

Chapter 7

A Methodology for Constructing Subjective Probability Distributions with Data

John Quigley and Lesley Walls

Abstract Our methodology is based on the premise that expertise does not reside in the stochastic characterisation of the unknown quantity of interest, but rather upon other features of the problem to which an expert can relate her experience. By mapping the quantity of interest to an expert's experience we can use available empirical data about associated events to support the quantification of uncertainty. Our rationale contrasts with other approaches to elicit subjective probability which ask an expert to map, according to her belief, the outcome of an unknown quantity of interest to the outcome of a lottery for which the randomness is understood and quantifiable. Typically, such a mapping represents the indifference of an expert on making a bet between the quantity of interest and the outcome of the lottery. Instead, we propose to construct a prior distribution with empirical data that is consistent with the subjective judgement of an expert. We develop a general methodology, grounded in the theory of empirical Bayes inference. We motivate the need for such an approach and illustrate its application through industry examples. We articulate our general steps and show how these translate to selected practical contexts. We examine the benefits, as well as the limitations, of our proposed methodology to indicate when it might, or might not be, appropriate.

7.1 Introduction

Our goal is to acquire a probability distribution consistent with an expert's belief about the true value of a quantity of interest. In this chapter we explain how to construct such a prior probability distribution using observed data by adopting an empirical Bayes method embedded within an elicitation process to achieve consistency between the distribution obtained and the judgement of an expert. Motivated by the need to elicit subjective distributions within real industry applications, the

J. Quigley (✉)
University of Strathclyde, Glasgow, UK
e-mail: j.quigley@strath.ac.uk

L. Walls
Department of Management Science, University of Strathclyde, Glasgow G4 0GE, UK
e-mail: lesley.walls@strath.ac.uk

methodology is grounded in core theoretical principles and aims to provide a useful, scientifically sound approach.

Core to our reasoning is the consideration of the ways in which an expert might assess uncertainty through analogy with similar events. In this respect we adhere to the view expressed by David Hume (1748) who wrote that “*All our reasonings concerning matter of fact are founded on a species of analogy*”. Others have acknowledged the role of empirical data for similar events in making assessments of uncertainty. For example, Kahneman and Lovallo (1993) proposed using empirical data as a means of correcting for overconfidence and optimism bias which might exist when an expert is asked to express her subjective assessments directly. Inherent in their so-called outside view is the mapping between the observed histories of the similar events and the future histories of the events associated with the quantity of interest. Practically, such an approach can be operationalised in various ways, including as a read-across process as discussed in EFSA (2015). Earlier, Koriati et al. (1980) articulated three stages for elicitation of probability judgements from an expert: first, memory is searched for relevant information; second, evidence is assessed to arrive at a feeling of uncertainty; and third, the feeling has to be mapped onto a conventional metric. However, they recognised that an expert’s lack of experience in performing the internal mapping between feeling and a metric might lead to a corresponding lack of reliability, and/or incoherence, in the probabilistic expression of uncertainty. This chapter contributes a methodology consistent with an outside view which builds upon the initial stages of a probability elicitation but avoids the need for an expert to make an internal mapping. We aim to systematically support an expert to perform an appropriate mapping by grounding an analogy assessment in domain knowledge to select relevant empirical data for similar events which can then be translated into a defensible subjective probability distribution.

We begin by describing selected industry examples where both the need to express uncertainty about a quantity of interest and the opportunity for an expert to match the event to be predicted with an analogous pool of events exists. By abstracting from these examples, and by drawing upon theory from the wider literature, we present general steps for eliciting a subjective probability distribution using empirical data. The rationale and activities involved in each step are explained. Examples of implementing our approach illustrate how the general principles can be applied. We conclude by examining the benefits and shortcomings of our proposed approach to provide some insight on when it can be useful, when it might not be applicable, and issues to consider during implementation.

7.2 On the Nature of the Problem

7.2.1 Motivating Industry Challenges

Let us consider two industry contexts. Both examples are simplifications of real issues for which probabilities are required for variables within models developed to

support management decision-making. Here, we focus only upon issues related to the expression of the prior probability distributions.

First, consider a situation where a supply chain manager has procured a new supplier and wishes to assess the uncertainty in the true non-conformance rate of the parts to be supplied as an input to modelling quality related decisions (Quigley et al. 2018). The manager is uncomfortable making subjective probability assessments because the concept of quantifying some outcome that will in time be observable is cognitively challenging. But she is able to match the new supplier with similar existing suppliers since all have been subject to the standard procurement process. Hence the manager is relatively more comfortable in making analogy assessments between suppliers in terms of characteristics that might impact their performance. This judgement guides the creation of a relevant data set for existing suppliers providing a comparator pool that can be used to estimate a prior distribution. Of course, the uncertainty in the non-conformance rate of the new supplier represented by the estimated prior distribution should be checked for consistency with the beliefs of the supply chain manager.

Now consider a context where a new engineering design for an aerospace system is being developed as a variant of an earlier generation product (Walls et al. 2006). Typically the designers match the functionality of the new design specification and existing products to assess what aspects of the existing designs can be transferred. In addition, innovations relating to technologies, materials, processes and such like are introduced to create a new system design. The designers are asked to provide estimates of the probabilities associated with key failure modes of the new system design as part of a reliability analysis which in turn impacts the development budget. As in our first example, the designers are not entirely comfortable in expressing subjective probabilities. In part, this is because their mind-set implies designs are created to function not to fail hence thinking through negative outcomes is challenging. But also, because assessing probabilities arising from the myriad of scenarios across which uncertainty might be manifested is cognitively complex. Since the designers naturally match the functionality of the new system to analogous existing system designs we build upon this natural comparison to obtain our probability assessments of the failure of the new system to function as required. We take as our primitive for expert judgement the engineering relationship between the new and heritage system designs so that we can select relevant operational experience data from earlier generation products for the latter to obtain an empirical prior distribution for given failure modes of the former.

7.2.2 Generalisation of the Problem

Abstracting these two industry contexts allows us to establish three common features of our elicitation problem.

First, we consider situations in which we are effectively anticipating data about events that might be realised in future and for which there exists observed data for analogous entities. For example, the number of non-conformances in future

parts delivered by a new supplier or the number of failure events in the future operational use of the new system design. In each situation data are available for existing suppliers or systems, and a data set associated with the new supplier or system will become available, at least in principle if not also in reality.

Second, we can articulate a set of models to explain the variability in the anticipated data set. That is, the data set that comprises the future event history for the quantity of interest that does not yet exist but might be realised. This model family is indexed by parameters to describe the variability in the data generating process (DGP) associated with the event history. For our examples, a simple probability model for the DGP could be a Poisson distribution parameterised by the underlying true rate. For the supplier non-conformance and the new system development examples, the Poisson model describes the count of the non-conformances and the count of failure events per unit time parameterised by the true non-conformance rate and the true failure rate respectively. The true rate is not known with certainty, therefore we can represent the uncertainty in the parameter using a prior distribution if we follow a Bayesian approach.

Third, we require expert judgement to specify the prior probability distribution representing the uncertainty in the quantity of interest. For example, the prior distribution provides a set of plausible values representing the uncertainty about the true non-conformance rate of the new supplier or the true failure rate of the new system design. The challenge is to elicit a prior distribution so that it is meaningful and defensible, making appropriate use of expert judgement.

7.2.3 Implications of Inference Principles for Elicitation

If we approach elicitation from a Bayesian perspective then we are effectively asking an expert to map her beliefs about the quantity of interest onto a mechanism where the uncertainty is fully understood. This mechanism can be conceptualised by, for example, chips or a probability wheel (Spetzler and Stael von Holstein 1975), all of which translate to asking questions during elicitation to obtain an answer to a question such as ‘what is the probability of a non-conforming part being delivered by the new supplier?’. The elicitation intends to encourage an expert to think about a self-consistent betting regime. Take a simple probability wheel conceptualisation, as shown in Fig. 7.1. If an expert states there is a 50% chance that the next part delivered by the new supplier is a non-conformance then we could map this outcome to the white or black implying that an expert is effectively mapping her belief as a bet she is willing to take onto a mechanism whose stochastic characteristics are fully known.

But what happens when an expert more naturally makes analogies to her experience related to, say, past suppliers based on an assessment of similarity between characteristics believed to be influencing quality performance. Based upon the evidence of achieved performance for similar suppliers for whom empirical data are available, we can construct a class of plausible non-conformance rates for the new supplier. In this situation, an expert is essentially forming a comparator data set

representing the extent of her knowledge about the uncertainty in the true rate. More abstractly, we can say the expert needs to assess the characteristics of a DGP for the non-conformance rate of the new supplier so that a comparator pool of DGPs for which empirical data already exists can be identified. We argue that this matching of the DGPs for the new and similar existing suppliers represents the extent to which we can make reliable use of expert judgement. Achieving a match implies that the probability distribution representing the variation in the comparator pool allows us to empirically estimate the prior distribution for the true non-conformance rate.

Theoretically, we reason that if the comparator pool reflects the beliefs of an expert then, as the number of DGPs within the pool increases, the empirical distribution characterising the uncertainty in the quantity of interest will converge to the subjective prior distribution obtained through mapping to a probability mechanism that is fully understood; see Fig. 7.1. Practically, of course, constraints are likely to exist on the amount of experience which can be accumulated by an expert meaning that an infinite pool is infeasible which in turn implies that we lack complete understanding of the probability mechanism. If expert judgement is based on finite pools, or equivalently incomplete experience, then this leads us to question the general adequacy of a prior distribution elicited solely using subjective expert judgement. To address the challenges of some practical contexts, such as those discussed in our motivating examples, we propose an alternative approach that aims to make use of an expert’s judgement as well as relevant empirical data with the goal of eliciting a meaningful prior distribution for parameter uncertainty. Our proposed approach is grounded in the method of empirical Bayes inference.

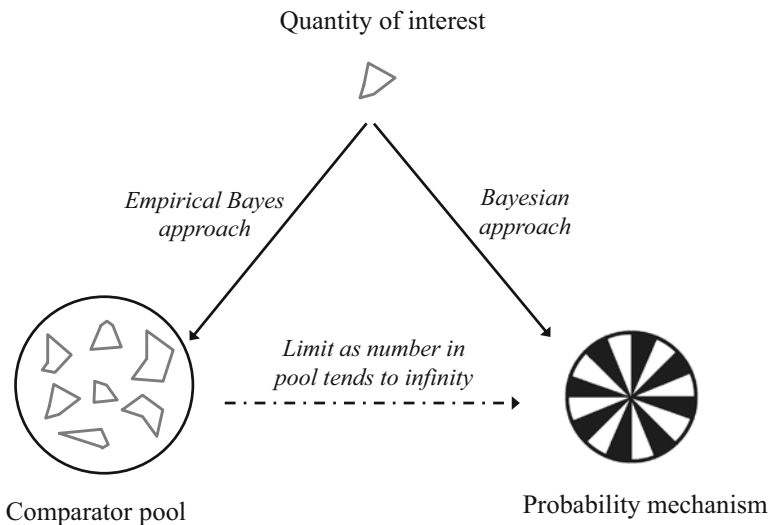


Fig. 7.1 Empirical Bayes and Bayesian reasoning for subjective probability elicitation

7.2.4 Principles of Empirical Bayes Inference

Figure 7.2 illustrates the concepts of empirical Bayes inference. Multiple data generating processes (i.e. the m DGPs) are required to form a comparator pool of data for the quantity of interest. Each DGP is described by a family of probability models for which empirical observations are available to support parameter estimation. We use the term family deliberately since the probability models are all of the same type (e.g. Poisson) but the parameter values of each distribution can differ to characterise the variation in each individual DGP. Importantly in our context, the empirical data across all DGPs are pooled to estimate the parameters of the prior distribution, which represents the variation in the comparator pool. For example, if the probability model family for the DGPs is a Poisson distribution parameterised by the non-conformance rate, then the empirical prior mean estimated by pooling data provides a point estimate of the true non-conformance rate of the new supplier while the full prior probability distribution characterises the uncertainty.

Although not the focus of this chapter, it is worth mentioning that Bayes theorem can be used to generate a posterior distribution by updating the prior distribution in light of empirical data for a given DGP, whether the DGP relates to the events for a new or an existing entity, such as a supplier. In general, the posterior estimate will be a weighted average of the comparator pool and the individual estimate, where the weighting depends on the degree of experience. Typically less weight is given to an individual and more weight to the pool for those DGP with limited histories, with greater weight given to an individual with more data.

In summary, empirical Bayes adopts the same basic steps as a Bayesian methodology by articulating a prior distribution and having the capability of updating

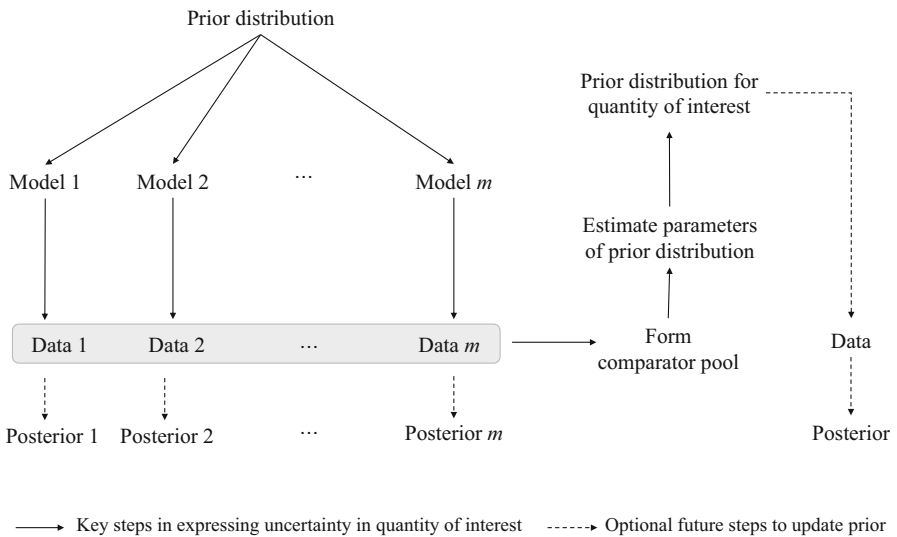


Fig. 7.2 Rationale of an empirical Bayes approach to obtain a prior distribution

this prior with data to generate a posterior distribution. The difference is that under empirical Bayes the prior distribution is estimated using observed data for a comparator pool while a full Bayes approach uses a subjective prior distribution. The roots of empirical Bayes reasoning can be traced to von Mises (1942), with Robbins (1955) formalising the terminology and providing the first serious study of the method within a non-parametric framework. Further details about empirical Bayes can be found in the seminal papers by Good (1965), Efron and Morris (1972), Efron and Morris (1973), Efron and Morris (1975), Efron et al. (2001). While Carlin and Louis (2000) as well as Efron (2012) provide introductory texts.

7.3 General Methodological Steps

We propose a five step approach to obtain the prior distribution using relevant empirical data, as shown in Fig. 7.3.

7.3.1 Characterise the Population DGP

We begin by identifying those factors characterising, what we call, the population DGP. This is the process generating the anticipated data or future events for the quantity of interest. This is an important step because it defines the criteria by which data sets (i.e. the sample DGPs) are subsequently selected for inclusion in the comparator pool used to construct the prior distribution.

The characterisation of the population DGP should be driven by problem domain experts, suitably facilitated by an analyst. An expert has an important role in this step because it is the expert who possesses substantial accumulated understanding of what is likely to influence the realisation of events and, with the support of the analyst, articulates the factors to provide the basis for similarity matching.

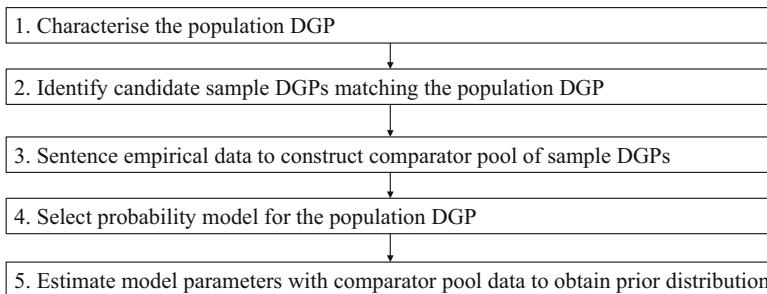


Fig. 7.3 Key steps in constructing a prior distribution using empirical data

The way in which we might approach characterisation of the population DGP can be considered a specific instance of the ideas inherent within the wider context of statistical sampling. Much has been written in the literature on, for example, survey sampling about the need to identify appropriate factors to define population characteristics to support sound inference and to share insights into how such factors might be identified within applications. See, for example, Cochran (1975).

There has also been consideration of this issue within the context of the so-called reference class problem (Reichenbach 1971). This is concerned with classifying an event such that appropriate data can be used to infer probability. According to the Oxford English dictionary a reference class within the context of probability theory and the philosophy of science is “*the class of entities sharing a property with respect to which a theory or a statement of probability is framed*”. Cheng (2009) has examined the challenges associated with identifying reference classes in legal practice where it is acknowledged that a finite number of possible (i.e. sample) DGPs exist for a given case and a key question is “*how does one choose the comparison group?*”. Although the paper is framed as inference within an adversarial context, with perhaps information asymmetry between the two opponents, many points raised have more general currency providing examples of defining the population characteristics, including some where an apparent lack of consideration of the appropriate factors to define the population DGP has resulted in misleading inference.

7.3.2 Identify Candidate Sample DGPs Matching Population

The argument underpinning the approach proposed by Cheng (2009) is that “*reference class-style reasoning is equivalent to using a highly simplified form of regression modelling*” where the factors characterising the population DGP can be switched on/off for candidate data sets effectively providing a means of making a relative assessment of relevance against a set of criteria. Cheng (2009) also points out that in practice the goal is not to find the optimal class but simply to find the best available data sets to make reasonable and timely inference. This is an important point in relation to our purpose since we are also likely to be constrained by the availability of a finite number candidate data sets. Also, unlike other types of statistical sampling, we are not in a position to collect primary data to match the characteristics of the target population. Instead we are matching the characteristics of the population DGP with secondary sources of data that already exist. Hence a formal means of matching the similarity between the population and the candidate sample DGPs (i.e. available data sets) in terms of the factors characterising the former is important and necessary.

7.3.3 *Sentence Empirical Data to Construct Sample DGPs*

Once the preferred data sources have been selected, the events recorded should be scrutinised in collaboration with a domain expert to assess the representativeness of the records for the type of experience to which our anticipated DGP will be exposed and to sentence these records, if required. The intention is to create a data set that is not only appropriate in terms of its similarity matching to the population characteristics but is also relevant in terms of the events and the circumstances under which these have been realised. The nature of sentencing can vary with application contexts and the associated modelling. For example, sentencing can include selecting records for events within a sub-set of the realised data set to form a representation of the anticipated experience or simply to screen out events that have been realised under unusual circumstances that are not representative.

More formally, we can reason through an assessment between the population and sample DGPs as follows. Although the data sets formed to create candidate sample DGPs are heterogeneous with respect to their stochastic characteristics (e.g. means and standard deviations), the expert should not be able to meaningfully discriminate between these DGPs based on any information other than their realisations. Care must be taken with the data analysis since the realisations within any DGP will be correlated as belonging to the same DGP, but the sets of event data records between DGPs are assumed independent. Confirming the suitability of the data records essentially requires checking that the predictive distributions for each DGP are independent and identically distributed. This can be achieved by conceptualising as a comparison of order statistics. Let $_jX_{i:n}$ denote the i th smallest value from a sample of n records from the j th DGP where $j = 0$ denotes the DGP associated with the quantity of interest for which an estimate of uncertainty is to be made. The comparator pool of sample DGPs will be appropriate, if an expert can confirm that based on the covariate information only the following statement is true:

$$\Pr(\text{Min}({}_0X_{i:n}, {}_1X_{i:n}, \dots, {}_mX_{i:n}) = {}_jX_{i:n}) = \Pr(\text{Min}({}_0X_{i:n}, {}_1X_{i:n}, \dots, {}_mX_{i:n}) = {}_kX_{i:n}),$$

$$\forall i, j, k, n.$$

In words, based on the reference factors used to characterise the population DGP, the minimum of any order statistic is equally likely to be generated from any of the sample DGPs; this is true for all order statistics and for all possible sample sizes. Practically this implies that when assessing the order statistics an expert may simply reflect upon whether the extremes and the typical values of the comparator data sets are appropriate for the quantity of interest.

7.3.4 Select Probability Model for Population DGP

The family of probability models considered suitable for describing the population DGP will be largely determined by the context so that it supports suitable inference, not only in terms of meaningful parameters but also in terms of mathematical and computational implementation.

For the two motivating examples, we indicated that the Poisson distribution is an appropriate simple model to describe the variation in the count of events and hence capture the aleatory uncertainty as the within-process variation. Since the prior probability distribution predicts the epistemic uncertainty in the true rate then choosing a conjugate parametric form leads naturally to the Gamma distribution to model the between-process variation across the comparator pool of sample DGPs.

Therefore it is important to select a model family that allows coherent representation of the variation both within and between the DGPs to articulate both the aleatory and epistemic uncertainties, even though it is the latter that is of primary interest to us in the elicitation context.

7.3.5 Estimate Model Parameters to Obtain Prior Distribution

Statistical inference to estimate the parameters of the model for the population DGP can be conducted using standard approaches such as Maximum Likelihood or Method of Moments (e.g. Klugman et al. 2012). The mathematics of the inferential procedure will depend upon the parametric form of the probability models. For example, Quigley et al. (2007) provides mathematical details of the statistical inference methods for the Poisson-Gamma model family.

The parameter estimates obtained using the data in the comparator pool formed from the sample DGPs allow the prior probability distribution to be fully specified.

7.4 Example Applications of the Elicitation Process

Two examples are presented. Both relate to industrial applications of risk and reliability analysis for which the quantity of interest relates to the frequency of events over time. We have deliberately selected examples where related probability models are chosen for the population DGP since it allows us to show how different application considerations give rise to adaptation and customisation of the general methodology. Each example presents distinct challenges in relation to the characterisation of the population DGP, the identification and sentencing of empirical data to create sample DGPs, and the method selected to estimate model parameters for the given the probability models. We present the examples in order of their relative complexity of the emergent elicitation issues. For this reason we focus our discussion on the distinctive elements of each example even though the elicitation for both examples did require careful consideration of each step.

7.4.1 Assessing Uncertainty in Supplier Quality

A project aimed to model the risk in supplier performance for a manufacturer of complex, highly engineered systems reliant on an extensive, international supply chain for parts and sub-assemblies. The modelling problem under consideration involved supporting decisions about whether, or not, to develop a supplier given only information gained about quality from company standard contracting and procurement processes. Quigley et al. (2018) describe the wider modelling methodology and results. Here we focus upon the elicitation of the subjective distribution representing the uncertainty in the quality performance of the new supplier, where quality is measured by the true non-conformance rate associated with parts delivered from the supplier to the manufacturer.

7.4.1.1 Characterise the Population DGP

To characterise the population DGP, we need to identify the reference factors in partnership with a suitably qualified expert. Taking an expert to be a person(s) with substantive experience in relation to the event for which uncertainty is to be assessed, then the natural set of experts for this problem are those staff within the manufacturing company with qualifications and experience in managing the supply chain and production operations.

As is common more generally (e.g. Slack et al. 2016), the manufacturer organises its parts supply base into coherent commodity groups each of which correspond to classes of technologies and processes. Such a classification allows managers to share the responsibility for the procurement and development of a set of suppliers within a given market. Importantly, it also implies that the manufacturer has already considered classification of parts in terms of common factors that are likely to influence the nature of the functional specification and hence the opportunity to conform (or not) with that requirement as a consequence of the type of part being supplied.

Much has been written about the types of factors affecting supplier quality and the risks associated with supply chain performance (Nagurney and Li 2016; Sodhi and Tang 2012; Talluri et al. 2010; Zhu et al. 2007). Hence secondary information about the possible types of factors which might influence the new supplier quality performance is available to the analyst leading the elicitation. Such information can be useful in preparing to elicit those factors which are considered by an expert to be influential for the case under consideration.

So who is our expert and how do we identify the factors believed to be important in characterising the population DGP? Our expert is a supply chain manager who possesses the experience of the day-to-day management of the supply base within the company as well as wider expertise in managing operations in similar organisational contexts. In this sense our expert is suitably qualified to share his knowledge and experience during the elicitation. Through multiple conversations

taking the form of semi-structured interviews, supported where appropriate with diagramming techniques, we have surfaced the expert's beliefs about influencing factors and the relationships between them. Factors identified include the nature of the part technology, design, production and shipping, including type, complexity and scale, as well as the nature of the supplier experience, capability, capacity and location.

7.4.1.2 Identify Candidate Sample DGPs Matching Population

Next we identify empirical data sets in terms of their match to the population characteristics as defined by the reference set of influential factors.

To manage operations, the manufacturer has databases containing empirical records associated with supplier and part details as well as their transactional data for events related to the placing and receiving of parts ordered for engineering projects, including the quality of parts received at goods inwards. More generally, such databases or enterprise resource planning (ERP) systems are core to managing operations (Gallien et al. 2015). They can be extensive both spatially, in terms of part/supplier coverage, as well as temporally, given the dynamic nature and scale of manufacturing production. This means that in terms of matching and subsequent sentencing of empirical data sets, we need to consider the records to be used in terms of both coverage of 'similar' event histories for suppliers and also the relevant time window in order to obtain a reliable predictor of the uncertainty in the true non-conformance rate of the new supplier.

In our application, our choices about possible matches to the population DGP includes event history data for a super-set of all suppliers, a set of suppliers within the commodity group to which the new supplier belongs, a sub-set of this commodity group defined by those suppliers/parts possessing common identifiable factors. We have used the commodity group data as the basis for our candidate samples from the population DGP because this best matches those factors believed by our expert to most influence the supplied part quality. The commodity grouping confounds the influential effects of part technology and processes on the opportunity to deviate from conforming to functional specification. We discounted the other two alternatives mentioned for the following reasons. Using a super-set of all suppliers mixes multiple groupings each with different degrees of opportunity and so would tend to overestimate the uncertainty in the true non-conformance rate of the new supplier. Using a sub-set of suppliers within the commodity group might underestimate the uncertainty since the reduction could only be based on recorded factors such as geographical location, which experts judge to be less influential than other factors such as supplier production capacity and loading which are not directly observable.

While our decision to select particular data sets has been based upon the judgement elicited from the domain expertise of the supply chain manager, we have also been able to explore the degree of historical influence of certain recorded factors on the variation in the observed non-conformance rate of existing suppliers using,

for example, regression modelling. Although not an exhaustive analysis since the covariate information is incomplete, the findings of such data analysis can help us to challenge and to elicit judgements from an expert.

The choices we make in selecting data sets will ultimately affect the number of sample DGPs we use to estimate the prior distribution. For example, taking the commodity group of 35 suppliers as a baseline, then by definition there will be more (less) candidate sample DGPs in the super-set of all suppliers (sub-set of the commodity group). Obviously, the number of sample DGPs, as well as the amount of data in each, will impact the degree of sampling error and hence inference.

7.4.1.3 Sentence Empirical Data to Construct Sample DGPs

Having selected the candidate data associated with the existing suppliers, we now require to finalise the set of event records for each supplier in order to form the comparator pool of sample DGPs to be used for inference.

Two types of data sentencing are needed. First, to choose the relevant records from past event data. Second, to cleanse the selected records to deal appropriately with any anomalies whether they arise due to data recording errors or unusual circumstances affecting the suppliers. The latter is standard statistical data preparation, therefore we focus discussion on the former.

Since the purpose of selecting the data records is to form a distribution representing the uncertainty in the unknown true non-conformance rate, we need to consider historical events for existing suppliers only insofar as they are likely to be reliable predictors of the future for the new supplier. Hence again expert judgement will be vital in assessing the relevance of choosing data from different time horizons. In our application, data are recorded daily but management reports use summaries on weekly, monthly, annual windows associated with different purposes bringing a tendency for the expert to anchor upon conventional time frames. Given the length of our engineering procurement projects, which last several years during which there is turnover in the supplier base, we elect to use time windows defined on annual basis on our initial sentencing of the data. Of course, there can be a tension between the relevance of the time windows selected and sample size given that focussing on the recent past implies a shorter sampling history than had we chosen a longer time horizon. However this is a trade-off that needs to be made since relevance of the selected events over time is preferred to simply more event data per se.

In our application we agree upon a data set to represent the sample DGPs that includes the number of non-conformances over the specified annual time intervals for 35 similar suppliers. Although not shown in its raw form, there is a degree of heterogeneity in the data from the comparator pool and this is used to capture the distribution from which the new supplier's 'future experience' can be considered to be randomly selected.

In assessing the order statistics between the DGP associated with the true non-conformance rate of the new supplier and the candidate sample DGPs for the existing suppliers, it is sufficient to assess whether the minimum rate for the new

supplier is equally likely to be from any of the existing suppliers, if we are assuming a Poisson-Gamma probability model. However, to assess the parametric distributional assumptions requires the expert to reflect upon the order statistics more fully as described in Sect. 7.3.3. For example, if an expert identifies that one supplier is much more volatile than another but each have similar median performances, then this would indicate the distributional assumptions are in question.

7.4.1.4 Select Probability Model for Population DGP

We use a Poisson-Gamma probability model because it provides a flexible family capable of representing a wide class of patterns of uncertainty and, as a conjugate of the Poisson, computations are easily supported (Carlin and Louis 2000). Given the prior is estimated empirically it is also possible to check the statistical fit of this assumed model family by, for example, comparison of the observed and expected percentiles of the fitted predictive distribution.

Figure 7.4 shows an annotated version of the empirical Bayes approach, originally given in Fig. 7.2, for this supplier non-conformance rate application.

More formally, denote the number of non-conformances $N_i(t_i)$ accumulated by time t_i for the i th supplier to be conditionally independently Poisson distributed with mean $\lambda_i t_i$. We follow an empirical Bayes methodology, whereby a two stage hierarchical model is assumed, such that the rate for each supplier, $\Lambda_i, i = 1, 2, \dots, m$, is treated as though independent and identically distributed (i.i.d.) from a continuous prior distribution, the form of which is assumed to be Gamma with shape parameter α and scale parameter β :

$$\Lambda_i \stackrel{i.i.d.}{\sim} G(\alpha, \beta)$$

$$N_i | \Lambda_i = \lambda_i \stackrel{indep}{\sim} Po(\lambda_i t_i).$$

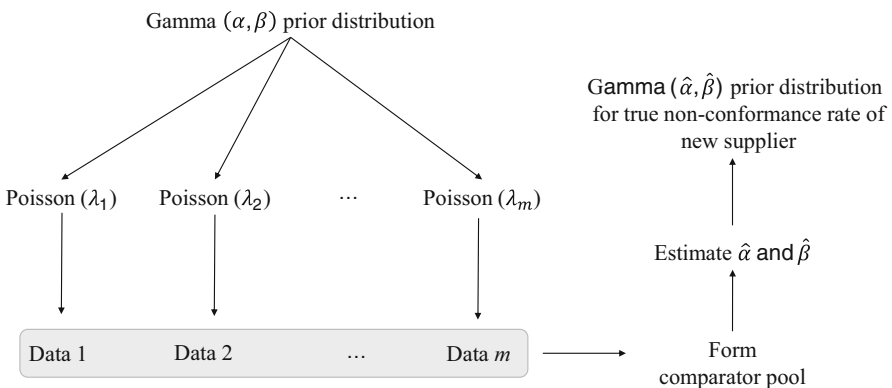


Fig. 7.4 Empirical Bayes reasoning for the Poisson-Gamma probability model for supplier non-conformance rate

7.4.1.5 Estimate Model Parameters to Obtain Prior Distribution

The parameters of the prior distribution, α and β , are estimated using the empirical data in the sample DGPs by calculating the predictive distribution, which then forms the basis for the likelihood function for the model. For our Poisson-Gamma model the predictive distribution takes the form of the Negative Binomial distribution (Greenwood and Yule 1920):

$$P(N_i(t_i) = n_i | \alpha, \beta) = \int_0^\infty \frac{(\lambda_i t_i)^{n_i} e^{-\lambda_i t_i}}{n_i!} \frac{\beta^\alpha \lambda_i^{\alpha-1} e^{-\beta \lambda_i}}{\Gamma(\alpha)} d\lambda$$

$$= \frac{\Gamma(n_i + \alpha)}{\Gamma(\alpha) n_i!} \left(\frac{\beta}{\beta + t_i}\right)^\alpha \left(\frac{t_i}{\beta + t_i}\right)^{n_i}, \quad \alpha > 0, \beta > 0, t_i > 0, n_i = 0, 1, 2, \dots$$

Following Arnold (1990), a likelihood function for the data can be constructed by taking the product of the predictive probability functions for the i th supplier evaluated at each of the associated realisations of non-conformance events for that supplier:

$$L(\alpha, \beta) = \prod_{i=1}^m \frac{\Gamma(n_i + \alpha)}{\Gamma(\alpha) n_i!} \left(\frac{\beta}{\beta + t_i}\right)^\alpha \left(\frac{t_i}{\beta + t_i}\right)^{n_i}.$$

Thus the Type 2 (Good 1976) Maximum Likelihood Estimators (MLE) of the pool parameters, denoted by $(\hat{\alpha}, \hat{\beta})$, can be obtained as confidence regions for the parameters.

Figure 7.5 shows the form of the Gamma prior Probability Density Function (PDF) obtained for the data in our comparator pool of sample DGPs and an

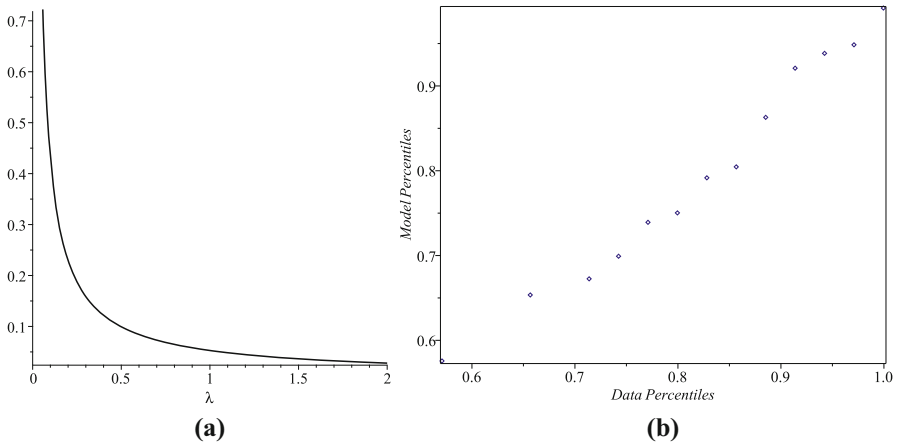


Fig. 7.5 (a) Estimated prior PDF for λ , the true non-conformance rate of the new supplier and (b) Fit of the Poisson-Gamma model to comparator pool data based on the predictive distribution and empirical percentiles

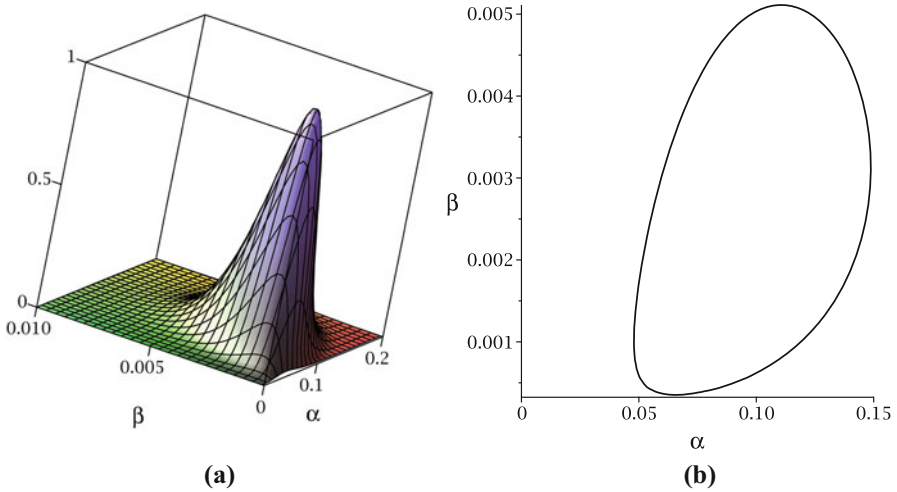


Fig. 7.6 (a) Relative Likelihood function and (b) 95% confidence region for pool parameters

assessment of the statistical adequacy of the probability model for the data used. The prior distribution in Fig. 7.5a has a long right tail with a prior mean of nearly 50 non-conformances per unit time. The plot of observed and expected percentiles based on the predictive distribution in Fig. 7.5b indicates a reasonable fit of the Poisson-Gamma model to the empirical data given that the points fluctuate around the 45° line. Thus the data used is consistent with the probability model selected for the population DGP.

Figure 7.6 illustrates the relative likelihood function and the associated 95% confidence region for the pool parameters estimated from the 35 sample DGPs. In Fig. 7.6a the peak corresponds to the Maximum Likelihood Estimates (MLE) and is assigned a value of one from which the likelihood of any combination of parameter values are measured relatively. Figure 7.6b shows that, based on the 95% confidence region, α is between 0.048 and 0.148 while β is between 0.00035 and 0.00511. Moreover, the parameter estimates are not independent since some the coverage of some pairings are not within the confidence region, in particular the high values of α coupled with the low values of β .

Figure 7.7 provides a pointwise 95% tolerance interval for the prior Cumulative Distribution Function (CDF). The long right tail of the distribution is evidenced by the steep climb of the CDF followed the relatively flat growth. The MLE of the CDF provides an estimate of the probability that the true rate, λ , is less than a specified value. For example, although not discernible in the plot, there is a 0.4 probability that λ will be less than 0.01. More apparently, as λ increases to 0.1 and again to 1 then the corresponding cumulative probabilities rise to 0.49 and 0.6, respectively.

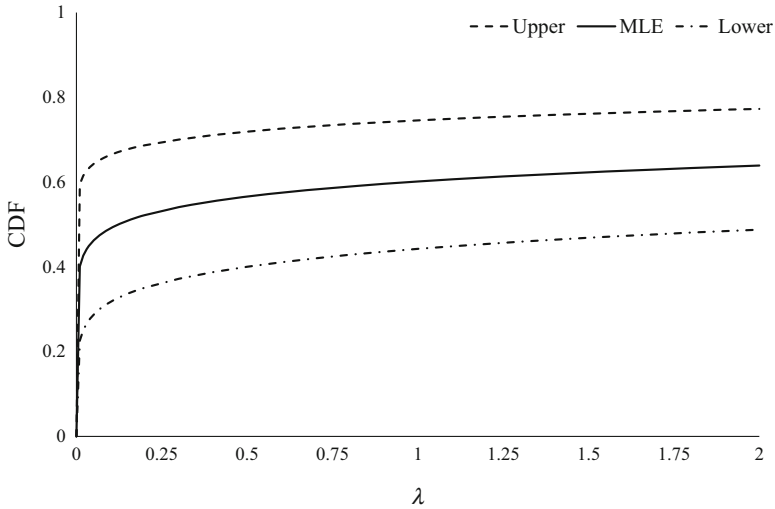


Fig. 7.7 Piecewise 95% tolerance interval for the CDF of new supplier true non-conformance rate

More interesting is the width of the tolerance intervals. For example, over the range of λ values between 0.01 and 1 the width remains relatively constant between 0.36 to 0.30, respectively; that is, approximately one third. This has implications for the degree of uncertainty in the prior distribution. In this example, our tolerance intervals allow us to acknowledge the uncertainty in the prior distribution estimated from the 35 sample DGPs in the comparator pool.

Although the prior distribution is estimated empirically using the data selected by expert judgement, we also require the expert to assess whether the estimated prior distribution adequately represents his beliefs about the uncertainty in the true non-conformance rate of the new supplier. In our application we make such an assessment by providing the expert with visual feedback, say, in the form of the prior distribution plot. Since the methodology has been developed for estimating the prior parameters from data, it is also relatively easy to provide alternative prior distributions based on different selection of data, as might correspond to different matching of sample DGPs from the population. For example, we have shown the findings based on selection of a pool of suppliers from the same commodity group. But it is also possible to generate equivalent plots using, say, the super-set of all suppliers or a sub-set of commodity group suppliers to present an expert with alternative prior distributions representing different degrees and patterns of uncertainty. Such a tactic provides a form of internal consistency checking between the explicit reasoning about the influential factors affecting the uncertainty in the non-conformance rate and the representation of these beliefs in the form of a prior probability distribution.

7.4.2 Assessing Uncertainty About Reliability of an Engineering Design

Our second example is based on a project to model the reliability of an engineered unit during its design and development phase. The unit will be part of a new generation aircraft. The ultimate purpose of modelling is to support decisions about the efficient allocation of resources to grow the reliability performance of the unit design to meet its required specification (Walls and Quigley 1999; Walls et al. 2006; Johnston et al. 2006; Wilson and Quigley 2016). The modelling approach adopted requires elicitation of the sources of uncertainty regarding any design weaknesses and the time to their realisation as failures if not removed or mitigated. The design under consideration is a variant of an established product family and so the manufacturer has extensive operational data on performance of earlier generations of the unit type. Such data contains information about all life events for each unit within a fleet, including entry into service, failure and maintenance events.

In order to identify possible weaknesses of the new unit design, structured expert judgement is elicited from relevant engineers to both express their concerns and to quantify the uncertainties about the existence of these concerns as subjective probabilities. The process supporting this subjective elicitation is given in Walls and Quigley (2001) while reflections on the practice of implementation are given in Hodge et al. (2001). Specifically for this example, a representative selection of thirty engineers have been interviewed, including designers, programme managers, as well as specialists in components, environmental test, procurement, and manufacture. These engineers have identified their concerns and assessed their chance of occurrence in system operation resulting in a subjective Poisson prior distribution with means ranging from approximately 3–11 across different classes of engineering concern.

Our focus in this chapter is upon the expression of an empirical prior distribution for the epistemic uncertainty associated with the time to realisation of engineering concerns as failures which can be estimated from relevant observational data from variants of the unit design already in service. A different expert to those involved in sharing engineering judgement about the nature of concerns is involved in providing judgement about the selection of the empirical data to be used to model the failure occurrences within specified time intervals. The expert working with the analyst to develop the empirical priors assumes a more systems level view of the new unit than those engineers who had provided judgements about the nature of epistemic uncertainties in relation to the concerns about the new design. The expert assuming the role in empirical data preparation is an experienced technical engineer with a breadth and depth of experience of the product family. Earlier he supported the facilitation of the subjective judgement from domain experts about design concerns from a systems perspective and so provides a link in interpreting the engineering detail about specific design issues with the observational data available for product families.

Quigley and Walls (2011) describe the full methodology for combining the subjective prior distribution on engineering concerns with empirical prior distributions to support reliability growth decision making. Here we consider the application steps in constructing the prior distribution only.

7.4.2.1 Characterise the Population DGP

The nature of the engineering concerns are pivotal to the characterisation of the population DGP since these concerns capture the potential for types of failure to occur due to a mismatch between the conceptual design ‘strength’ and the ‘stresses’ to which it will be exposed. Engineering concerns may relate to, for example, choices about electronic component rating, material characteristics, manufacturing processes, topology and so on. More generally concerns relate to aspects of the design, manufacture, operation and maintenance where opportunities for stressors to challenge the intended functionality of the unit might arise.

There are, of course, more tangible factors that might characterise the population DGP in the form of the specified requirements of the unit design. Such requirements will articulate the function, environment, duration as well as other influential features of the design specification. It is based on such factors that design engineers might select a base design from an existing product family in order to develop a new variant (Pahl and Beitz 2013). While such factors can also aid characterisation of the population DGP, it might be too naïve to consider them only since they effectively represent the factors that drive the choices of the designers in developing a new unit. It is the consequences of these design and other choices in engineering the unit that give rise to concerns.

In essence, the concerns represent the epistemic uncertainties of the engineers about the ability of the new design to function as intended in its operational environment. The nature of how the concerns will be realised as failures provides a means of characterising a sub-population DGP which is needed because each type of concern will be associated with a distinct pattern of realisation. For example, if the electronic components are insufficient for the operating stresses then this concern is likely to be realised early in service as a form of shock failure, whereas material characteristics may imply a faster rate of degradation than intended, resulting in a failure later in service but before the anticipated lifetime of the unit and so sooner than desired.

The elicitation of engineering concerns is very important because it allows us to understand the possible effects of failures that might be realised due to the causal reasoning from design choices through to operational functioning. This understanding allows us to define the reference factors relevant to each class of concern in terms of their temporal realisation as failures and so specify the characteristics of the population DGP at a sub-population level.

7.4.2.2 Identify Candidate Sample DGP Matching Population

As mentioned, the company has operational data on life events for related products within the unit family. For earlier generations of the unit design, no elicitation of engineering concerns had been formally conducted although other forms of reliability analysis are available which provide insight into anticipated failure modes and why these did or might have occurred. To identify our candidate data sets we need to consider the concerns elicited for the new design and the equivalent data for past heritage designs given the relative similarity between design variants in terms of the consequences of the choices about externalities of function, form and environment so that we understand the relative opportunities for vulnerabilities to exist and to be experienced.

In our application, we identify several existing unit designs for which there are data sets offering candidate sample DGPs. However, there is not a one-to-one similarity match between the full set of concerns, and the reasons for these concerns, between the new unit design and the existing units. This is not unexpected given we have characterised the population DGP at the sub-population level. In this context, the sources and coverage of our data sets for the candidate sample DGPs can vary for different sub-populations depending on the classification of the engineering concerns.

Figure 7.8 summarises the principles underpinning the formation of the sample DGP for this example. Each concern class in the centre of the diagram represents a sub-population DGP defined in terms of common reference factor settings for the engineering concerns. The links between the individual concerns and the classes

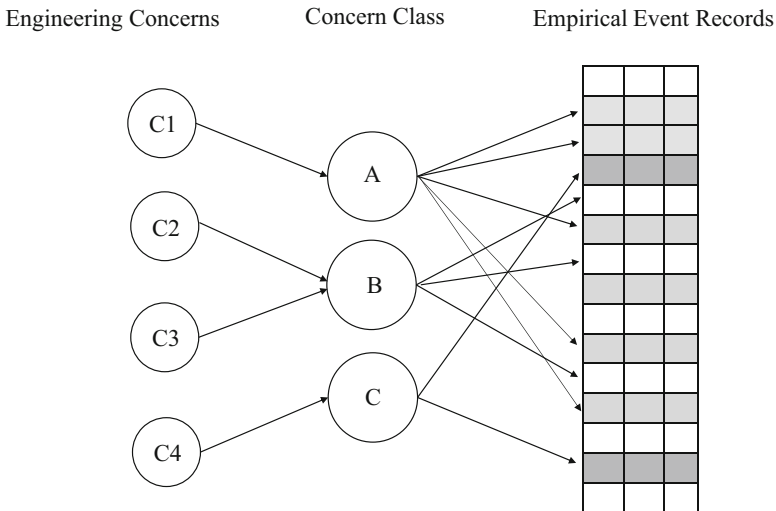


Fig. 7.8 Conceptual relationship between engineering concerns for new unit design and event history records for related designs

represent the mapping between the engineering judgement and their expression as reference factors for the similarity matching with heritage unit designs which have empirical data. The data records shown on the right side are in the form of event histories with rows corresponding to accumulated time to an observed failure and columns giving covariate information, such as heritage unit and failure mode. The shading of the records indicates the distinct data sources selected for different heritage units that are candidate sample DGPs. The links between the concern classes and individual events represents the failure data records to be included in the sample DGPs. If a class contains only one concern for which there is a match between the relevant reference factors and the observed failure event codings, then this can be conceptualised by mappings such as $C1 \rightarrow A$ and $C4 \rightarrow C$. A class might be formed if individual concerns can be meaningfully grouped in terms of their likely pattern of realisations through time, as shown by the mapping $C2 \cup C3 \rightarrow B$. For example, specific unit build vulnerabilities can be grouped together if the pattern of realisation of the resulting shock failures due to manufacturing issues are assessed by the engineering experts to be the same.

7.4.2.3 Sentence Empirical Data to Construct Sample DGPs

So far we have built up our argument in terms of defining our population DGP in order to match suitable sample data sets. For this reason we show the mapping from concerns to classes to empirical records in Fig. 7.8. However, as we acknowledged earlier, sentencing data is a craft built upon scientific principles, hence we also need to take into account the state of the empirical data sets into consideration during the process of constructing the sample DGPs.

In our example, the empirical data sets included the individual unit reference, accumulated flight hours, date stamp, number of flight cycles, type of aircraft, operator, fault type, failure mode code, failure effect code, text description of event occurrence, amongst others. Having data that describes the context, the nature and a classification of that event is not atypical in a reliability engineering context (e.g. Cooke 1996). In particular, the classification of events is embedded in the manner in which much historical failure event data has been stored and shared both within organisations and at industry sector levels (e.g. Rausand and Hoyland 2004). Although it can be convenient to use the standard classes within the empirical data set to define the classes of engineering concerns, we urge caution in simply automatically back-fitting. It is important to define classes grounded in the nature of the engineering concerns for the new design for which the prior distribution, and ultimately the reliability, are assessed. Even with historical data, rich information can be found in event descriptions to form sets of records that match to appropriate classes, which might be a sub-set of the standard grouping of events. As an example, consider a situation with two distinctive engineering concerns articulated in relation to some electronic components in the unit. One concern might relate to the geometry of one component's siting and another concern might relate to the material properties of another component. These concerns, should they exist as real

problems, are reasoned to be realised in different ways since the former will be likely to occur more quickly as it will be vulnerable to operating stresses within a flight cycle, while the latter might be realised more slowly since events are more likely to occur as experience is accumulated between flight cycles. The empirical data set categorises all events related to the electronic components together and so mixes the time to failure distributions that relate to the concerns. If sufficient information is available from the textual description then the records within the electronics components categories can be partitioned into more appropriate classes that better match the population characteristics of the concerns.

It is possible that using empirical records from past units to assess the times to failure of some engineering concerns is judged to be inappropriate by the engineering experts. This might occur when there are novel aspects of the design for which reasoning through the physical science of the failure mechanisms might provide a better assessment of uncertainties. Within the context of probabilistic risk analysis for engineering design, Fragola (1996) introduces the notion of “*tolerance uncertainty*” which relates to this issue. Tolerance uncertainty corresponds to an engineering expression of the relevance of historical failure data in relation to an anticipated failure mode for a new design so that credible choices are made about the selection of relevant data for analysis. Following this logic, we are effectively arguing that if the empirical data on observed events for related designs are judged by the engineering experts to be tolerable assessments of the anticipated occurrences of failures due to an engineering concern for a new design then the empirical data can be selected to form the prior distribution. However, empirical data should not be used if it is judged by the engineering experts to be intolerable since this implies an alternative source, such as subjective assessments of uncertainty based on understanding of the underlying science supplemented by engineering analysis and test data, are arguably more justifiable.

Focusing upon the use of empirical data only, then like our first example, choices also need to be made about issues relating to the boundaries of data in terms of time and coverage as well as treatment of data anomalies. In this example we need to consider the inclusion or exclusion of data from particular units within the fleet for the existing design that is to be used to inform the prior. Some units might be spares and so experience long periods in storage followed by short periods of intensive use and so have unusual operational profiles compared with the majority of units which will be operated on aircraft in very similar flight patterns. Also, choices need to be made about the time windows over which empirical data will be selected. In this context there will be considerable relative stability for long periods given the nature of the certification and operational use of aircraft, however there can be scheduled upgrades which roll out part design changes across the fleet and so should be taken into consideration if it affects particular engineering concerns.

For our example, we have used an empirical source data relating to over 400 heritage units and extracted records relating to events occurring over several years. For this stage of the modelling we work closely with the engineering expert who possesses the expertise and extensive experience in the design process and technology together with the responsibility for managing the reliability development

programme. The empirical data selection and sentencing is led by the analysts who drive the methodological approach but the choices made are based upon the judgement of our expert. Ultimately we have created a data set containing the times to first occurrence of events within each of eight classes relating to the engineering concerns surfaced.

We also partition the operating time horizon into five intervals with natural break-points corresponding to the accumulated flying hours at nominal inspection periods associated with units of different ages. This choice was made for a combination of engineering and modelling reasons. The engineers are most interested in the likelihood of failures occurring during stages of a unit life, while the analysts are thinking ahead to candidate probability models which will be consistent with the data and the wider purpose of analysis. Further, since we have partitioned time into five mutually exclusive intervals, the expert has to assess the equivalence of the probabilities for each interval as a means of operationalising the assessment of the equivalency of the order statistic distributions for each DGP in the comparator pool, as described in Sect. 7.3.3.

7.4.2.4 Select Probability Model for Population DGP

The reliability in this example is taken to be a measure of the duration of unit failure free operating time and is parameterised by both the engineering concerns and their time to realisation. More formally we can write this as follows. Let J denote the number of concern classes, N_j represent the number of concerns in class j that will be realised as failures and let I denote the number of mutually exclusive and exhaustive partitions of the distribution of times to realisation of concerns. Then the prior distribution is sought on the $(I \times J)$ matrix, denoted by \underline{P} , whose (i, j) element, denoted by p_{ij} , represents the probability that a concern associated with class j will be realised in the i th epoch. Hence the probability that a unit will not fail by time t_0 , denoted by T_u , conditioned on the matrix \underline{P} , and the vector $\underline{N} = \{N_1, \dots, N_J\}$ is given by:

$$P(T_u > t_0 | \underline{N}, \underline{P}) = \prod_{j=1}^J \left(1 - \sum_{i=1}^{t_0} p_{ij} \right)^{N_j}.$$

A multinomial distribution provides a simple and reasonable model to describe the sampling variation in the number of failures within time partitions of the concern classes. Each interval is assigned a parameter to measure the chance that a failure arising due to a concern would be realised in that time interval and the set of probabilities for any failure class are constrained to lie within a simplex. Further, the vectors of probabilities across classes are assumed to be independent and be Dirichlet distributed. We seek the empirical prior on these Dirichlet distributions; one for each class.

7.4.2.5 Estimate Model Parameters to Obtain Prior Distribution

A likelihood function to obtain Type 2 MLE for a concern class can be derived by first taking the product of all multinomial distributions for each sample DGP in the comparator pool and subsequently taking the expectation with respect to the Dirichlet prior.

Let m_{ik} denote the observed number of failures realised in time period i from the k th sample DGP created after sentencing the relevant historical data and let \underline{M} denote the corresponding matrix of data. The likelihood function for the k th DGP, which is a function of the vector $\underline{P}_k = (p_{1k}, \dots, p_{Ik})$, can be expressed as:

$$L_k(\underline{P}_k) = \binom{\sum_{i=1}^I m_{ik}}{m_{1k}, \dots, m_{Ik}} \prod_{i=1}^I p_{ik}^{m_{ik}}.$$

Following Ng et al. (2011), we assume the conjugate prior of the multinomial distribution to be the Dirichlet distribution of the form:

$$\pi(p_1, \dots, p_I) = \frac{\Gamma\left(\sum_{i=1}^I a_i\right)}{\prod_{i=1}^I \Gamma(a_i)} \prod_{i=1}^I p_i^{a_i-1}, p_i \geq 0, \sum_{i=1}^I p_i = 1, a_i > 0.$$

By taking the expectation of the likelihood equation with respect to the Dirichlet prior distribution, the new likelihood is obtained as a function of the parameters in the prior distribution and is given by:

$$L(a_1, \dots, a_I) = \prod_{k=1}^K \frac{\Gamma\left(\sum_{i=1}^I a_i\right)}{\prod_{i=1}^I \Gamma(a_i)} \cdot \frac{\Gamma(a_i + m_{ik})}{\Gamma\left(\sum_{i=1}^I a_i + m_{ik}\right)}, a_i > 0$$

from which the Type 2 maximum likelihood estimates (MLE) of the a_i can be obtained.

Table 7.1 gives the Type 2 MLE of a_i together with the empirical prior mean proportion of failures in each time period and the proportion of failures observed in each of the eight classes corresponding to engineering concerns. Note that the empirical Bayes inference does not impose a monotonic function on the form of the prior distribution.

Table 7.1 Estimates of empirical Dirichlet prior distribution parameters for unit concern classes

Time interval	MLE of a_i	EB prior estimates of \underline{P}	Observed proportion of events in classes							
			C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
1	6.74	0.28	0.34	0.20	0.00	0.30	0.50	1.00	0.33	0.18
2	3.27	0.14	0.09	0.00	1.00	0.07	0.00	0.00	0.00	0.08
3	8.08	0.34	0.27	0.20	0.00	0.50	0.50	0.00	0.33	0.49
4	4.66	0.20	0.10	0.60	0.00	0.13	0.00	0.00	0.33	0.25
5	1.03	0.04	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00

7.5 Summary and Conclusions

We have proposed a methodology for elicitation that aims to preserve the character of early stage judgements by using empirical data, where possible, to express a probability assessment of uncertainty consistent with an expert’s beliefs. Our rationale is based upon the premise that expertise does not reside in the stochastic characterisation of the events, but rather upon other problem features to which an expert can relate her domain knowledge. Thus we map the quantity of interest to an expert’s experience when there are associated data sets to support the quantification of uncertainty. Empirical Bayes inference is used to estimate the prior probabilities with the relevant observational data to provide a distribution representing the epistemic uncertainty about the quantity of interest.

We contribute a methodology consistent with an outside view of uncertainty assessment as discussed by Kahneman and Lovallo (1993). Our approach avoids imposing conformance upon an expert when assessing uncertainties probabilistically. Therefore it is capable, in principle, of overcoming some biases acknowledged to exist when an expert makes a subjective assessment through an internal mapping to an assumed probability mechanism.

7.5.1 Methodological Steps

Table 7.2 summarises our methodology in the five key steps, which can be summarised by the acronym CISSE corresponding to the initial verbs associated with the purpose of each step. The tasks involved in translating the general steps to an application are described to provide an analytical guide. Cross-references to the choices made for the two example applications are provided for illustration. Specifically, the role of the expert within each step is shown to highlight how subjective judgement is kernal to obtaining a meaningful prior distribution estimated using empirical data.

Table 7.2 CISSE methodology to construct a subjective probability distribution with data

Step	Acronym	Objective	General description	Expert role	Example 1	Example 2
1	C	Characterise the population DGP	Define the reference factors believed to characterise the future history of the quantity of interest	Provide domain knowledge to specify reference factors	Identify factors influencing new supplier quality measured by non-conformances	Identify factors influencing realisation of engineering concerns as failures
2	I	Identify candidate sample DGPs matching population	Match empirical data for entities with observed event histories to the future history of the quantity of interest based on the reference factors	Advise on candidate data sets and reasons for matching	Match to data sets for existing suppliers	Match to data sets for related designs
3	S	Sentence empirical data to construct sample DGPs	Make choices about which events for the selected data sets provide relevant empirical data for the comparator pool	Assess relevance of observed events to possible future events	Choose records for matched suppliers	Choose records for matched concern classes
4	S	Select probability model for population DGP	Choose appropriate probability model for the population to which the future and observed histories are believed to belong	Advise on credibility of model assumptions	Select Poisson-Gamma model for non-conformances	Select multinomial-Dirichlet model for realisation of failures in time windows
5	E	Estimate model parameters with sample DGPs to obtain prior distribution	Obtain empirical prior representing the future history for the quantity of interest estimated using the data for the comparator pool	Verify empirical prior is valid expression of uncertainty in quantity of interest	Estimate prior empirically and check consistency with expert's belief about uncertainty in new supplier non-conformance rate	Estimate prior empirically and check consistency with expert's belief about uncertainty in time to failure of new engineering design

7.5.2 *Effect of Sample Size on Prior Distribution*

Given our reliance on empirical data, there is an obvious question relating to the impact of ‘sample size’ on estimation and hence upon the representation of uncertainty in the prior distribution. Heuristically we can appreciate that there are two competing sample size effects. The relationship between these effects on inference might be complicated but we can reason through the effects of the choices we make in steps 2 and 3 by considering the effect of the length of a sample DGP and the number of sample DGPs separately.

Firstly, as the number of sample DPGs increases then the sampling variation in estimating the parameters of the prior distribution will reduce. This implies that the confidence regions for the comparator pool parameters will be tighter, and the associated tolerance intervals of credible prior distributions consistent with the empirical data will be narrower, when a larger number of sample DPGs are selected to match the population DGP. For example, we showed analysis of the sampling variation on the pool estimates based on the 35 suppliers used in our first example application. Had we used a larger (smaller) number of data sets providing equivalent similarity matches then we would expect the tolerance intervals to be narrower (wider) than those shown in Fig.7.7.

Secondly, as the history of a sample DGP increases then this will primarily reduce parameter estimation error associated with that individual DGP with only a marginal error reduction in estimates for the comparator pool. For example, an empirical Bayes estimate of the non-conformance rate for an individual supplier will be affected more by changes in the length of the event history for that supplier than the corresponding estimates based on the comparator pool which provides the prior distribution for the true rate of non-conformance of a new supplier.

We emphasise that our reasoning is limited to consideration of the mutually exclusive effects of the number and length of sample DGPs. However, it is important to appreciate such sample size effects because of the resulting implications for the degree of uncertainty inferred in the empirically constructed prior distribution. It is possible, as shown for our first example, to quantify the effects of sampling error allowing us to appreciate the implications for the assessment of uncertainty.

7.5.3 *Caveats and Challenges*

We acknowledge some caveats associated with our approach. Importantly, it will only be feasible in situations where it is possible to construct a comparator pool of data consistent with an expert’s articulation of the reference factors defining the population DGP. This might not always be the case. For example, radical innovations leading to very novel engineering designs in a reliability context, or long term predictions in a supply chain management context are problem contexts for which our approach is less credible. More generally, if no candidate sample

DGPs can be identified then constructing a prior through our proposed empirical approach should not be pursued. Even when comparator pools do exist then the analyst has considerable responsibility in ensuring that the data used are relevant and defensible given the impact of making choices about candidate data sets and forming relevant sample DGPs. A formal means of allowing an expert to assess the credibility of the empirical prior provides a degree of mitigation against this risk.

It is well known that empirical Bayes inference improves as comparator pool homogeneity increases (e.g. Efron 2012; Carlin and Louis 2000). Here we have constructed sample DGPs through a process involving subjective expert judgement. It is possible to scientifically aid the homogenisation process by including homogenisation factors within the probability model. See, for example, Quigley et al. (2011) who examine the role of expert judgement to specify homogenisation factors.

In our example applications we have illustrated the choices made during elicitation using the type of empirical data available at the time of analysis. Both contexts considered relate to scenarios where extensive data already exists but has not been fully utilised to understand the degree of uncertainty associated with the quantities of interest relevant to engineering development and operational decision-making. There are potentially interesting challenges affecting data selection and sentencing with more extensive or unstructured data that might be available in future. For example, in a reliability context many engineering systems are fitted with many sensors implying more empirical data is available for covariates (Meeker and Hong 2014) that may relate to the reference factors that define the population characteristics. Such explanatory data from sensors and other automated data collection might be used to support more effective and/or efficient formation of comparator pools.

Since we have proposed and illustrated how to construct a 'subjective' probability distribution using data, we conclude by emphasising the importance of engaging an expert in key steps. The nature of our approach also requires us to examine the roles of both the subject domain experts and the analytical experts because both make choices that impact the prior probability distribution obtained. The analyst makes choices in our methodology, as indeed in any elicitation process, in relation to issues such as who are experts, how should they be engaged and how should their judgements be credibly expressed. However we also require the analyst to be actively engaged in data preparation, probability model and inference method selection. Most importantly, where possible, we are not asking an expert to express his or her uncertainty about some event of interest as a subjective probability. Rather we advocate using the subject domain expertise to structure the characteristics of the population DGP for the quantity of interest and to be involved at the key stages of matching candidate data sets, sentencing records and assessing the credibility of probability distributions, both in terms of any underlying assumptions and the resulting profile of uncertainty. Our goal is to construct empirically a probability distribution that is consistent with the subjective assessment of uncertainty about a relevant quantity by the expert.

Acknowledgements We would like to thank the many engineers and managers from various companies who have been involved in challenging and evaluating our approach in practical decision-making contexts. Their engagement has helped us develop our scientific thinking into an operational process.

References

- Arnold S (1990) *Mathematical statistics*. Prentice-Hall, Englewood Cliffs
- Carlin BP, Louis TA (2000) *Bayes and empirical Bayes methods for data analysis*. Chapman & Hall/CRC, Boca Raton
- Cheng EK (2009) A practical solution to the reference class problem. *Columbia Law Rev* 109(8):2081–2105
- Cochran W (1975) *Sampling techniques*. Wiley, New York
- Cooke RM (1996) The design of reliability databases Part 1 - review of basic design concepts. *Reliab Eng Syst Saf* 51(2):137–146
- Efron B (2012) *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol 1. Cambridge University Press, Cambridge
- Efron B, Morris C (1972) Limiting the risk of Bayes and empirical Bayes estimators - Part II: the empirical Bayes case. *J Am Stat Assoc* 67(337):130–139
- Efron B, Morris C (1973) Stein's estimation rule and its competitors - an empirical Bayes approach. *J Am Stat Assoc* 68(341):117–130
- Efron B, Morris C (1975) Data analysis using Stein's estimator and its generalizations. *J Am Stat Assoc* 70(350):311–319
- Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* 96(456):1151–1160
- EFSA (2015) Scientific opinion on the risks for public health related to the presence of chlorates in food. *EFSA J* 13(6):4135
- Fragola JR (1996) Risk management in US manned spacecraft: from Apollo to Alpha and beyond. In: Perry M (ed) *Proceedings of the product assurance symposium and software product assurance workshop*, EAS SP-377, European Space Agency, pp 83–92
- Gallien J, Mersereau AJ, Garro A, Mora AD, Vidal MN (2015) Initial shipment decisions for new products at Zara. *Oper Res* 63(2):269–286
- Good IJ (1965) *The estimation of probabilities*. Research monograph, vol 30. MIT Press, Cambridge, MA
- Good IJ (1976) The Bayesian influence, or how to sweep subjectivism under the carpet. In: *Foundations of probability theory, statistical inference, and statistical theories of science*, Springer Netherlands, New York, pp 125–174
- Greenwood M, Yule GU (1920) An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J R Stat Soc* 83(2):255–279
- Hodge R, Evans M, Marshall J, Quigley J, Walls L (2001) Eliciting engineering knowledge about reliability during design-lessons learnt from implementation. *Qual Reliab Eng Int* 17(3):169–179
- Johnston W, Quigley J, Walls L (2006) Optimal allocation of reliability tasks to mitigate faults during system development. *IMA J Manag Math* 17(2):159–169
- Kahneman D, Lovallo D (1993) Timid choices and bold forecasts: a cognitive perspective on risk taking. *Manag Sci* 39(1):17–31
- Klugman SA, Panjer HH, Willmot GE (2012) *Loss models: from data to decisions*. Wiley, New York
- Koriat A, Lichtenstein S, Fischhoff B (1980) Reasons for confidence. *J Exp Psychol Hum Learn Mem* 6(2):107–118

- Meeker WQ, Hong Y (2014) Reliability meets big data: opportunities and challenges. *Qual Eng* 26(1):102–116
- Nagurney A, Li D (2016) *Competing on supply chain quality*. Springer, Berlin
- Ng KW, Tian GL, Tang ML (2011) *Dirichlet and related distributions: theory, methods and applications*, vol 888. Wiley, New York
- Pahl G, Beitz W (2013) *Engineering design: a systematic approach*. Springer, Berlin
- Quigley J, Walls L (2011) Mixing Bayes and empirical Bayes inference to anticipate the realization of engineering concerns about variant system designs. *Reliab Eng Syst Saf* 96(8):933–941
- Quigley J, Bedford T, Walls L (2007) Estimating rate of occurrence of rare events with empirical Bayes: a railway application. *Reliab Eng Syst Saf* 92(5):619–627
- Quigley J, Hardman G, Bedford T, Walls L (2011) Merging expert and empirical data for rare event frequency estimation: pool homogenisation for empirical Bayes models. *Reliab Eng Syst Saf* 96(6):687–695
- Quigley J, Walls L, Demirel G, MacCarthy B and Parsa M (2018) Supplier quality improvement: the value of information under uncertainty. *Eur J Oper Res* 264(3):932–947
- Rausand M, Hoyland A (2004) *System reliability theory: models, statistical methods and applications*. Wiley, New York
- Reichenbach H (1971) *The theory of probability*. University of California Press, Berkley
- Robbins H (1955) An empirical Bayes approach to statistics. In: *Proceedings of the third Berkley symposium mathematical statistics and probability 1*, University of California Press, Berkley, pp 157–164
- Slack N, Brandon-Jones A, Johnston R (2016) *Operations management*, 8th edn. Pearson
- Sodhi MS, Tang CS (2012) *Managing supply chain risk*. Springer, Berlin
- Spetzler CS, Stael von Holstein CAS (1975) Exceptional paper-probability encoding in decision analysis. *Manag Sci* 22(3):340–358
- Talluri S, Narasimhan R, Chung W (2010) Manufacturer cooperation in supplier development under risk. *Eur J Oper Res* 207(1):165–173
- von Mises R (1942) On the correct use of Bayes' formula. *Ann Math Stat* 13(2):156–165
- Walls L, Quigley J (1999) Learning to improve reliability during system development. *Eur J Oper Res* 119(2):495–509
- Walls L, Quigley J (2001) Building prior distributions to support Bayesian reliability growth modelling using expert judgement. *Reliab Eng Syst Saf* 74(2):117–128
- Walls L, Quigley J, Marshall J (2006) Modeling to support reliability enhancement during product development with applications in the UK aerospace industry. *IEEE Trans Eng Manag* 53(2):263–274
- Wilson KJ, Quigley J (2016) Allocation of tasks for reliability growth using multi-attribute utility. *Eur J Oper Res* 255(1):259–271
- Zhu K, Zhang RQ, Tsung F (2007) Pushing quality improvement along supply chains. *Pest Manag Sci* 53(3):421–436