleh A., Bier V. & Apostolakis G., 1988, "A Critique of Current Practice for the Use of Expert Opinions in Probabilistic Risk Analysis", *Reliability Engineering and System Safety*, 20:63-85

Mosleh A., 1991, "Common Cause Failures: An Analysis Methodology and Examples", *Reliability Engineering and System Safety*, 34:249-292

Paté-Cornell M. & Dillon R., 2001, "Probabilistic Risk Analysis for the NASA Space Shuttle: a Brief History and Current Work", *Reliability Engineering and System Safety*, 74:345-352

Paté-Cornell M., 2002, "Risk and Uncertainty Analysis in Government Safety Decisions", *Risk Analysis*, 22:633-646

Phimister J. R., 2005, "Flirting with disaster", *Issues in Science and Technology*, 22:31-35

Pietzsch J., Paté-Cornell M. & Krummel T., 2004, "A Framework for Probabilistic Assessment of New Medical Technologies", *Proceedings of the 7th International Conference on Probabilistic Safety Assessment and Management (PSAM 7 – ESREL '04)*, Berlin, Germany, June 14-18, Editors: C. Spitzer, U. Schmocker, and V. N. Dang, Springer, London, United Kingdom.

Rausand M. & Hoyland A., 2004, System Reliability Theory, *Wiley Interscience*, New Jersey, USA

Siu N. & Apostolakis G., 1986, "Modeling the Detection Rates of Fires in Nuclear Plants: Development and Application of a Methodology for Treating Imprecise Evidence", *Risk Analysis*, 6:43-59

Siu N., 1990, "A Monte Carlo Method for Multiple Parameter Estimation in the Presence of Uncertain Data", *Reliability Engineering and System Safety*, 28:59-98

Siu N. & Kelly D., 1998, "Bayesian Parameter Estimation in PRA", *Reliability Engineering and System Safety*, 62:89-116

Tan Z. & Xi W., 2003, "Bayesian Analysis with Consideration of Data Uncertainty in a Specific Scenario", *Reliability Engineering and System Safety*, 79:17-31

Vitali R. & Lutomski M., "Derivation of Failure Rates and Probability of Failures for the International Space Station Probabilistic Risk Assessment Study", *Proceedings of the 7th International Conference on Probabilistic Safety Assessment and Management (PSAM 7 – ESREL '04)*, Berlin, Germany, June 14-18, Editors: C. Spitzer, U. Schmocker, and V. N. Dang, Springer, London, United Kingdom.