Empirical Bayes Estimates of Development Reliability for One Shot Devices

John Quigley, Tim Bedford, Lesley Walls
Department of Management Science, University of Strathclyde, Glasgow

This article describes a method for estimating the reliability of a system under development that is an evolution of previous designs. We present an approach to making effective use of heritage data from similar operational systems to estimate reliability of a design that is yet to realise any data. The approach also has a mechanism to adjust initial estimates in the light of sparse data that becomes available in early stages of test. While the estimation approach, known as empirical Bayes is generic, we focus on one shot devices as this was the type of system which provided the practical motivation for this work and for which we illustrate an application.

Introduction

This paper is motivated by the problem of predicting the probability of successful operation for a one shot device during its development. The device is a variant of existing operational systems, and hence heritage event data exists, albeit limited. This work developed from an industrial application in which the client and manufacturer were interested in understanding both the overall success probability and the current risk drivers in order to support the development process further. Hence a model was required to: measure the contributions of events to identify the main drivers of unreliability; quantify uncertainties in the estimated success probabilities under different scenarios; and update estimates using observations generated from development tests. In this paper we focus upon the first of these objectives, which requires us to estimate the probability of successful operation of the system on demand and to identify the contributing events that may drive reliability.

Classical estimates of reliability use observed data and, especially when sample sizes are small, errors in the estimates obtained can be large. For example, Figure 1 illustrates possible estimation errors arising when a traditional approach to estimating the probability of success as the ratio of the number of occurrences of success events to the total number of trials. In this illustration, we assume that there have been 50 trials and so the resulting possible estimates of success probability will increase from 0 (if we observe 0 successes in 50 trials) to 1 (assuming we observe 50 successes in 50 trials) in increments of 0.02 (i.e. 1 in 50). The error is measured as the estimate of the success probability minus the true value, which we have assumed to be 0.1. Figure 1 shows the distribution of the probability of error under the assumed scenario where the probability of the errors are measured in the height of the bars. For example, there is a probability of approximately 0.18 that we would correctly estimate the probability of a success to be 0.1 and hence have an error of 0. The main point we make is that even with a sample size of 50, which may be larger than will be experienced in many practical contexts, the range of the error can be large. In the case of this example, the error ranges from -0.1 to 0.1. Such error is of particular concern when we seek to identify and compare different reliability drivers since it can lead to incorrect prioritisation.

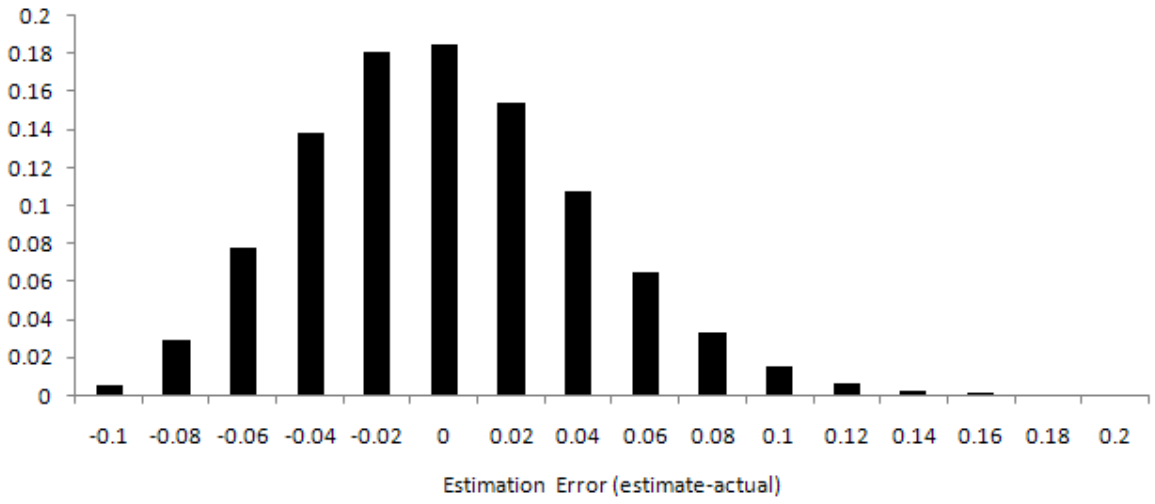

Figure 1 Distribution of possible estimation errors for a sample of 50 trials and an assumed underlying probability of success set to 0.1

An alternative approach to classical estimation is to use methods which rely on subjective probability assessments, such as Bayesian methods; see, for example Martz and Waller (1991). However Bayesian methods require considerable problem structuring to assess uncertainties with all possible events and hence formulate a suitable prior distribution. While problem structuring can be beneficial since insights gained can inform qualitative reliability assessment, the quantification of an engineering expert's uncertainty can result in a significant cognitive burden and so needs to be managed with care.

While Bayesian priors are usually constructed from subjective beliefs about the value of an event probability (Bedford et al 2008), empirical Bayes (Carlin and Louis 2000) provides a means of pooling observed data from various sources related to heritage systems to form an empirical prior. It is within an empirical Bayes framework that we anticipate greatest benefit for reliability estimation during development using heritage data from similar products. This is because it allows us to use expert judgment to support the qualitative aspects of modelling, such as explicating the differences and similarities of the new design with heritage products, but allows us to use relevant historical data to implement an empirical Bayes estimation procedure to quantify the probability of success.

The application of empirical Bayes (EB) within the context of risk and reliability is not new. For example: Martz and Waller (1991) discuss the technique generally; Vesely et al (1994) discuss an application to emergency diesel generators for binary data; Vaurio (2002, 2005) uses EB for estimating the rate of common cause failures; Ferdous et al (1995) use EB to support inference for the Weibull distribution within a software reliability growth context; Grabski and Sarhan (1996) combine spline density estimates for prior distributions with EB for inference with the exponential distribution; Vaurio and Jankala (2006) use EB within a Poisson modelling framework; and Sohn (1999) makes use of the methodology with response surface modelling of categorical quality characteristics of possible designs; Quigley et al (2007) used EB to estimate the probability of low frequency, high consequence events in a railway safety model.

This paper describes the proposed methods developed to support empirical Bayes estimation for the probability of successful operation of a one shot system modelled by a fault tree during its development. The full modelling process is illustrated through a (de-sensitised) industrial example. The originality of the approach described within this paper is the combination of the elicitation of the prior distributions for a new one-shot system by identifying aspects of relevant heritage systems using subjective engineering judgment and quantifying the prior distributions with historical data through EB. This reduces the cognitive burden on the engineering experts as well as supporting inference on the likelihood of events that have not been experienced by heritage systems.

Summary of Empirical Bayes Process

Figure 2 presents the process for constructing the EB estimate. Since we are concerned with success/failure outcomes on test or trial, the probability model that we use to describe the variability in the data is a Binomial distribution. We assume we have *m* different basic events, each with a unique probability of occurrence and we denote the probability of the i^{th} event by p_i. We seek to describe the variability of the p_i's with a distribution, which we estimate by pooling data and refer to as the empirical prior distribution. This is similar the prior distribution in Bayesian analysis except that it is not elicited from experts but inferred from observed data, which is made possible by the assumption that we have several, i.e. *m*, independent samples from this prior distribution. Once the prior distribution has been estimated we can use Bayesian updating to estimate a posterior distribution for each event, thus we have *m* different posterior distributions.

Parametric example

Let $\pi(p)$ denote the probability density function for an event, chosen at random from the *m* basic events identified, having probability of occurrence *p*. Assuming that the number of events has a Binomial distribution implies that a computationally convenient prior distribution is a Beta distribution of the form:

$$\pi(p)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}, \quad \alpha,\beta>0, \quad 0<p<1$$

Since the unknown probability of an event can be assumed to be selected at random from a Beta distribution, an average of the Binomial distributions weighted against the prior distribution provides the probability distribution of the number of events N_i that will occur for a given event *i*, based only on our knowledge of the pool. Mathematically this can be simplified to a Polya-Eggenberger distribution, which describes the variability in the pooled data so can be used to provide estimates $\left(\hat{\alpha},\hat{\beta}\right)$ of (α,β):

$$\begin{aligned}\mathrm{P}(X=\mathrm{x}) &= \int_0^{\infty}\binom{n}{x}p^{x}(1-p)^{n-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}dp \\ &= \frac{n!\Gamma(\alpha+\beta)\Gamma(\alpha+x)\Gamma(\beta+n-x)}{(n-x)!x!\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)}\end{aligned}$$

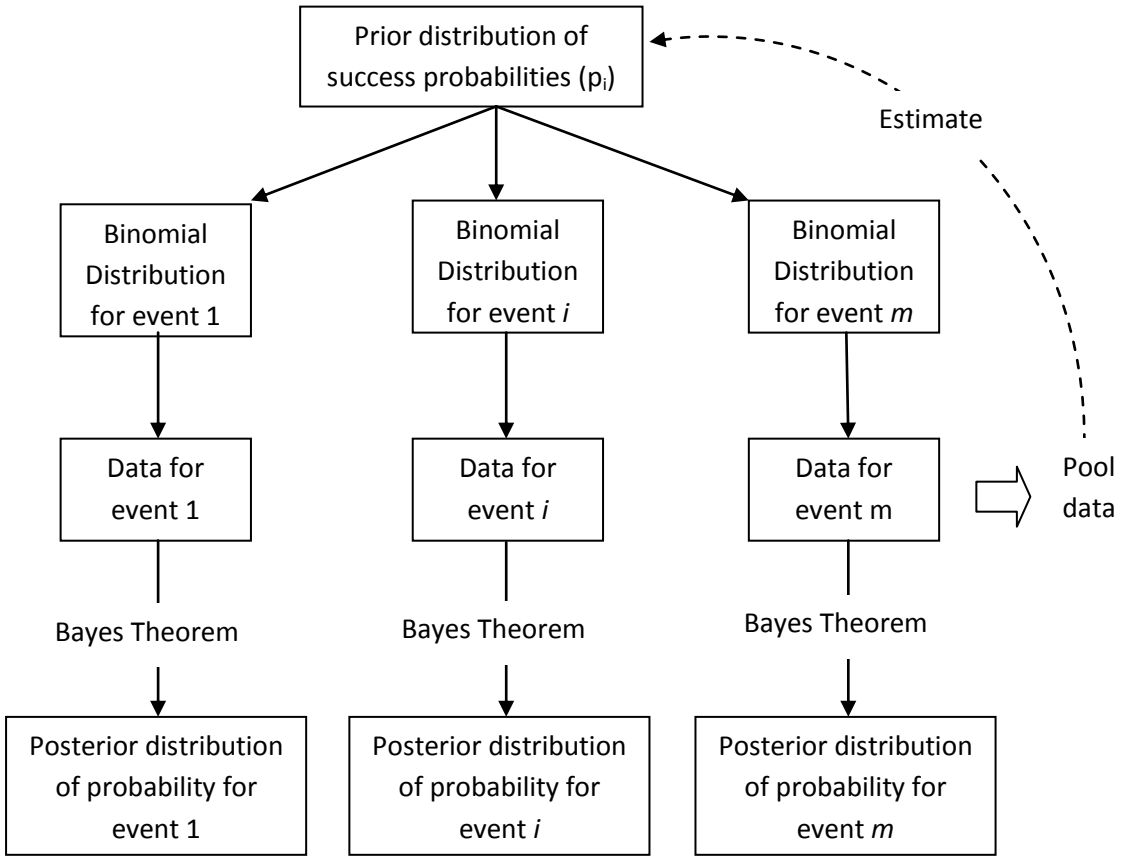

Figure 2 Empirical Bayes Process for a one-shot system where there are *m* basic events in the fault tree model

Given observed data of the form x_i events within n_i trials and the Beta prior distribution, the posterior distribution for basic event *i* updates to a Beta distribution given by:

$$\pi\left(p_i \middle| x_i, n_i, \hat{\alpha}, \hat{\beta}\right) = \frac{\Gamma\left(\hat{\alpha}+\hat{\beta}+n_i\right)}{\Gamma\left(\hat{\alpha}+x_i\right)\Gamma\left(\hat{\beta}+n_i-x_i\right)} p_i^{\hat{\alpha}+x_i-1}\left(1-p_i\right)^{\hat{\beta}+n_i-x_i-1}$$

The point estimate for the probability of the i^{th} basic event being realised at the next opportunity is derived from the expectation of the posterior distribution resulting in:

$$E\left(p_i \middle| x_i, n_i, \hat{\alpha}, \hat{\beta}\right) = \frac{\hat{\alpha}+x_i}{\hat{\alpha}+\hat{\beta}+n_i} = \frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}}\left(1-Z\right)+\frac{x_i}{n_i}Z$$

where:

$$Z = \frac{n_i}{\hat{\phi}+n_i}$$

$$\hat{\phi} = \hat{\alpha}+\hat{\beta}$$

Therefore, the estimate of the probability of an event occurring is a weighted average between the observed proportion of the individual experience (i.e. x_i/n_i) and the mean of the pool or prior distribution (i.e. $\hat{\alpha}/\left(\hat{\alpha}+\hat{\beta}\right)$). The weight applied to the individual observation tends to 1 as the number of opportunities for the event to be realised (i.e. n_i) increases. The weight applied to the individual experience rather than the pooled experiences is influenced by the heterogeneity within the pool. If the pool consists of a collection of heterogeneous events, (i.e. there is great variation in the p_i's) then $\hat{\phi}$ will be small and more weight will be applied to the individual event. Moreover, if greater variation in the data is expected for individual events, then less weight is given to the individual experience and more to the pooled experience.

Illustrative Example

System and its reliability requirements

The new system for which reliability is to be estimated is in an advanced state of design, having recently had initial tests. A number of issues have been identified surrounding its reliability since the new system is largely based on an existing system which is already in service and for which two major problems have been identified that should be removed through the re-design. There are a number of additional substantial design changes to the system, implying that some subsystem elements are not simple developments of the existing design.

This study considers only the probability of achieving a successful operation when the decision to use the system has been taken and the signal has been transmitted to the system. Since the system is still under design, and information such as a FMECA is not available, it has been necessary to conduct this study at a fairly high level. This means that the study can highlight areas of concern with the model, and provide a broadbrush overview of system reliability. However it cannot include detailed modelling of the subsystems. Such a detailed model would not however support the purpose of supporting management decisions about the system, as it could only be built in retrospect. This implies that the model built here is indicative rather than a hard forecasting model.

The general steps in model development are as follows:

- Development of high level fault tree including agreement with engineering team members on the structure and on definition of events;
- Interviews with engineering team members to gain insights into sources, main drivers, ownership, potential testing of uncertainties, and agreement about relevant historical data;
- Construction of spreadsheet model to estimate success probability of the system and conduct sensitivity analysis to investigate uncertainties relative to requirements.

Structuring the fault tree

The approach adopted has been to allow the data collected from stakeholders to 'speak for itself' and hence ground the model formulation in experience of designing and operating the system. Modelling began by gaining agreement about the top event under study. The first draft of the fault tree was made by the Strathclyde team on the basis of information gathered from an initial design

team meeting. The draft has been discussed in numerous telephone and face-to-face interviews with both the client and the manufacturer leading to evolution of the mature fault tree.

The fault tree developed is relatively "shallow" in the sense that it has not been developed to a deep level of complexity and consists largely of 'OR' logic gates as there is little redundancy at the level of modelling. The fault tree comprised 31 basic failure events.

Figure 3 summarises the key steps in the elicitation of the fault tree structure and probability of events. Experts were selected from both the manufacturer and client to provide coverage of the lifecycle demands upon the system. A formal elicitation proforma was used to structure interviews capturing reasons for expert selection.

Eliciting event probabilities

Despite the design changes between the system generations, it transpired that many events for the new system could be conditioned upon operational experience for earlier generations, since there remains commonality between the types of failure mechanisms.

For those events where historical data are deemed appropriate, the related design(s) are identified and the reasons for conditioning upon them noted. Then the relevant sample size and number of observed events are elicited. For example, the 'sample size' relates to the relevant exposure of the system to risk of an event. As a default, this is noted in terms of the number of units exposed to operational use and hence having the potential to experience an event. However, the number of parts at risk is also recorded where considered appropriate. Since some failure mechanisms can occur only on demand, while others may be latent and hence observed on inspection, the values noted for the sample size correspond to the number of systems deemed relevant for each event. Conditional upon this definition of event sample size, the observed numbers of failures are noted. On occasions where there is uncertainty about the relevance of observed events, as in the case of failures that occurred before a modification, then this is noted and used to inform sensitivity analysis.

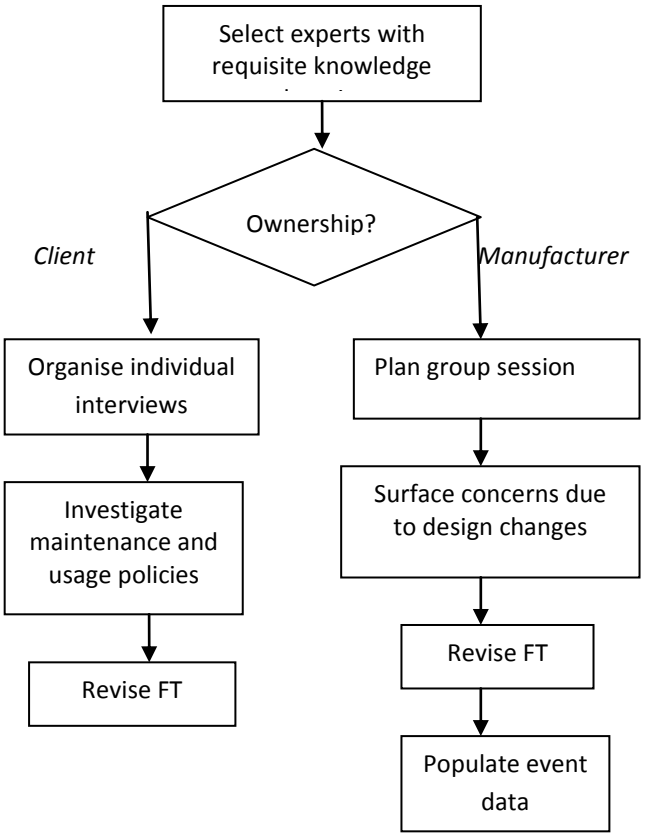

Figure 3. Process to elicit judgements from engineering experts

When empirical data are not considered relevant, then judgemental data about the uncertainty in the basic failure is elicited. The goal is to obtain information about the expected probability of realising a basic failure as well as lower and upper bounds on the uncertainty in that probability. Practically it is convenient to approach this by considering the number of realisations if 100 (or 1000 or 10000) systems are in use. This allows statements to the effect that the engineering expert would expect X events in 100 units; would be surprised if Y events in 100 units; would be disappointed if Z events in 100 units. From this data, we can extract the average probability of an event as well as lower and upper bounds on the uncertainty respectively. This approach is consistent with the fractile method for subjective probability assessments. Although only three fractiles have been used in this case –5%, 50% and 95% – it is deemed appropriate for the accuracy of modelling required. Further, the manufacturer experts appeared comfortable with providing expected numbers plus surprise limits. Indeed we have adopted their natural language descriptions of our statistical fractiles.

Since the client personnel interviewed were not comfortable in providing subjective judgement about the probabilities of basic failures for which they had ownership, the data used for all events is based on the manufacturer's knowledge. There are two exceptions: the basic failures associated with activating the system which was computed by other analysts; and an assessment of human

error made using the HEART method since this was considered outside the range of expertise of the expert group.

It is assumed that the degree of degradation is not sufficient to impact the reliability during operation since the historical evidence from previous generations is that there were no incipient failures on routine inspection.

Of the 31 basic failures, 23 were conditioned upon empirical data for previous designs, 6 were based on the manufacturer's subjective engineering judgement, 1 from generic handbooks and 1 is provided from other analysts. For some events it was noted that additional data would be provided from the outcome of future tests.

Most basic failures are described naturally with on-demand probabilities of failure, with the exception of three events where data was originally provided in the form of a rate of occurrence over the mission. These data were converted to a probability of an event to ensure commonality of the data structures for modelling and estimation. As these rates are associated to degradation processes taking place over an approximately fixed period, there was a natural way to convert these rates to probabilities.

An example of the typical data structures is given in Table 1. Note that the data presented are for illustration of the typical core characteristics only.

Table 1 Example of typical empirical and judgemental data structures for selection of basic events within the fault tree of one-shot system, note that numerical values are for illustration only

Event No.	*No. Units at Risk*	*No. Obs. Events*	*Lower Bound*	*Median*	*Upper Bound*	*Other Prob*
1.1	100	0				
1.2	50	0				
1.3			0.0001	0.001	0.005	
2.1	150	2				
2.2	20	1				
2.3			0.00001	0.0001	0.001	
3.1						0.02

Predictions of success probability

The EB methods derived in the previous section have been implemented within a spreadsheet model taking the data structures in Table 1 as input. An example of (de-sensitised) output is shown in Table 2.

Table 2. Empirical Bayes estimate of probability of success for fault tree top event

P(success)		0.92
Key reliability drivers		
Event	P(failure)	Potential accumulated improvement in P(success) if removed
3.5	0.03	0.95
1.4	0.02	0.97
2.3	0.01	0.98
1.5	0.005	0.985

Interestingly the key basic failures driving reliability correspond to those where there is a lack of empirical historical data and concern has been raised during the review of design changes.

Through elicitation of the fault tree and the event data, several sources of uncertainty have been investigated through sensitivity analysis. These include uncertainties about human reliability, parts count and intended actions. The analysis indicated that point estimates were robust for modelling conducted at unit and part level, that growth had occurred, and was forecast to continue, through development, and the impact of human reliability may be significant in operations.

Summary and Conclusions

The approach described in this paper has been extended to include estimates of the uncertainties in the estimates so that lower and upper bounds can be generated for the probability of success of the top event. Also, the Bayesian framework supports updating of the estimates in the light of observations from tests and analysis on the new system design. Such extensions are discussed in Bedford et al (2007).

To conclude we examine the benefits and shortcomings with our approach to EB estimation.

A key strength of this approach has been the acquisition of numerical assessments for prior distributions using data rather than stretching the cognitive ability of the engineers outside their zone of comfort. The engineering experts engaged in discussions about potential basic failures for the new system, and identified relevant historical data from which a prior could be constructed. Identifying the appropriate historical data was not an easy task, as this involved a basic event by event analysis of all heritage systems.

The method provides a numerical estimate of the probability of occurrence for events that have not been experienced by heritage systems but potentially could have been. This is an improvement over median based estimation procedures (Bailey, 1997), because the EB estimate is made relative to the pool and as such there is calibration to some extent for the collective.

However, the proposed method could be subject to abuse because the grouping together of basic events to form a pool from which EB estimates can be supported provides the potential for influencing the outcome. For example, by identifying basic events that are orders of magnitude lower in probability of occurrence than the pool, the overall pool mean will be lowered which contributes to the weighted average. While the proposed method does correct for this - a more

heterogeneous pool will put more weight on each basic event individual experience – it could still be open to abuse if the integrity of the analysis is not assured.

Acknowledgements

We would like to thank Ken Waylett, Chris Hankin and Martyn Jones from Airborne Systems Limited and the MOD staff for their contributions to this study.

References

R Bailey (1997) Estimation from Zero-Failure Data, *Risk Analysis*, 17, pp 375-380.

Bedford T, Quigley J and Walls L (2007) Expert Elicitation for Reliable System Design (with discussion), *Statistical Science*, 42, 428-462.

B Carlin and T Louis (2000) *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd Edition, Chapman and Hall/CRC.

J Ferdous, M Borhan Uddin and M Pandey (1995) Reliability Estimation with Weibull Inter Failure Times, *Reliability Engineering and System Safety,* 50, pp 285-296.

F Grabski and A Sarhan (1996) Empirical Bayes Estimation in the Case of Exponential Reliability, *Reliability Engineering and System Safety*, 53, pp 105-113.

H Martz and M Lian (1974) Empirical Bayes Estimation of the Binomial Parameter, *Biometrika*, 61, pp 517-523.

H. Martz and R Waller (1991) *Bayesian Reliability Analysis*, Kreiger Publishing Company.

T Bedford, J Quigley and L Walls (2006), Fault Tree Inference using Bayes and Empirical Bayes Methods, Proceedings of *ESREL 2006.* Portugal.

J Quigley, T Bedford and L Walls (2007) Estimating Rate of Occurrence of Rare Events with Empirical Bayes: A Railway Application *Reliability Engineering and System Safety* 92(5), pp 619-627.

SY Sohn (1999) Robust Parameter Design for Integrated Circuit Fabrication Procedure with Respect to Categorical Characteristic, *Reliability Engineering and System Safety*, 66, pp 253-260.

JK Vaurio (2002) Extensions of the Uncertainty Quantification of Common Cause Failure Rates, 78, pp 63-69.

JK Vaurio (2005) Uncertainties and Quantification of Common Cause Failure Rates and Probabilities for System Analyses, *Reliability Engineering and System Safety*, 90, pp 186-195.

JK Vaurio and KE Jankala (2006) Evaluation and Comparison of Estimation Methods for Failure Rates and Probabilities, *Reliability Engineering and System Safety*, 91, pp 209-221.

WE Vesely, SP Uryasev and PK Samanta (1994) Failure of Emergency Diesel Generators: A Population Analysis Using Empirical Bayes Methods, *Reliability Engineering and System Safety,* 46, pp 221-229.