Bayesian modelling of positively skewed data with particular application to the cost aspect of medical cost-effectiveness analysis

Peter Leonard Gregory

A thesis presented for the degree of
Doctor of Philosophy

Department of Probability & Statistics

July 2010

Summary of this thesis

The motivation for this research was the study of a medical cost data set from a clinical trial. If a Regulatory Body were to be accept the new intervention that has been proposed then a Health Care provider has to budget for future treatments for some members of the rest of the population.

In this Bayesian analysis we want to be able to calculate the expected value for one unobserved member of this finite population from its posterior predictive distribution by firstly establishing the parametric data model that best captures the positive skew characteristics of the costs. We then develop a novel approach to modelling the priors that enable an expert's prior beliefs to be elicited while permitting a limited analytical study of the model.

These techniques have been applied to recent medical data sets to establish their comparative efficiency when compared with classical estimators.

Acknowledgements

I would firstly like to thank those academics who have influenced me. I would not have developed my interest in statistics at all at the University of Birmingham without the positive benefits resulting from my contact with David Wishart and Henry Daniels. This led to me working on a project for Joe Gani and subsequently my first involvement with the Department of Probability and Statistics.

I would like primarily to thank my supervisor Tony O'Hagan for his influence on me during this, my second spell at the University of Sheffield . He possesses a deep understanding of Bayesian statistics that has been truly inspirational and he has known, rather like a conductor to an orchestra, when to pull which strings throughout the highs and lows that I have experienced. My thanks are also due to John Stevens for the pmdI+ data set and to Simon Dixon for his two data sets. When I had to turn to R then Jeremy Oakley was present to guide me in the right direction and also made the very timely release of SHELF. Last, but by no means least, I would like to thank John Biggins for offering his wise counsel when required.

Secondly I want to say how much I have enjoyed the company of my student colleagues during the last few years. We have held many social events together and formed long lasting friendships. Perhaps a particular high point was for the group of us who worked together on the very successful RSC2004 conference which was held in Sheffield.

Thirdly, I want to offer my thanks to my family. My parents encouraged me to become the first member of the the family to go to University where I met the lady who subsequently became my wife and best friend for the last forty plus years with many more to come we hope. She has given me emotional support when I needed it and is very pleased that "the completion of the thesis is in sight" ! I do now receive inspiration from my three granddaughters whose development is, in every sense of that word, a great source of interest, fascination and stimulation.

Contents

Chapter 1

Introduction

1.1 Historical background

The birth of Evidence Based Medicine can be traced back to the Crimea War and Florence Nightingale. Although born in Italy her grandmother lived in Sheffield and her family resided in Derbyshire for a while. The year 2010 contains two of the important anniversaries relating to her. The 15th June 1860 is the date when she admitted her first 15 students to her school for nursing at St Thomas' Hospital in London and 13th August is the centenary of her death.

Florence Nightingale was better known as "The Lady with the Lamp" which derives from a report in *The Times* quoted by Cook (1913) and popularised by a poem by Longfellow (1857). She was sent to the Black Sea in 1854 with thirty-eight nurses to provide medical care for the wounded in the war against the Russians. When she returned to England for two years she collected evidence about the death rates for injured soldiers in hospital, when she came to believe that most of the soldiers at the hospital were not dying primarily from their battle wounds but from the diseases that they contracted whilst under medical care in the military hospital.

To support her numbers she devised Polar area diagrams (see for example Nightingale (1859)) which, by the twenty-first century, have evolved into what we now know as Pie charts. In 1858 she became the first woman to be elected a Fellow of the Statistical Society of London, as the Royal Statistical Society was then known, in recognition of her contribution to statistics.

If we now move forward in time to around 1990 then the concept of Clinical Trials was well established, see for example Friedman *et al* (1998), and consisted of the following four phases

Phase	Description	Trialists	Testing
I	initial	volunteers	safety
II	clinical	patients	dosage & efficacy
III	multi-centre comparative clinical	patients	safety, efficacy & regulatory
IV	post-marketing	patients	side-effects

The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) was launched in 1990 to bring together the regulatory and pharmaceutical industry of Europe, Japan and the United States. Its mission is to "make recommendations towards achieving greater harmonisation in the interpretation and application of technical guidelines and requirements for pharmaceutical product registration, thereby reducing or obviating duplication of tesing carried out during the research and development of new human medicines".

ICH Guidelines are issued for Quality, Safety, Multidisciplinary and Efficacy topics - which includes Clinical Trials E7-E11.

So until around 1990 if efficacy improved, with no adverse safety effects, then a Health Care Provider could make a decision, post Phase III trials, to accept the new treatment or intervention.

However, the rising costs of new treatments that were being developed in an environment of limited, if not fixed, financial resources meant that Health Care Providers were forced to consider both cost and effectiveness and then to look at Cost-effectiveness by involving Health Economists. Indeed the National Institute for Health and Clinical Excellence (NICE) is an NHS organisation established in the UK in 1999 with a brief to consider both cost and effectiveness evaluations when carrying out technology appraisals for proposed new interventions.

The initial Cost-effectiveness investigations, when undertaken statistically, were made using a frequentist approach in papers such as Willan and O'Brien (1996) and Drummond and O'Brien (1993) and used the Incremental Cost-Effectiveness Ratio (ICER) to make treatment comparisons, where

$$\text{ICER} = \frac{\Delta_c}{\Delta_e}$$

and Δ_c and Δ_e will be defined in Section 1.2.

Quantifying uncertainty for the estimator of the ICER gave difficulties, see for example O'Brien *et al* (1994), because it was the ratio of two random variables and in Willan & O'Brien (1996) they resorted to Fieller's Theorem, see Fieller (1954). This approach was continued, see for example Willan (2001), but with the author beginning to recognise both Net Monetary Benefit, see Section 1.2, and the Bayesian approach. The Bayesian approach was argued in O'Hagan & Stevens (2002) with the advantages of the NMB and subsequently the CEAC, see Section 1.2, introduced.

So the natural framework for such analysis is Bayesian inference as was argued in Briggs (1999) and O'Hagan *et al* (2000) and the Bayesian approach has been developed, using retrospective analysis of Phase III studies at first.

All of the initial approaches to Cost-Effectiveness analysis used the simplifying assumption of normality for the the underlying cost and efficacy data, or at least that the sample size is large enough for sample means to be normally distributed.

One of the main research interests in this theses is to identify data models that fit the cost data better than a normal distribution and this topic will be pursued in Chapter 2.

1.2 Model definition

This model is introduced in O'Hagan *et al* (2001) and describes a clinical trial to compare two treatments, where we wish to assess whether treatment 2 is more cost-effective then treatment 1.

We have n_i patients in treatment group i ($i = 1$ or 2) where they will provide efficacy data e_{ij} and cost data c_{ij} ($j = 1, 2, \ldots, n_i$). Let $\mathbf{y}_{ij}$ be the vector $(e_{ij}, c_{ij})^T$ and the complete data set be $\mathbf{y} = \{\mathbf{y}_{ij} : i = 1, 2; j = 1, 2, \ldots, n_i\}$.

For any treatment i let the population mean efficacy be μ_i with population mean cost similarly defined as γ_i and covariance matrix Σ_i to allow for correlation between cost and efficacy for a given patient. Then the mean incremental efficacy of treatment 2 over treatment 1 is $\Delta_e = \mu_2 - \mu_1$ where Δ_c is the corresponding mean cost increment. If we let $\boldsymbol{\alpha}_i = (\mu_i, \gamma_i)^T$ then the parameters may be defined as $\boldsymbol{\theta} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \Sigma_1, \Sigma_2)$.

The Net Monetary Benefit is defined to be

$$\beta(K) = K\Delta_e - \Delta_c$$

where K is the maximum cost per unit improvement in efficacy that a health provider is prepared to pay. Hence if $\beta(K) > 0$ treatment 2 is more cost-effective than treatment 1.

The assumption made in O'Hagan *et al* (2001) is that it is reasonable to assume independence between patients conditional on the underlying parameters. Whilst their model assumes a simple structure and there are many circumstances when it is reasonable to assume independence between patients we do need to recognise that there will also be some situations when this would not be reasonable and the model used would need to capture the dependence between patients.

More recent research, see for example Grieve *et al* (2005) and Manca *et al* (2005), have addressed multinational randomised clinical trials and they have used multilevel models to represent the two sources of variation (between country and between patients, within a country). These models capture the more complex hierarchical structure that may be present within the data but still proceed using independence between patients conditional on the underlying parameters.

We will proceed here under the customary assumption of independence between patients conditional on the underlying parameters when our data model says that

$$\mathbf{y}_{ij}|\boldsymbol{\theta} \sim N(\boldsymbol{\alpha}_i, \Sigma_i) : i = 1,2; j = 1,2,\ldots,n_i$$

and after formulating appropriate prior beliefs we can then examine the posterior probability that $\beta(K)$ is positive, $Q(K)$ as

$$Q(K) = p\{\beta(K) > 0|\mathbf{y}\}$$

which, regarded as a function of K, is called the Cost-Effectiveness Acceptability Curve (CEAC). For the justification of the interpretation of this approach only within a Bayesian framework see Section 1.5.

1.3 The Cost-effectiveness plane

In Section 1.2 for treatment i we let the population mean efficacy be μ_i with mean cost similarly defined as γ_i. If we let the co-ordinate pair (μ, γ_i) represent treatment Ti then the Cost-effectiveness plane allows the population mean cost and effectiveness for many different treatments to be plotted in the same two dimensional plane as shown in Figure 1.1 for six treatments

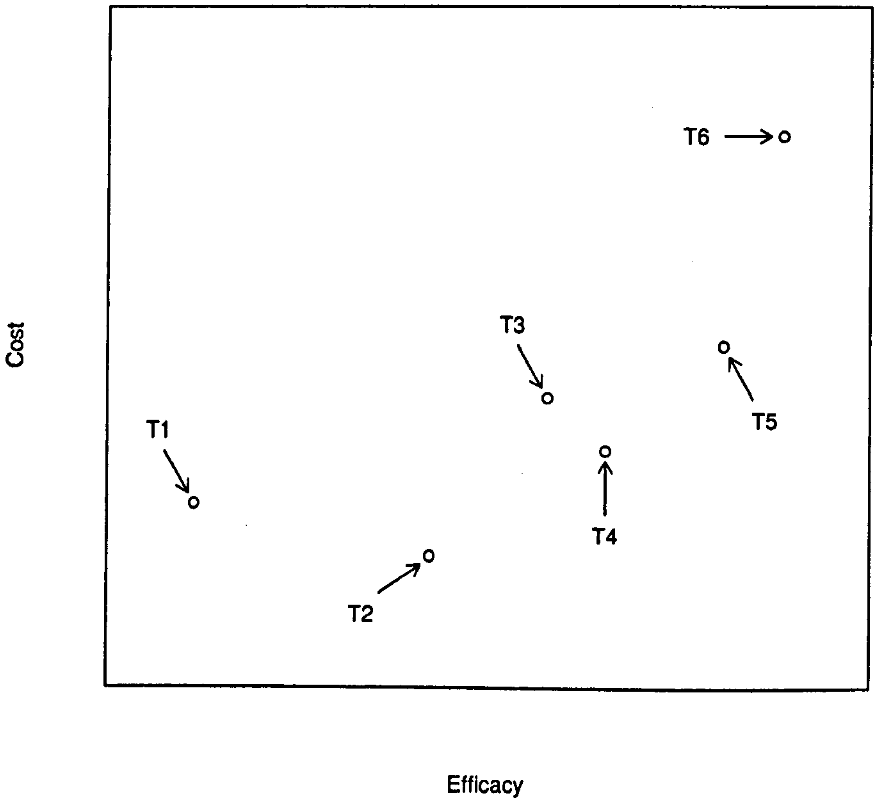


Figure 1.1: The Cost-effectiveness plane for treatments type A1 to A6

However, the usual practical situation will be the comparison of an existing treatment against an alternative proposed treatment. For this specification it is more useful to take the existing treatment as the base point and then consider the origin to be located at this point. The difference in cost and effectiveness between the existing treatment and any other treatment allows comparisons to be made between pairs of treatments.

1.4 The Incremental Cost-effectiveness plane

This concept was introduced by Black (1990). It enables graphical comparisons between pairs of treatments, taking the existing treatment as the base point and is a development of the The Cost-effectiveness plane shown in Figure 1.1. It is illustrated in Figure 1.2 below, where the diagonal continuous line represents $K\Delta_e = \Delta_c$

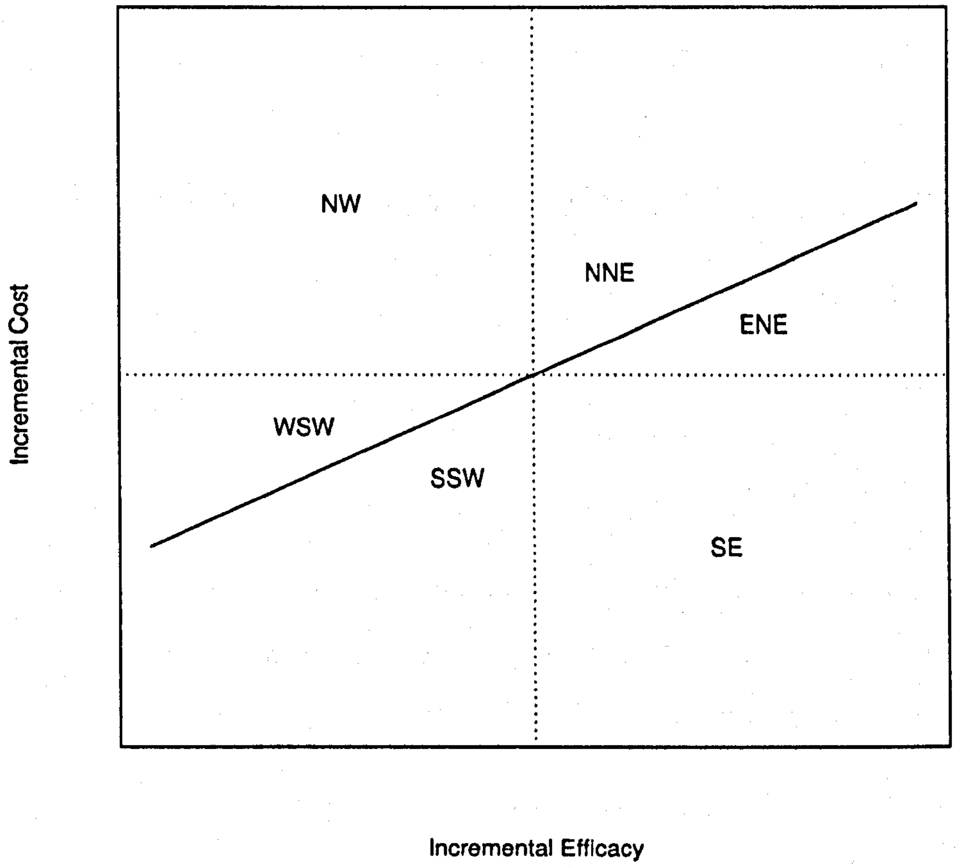


Figure 1.2: The Incremental Cost-effectiveness plane

The plane naturally divides into four quadrants labelled NE, NW, SW and SE for unique identification and when subdivision is required then the cartographical description is continued as, for example, the NE quadrant divides into NNE and ENE regions.

The line $\Delta_c = K\Delta_e$ partitions the plane where the area below this line can be seen to represent

$$\beta(K) = K\Delta_e - \Delta_c > 0$$

the region where treatment 2 will be preferred to (the existing) treatment 1 because

ENE region has a cost increase that is less than $K\times$ the efficacy gain
SE quadrant has a cost reduction accompanied by an efficacy gain
SSW region has a cost reduction that is greater than $K\times$ the reduction in efficacy.

1.5 The Cost-Effectiveness Acceptability Curve (CEAC)

The concept of the CEAC was introduced by Van Hout *et al* (1994) in a frequentist context.

We know from, Section 1.2 that the population mean incremental efficacy of treatment 2 over treatment 1 is $\Delta_e = \mu_2 - \mu_1$ where Δ_c is the corresponding population mean cost increment and the Net Monetary Benefit is defined to be

$$\beta(K) = K\Delta_e - \Delta_c$$

where K is the maximum cost per unit improvement in efficacy that a health provider is prepared to pay. Hence if $\beta(K) > 0$ treatment 2 is more cost-effective than treatment 1.

Van Hout *et al* (1994) introduced the CEAC as the value of the probability that treatment 2 is acceptable for some given fixed value of K or the probability that (Δ_e, Δ_c) falls within the acceptability region.

However, within frequentist statistics unknown parameters are not random variables following probability distributions and hence probability statements can not be made about their possible values.

The probability that (Δ_e, Δ_c) falls within the acceptability region is meaningful only in a Bayesian framework.

So after formulating appropriate prior beliefs we can then examine the posterior probability that $\beta(K)$ is positive, $Q(K)$, as

$$Q(K) = p\{\beta(K) > 0|\mathbf{y}\}$$

which, regarded as a function of K, is called the Cost-Effectiveness Acceptability Curve (CEAC).

This definition recognizes that the acceptability of treatment 2 depends on the value of K, the maximum cost per unit improvement in efficacy that a health provider is prepared to pay, which may or may not be known in practise. Hence it is more useful to determine $Q(K)$ over a range of values of K (and possibly different representations of prior information) as we can show in the chart below

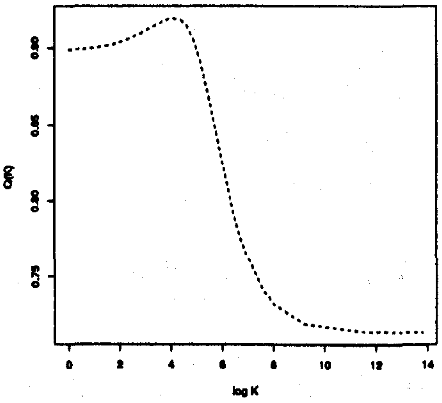


Figure 1.3: The CEAC : A chart to show a possible Q(K)

1.6 The pMDI data set

The CFC-propelled pressurized metered-dose inhaler pMDI has been used to treat Asthma for more than thirty years. However, amongst the perceived problems with pMDI-propelled inhalers is their contribution, however small, to the destruction of the ozone layer. This has led to the development of alternative inhalers including the inspiratory flow-driven multidose dry powder inhaler called Turbuhaler®.

Although a number of studies have shown specific advantages for Turbuhaler® over pMDI's, the study that was conducted by Pauwels *et al* (1996) was the first long-term study performed to compare the effectiveness of pMDI (treatment 1) against Turbuhaler® (treatment 2). A total of 1004 patients from 77 centres in 7 countries took part in the year long trial. The pooled results were evaluated for effectiveness by Pauwels *et al* (1996) using two common definitions of effectiveness.

When Liljas *et al* (1997) conducted their cost-effectiveness analysis, they had to recognise that some of the well known difficulties associated with cost comparisons between countries precluded the pooling of data. In their study the Canadian data was selected because the largest group of patients (445) came from that country. They showed that Turbuhaler® was dominant in the sense that it was both cheaper and more effective as well as reducing the contribution to the destruction of the ozone layer.

The UK data contained 58 patients receiving pMDI and 62 patients receiving Turbuhaler®. For each patient there was an observed measure of efficacy and an observed total cost for that trial period. This data set has been studied by O'Hagan *et al* (2001) using a Bayesian approach. The O'Hagan & Stevens (2003) paper studied the cost data for the pMDI+ patient group (those patients treated with pMDI and having a positive outcome ie no exacerbations) which contains 26 observations and is shown in Figure 1.4 below

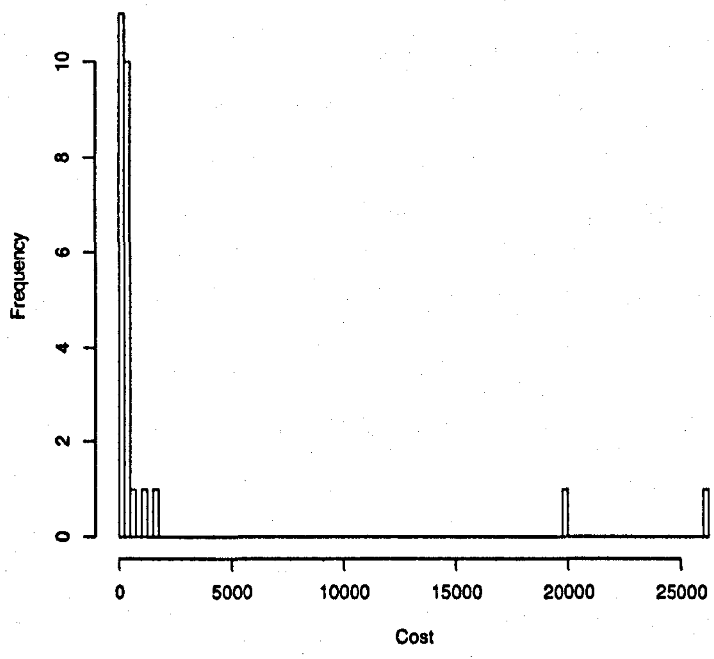


Figure 1.4: The pMDI+ data set

This data set contains positive values and is right, or positively, skewed with a few very large values. Whilst this would not be considered in any way atypical for a medical cost data set, the few very large values may or may not be present in another data set. Typically the data set may be small (in absolute numbers) and small in comparison with the size of our population.

It is of interest to note that the coefficient of skewness for the pMDI+ data set is 3.48 while, anticipating the analysis which is to come, for the log transformed pMDI+ data set it is 1.75. So even after the log transformation has been made some degree of positive skewness has been retained.

1.7 The structure of this thesis

The motivation for this research is the study of the pMDI+ data set introduced above with the main interest being how a Health Care Provider can budget for the cost of an intervention, as the posterior predictive mean, for unobserved members of a population.

The research will establish a suitable data model and will then develop prior beliefs that will satisfy three criteria. They will be shown analytically to produce values for the posterior predictive mean that will be shown to be finite, will allow numerical evaluation and will enable an expert's prior beliefs to be elicited.

The early work on Cost-effectiveness modelling made the convenient, but fairly unrealistic, assumption that costs followed a Normal distribution. The purpose of Chapter 2 is to explore realistic models for costs that capture the non-negative positive skew nature of the pMDI+ cost data set.

We will show that conducting a Bayesian comparison of candidate models tells us that the logNormal distribution is the most appropriate data model for the pMDI+ cost data set and indeed the logNormal distribution is frequently used as a financial parametric model.

It is in Section 2.4.2 that we first encounter a double integral of the form

$$m_\lambda(\mathbf{w}) = \int_0^\infty \int_{-\infty}^\infty \pi(\mu, \sigma^2) \left[f_\lambda(\mathbf{y})\right]^b d\mu d\sigma^2$$

where the random variable Y, observed as y, is transformed as $X = \lambda^{-1}(Y^\lambda - 1)$ where $X \sim \mathrm{N}(\mu, \sigma^2)$.

The type of integrand in the double integrals that we will encounter in this thesis naturally lead, for ease of evaluation, to determining the integral with respect to μ first.

In Chapter 3 we will examine inference for the posterior predictive mean from a finite population, using super-population theory to determine how to be able to predict future observations.

In Section 3.4.2 we will use WinBUGS to produce true values for the posterior predictive mean. When using WinBUGS we can only specify beliefs that are proper and hence we need to give numerical values to the parameters of the prior distributions which determines how informative are our prior beliefs.

Default priors can be considered a reference against which other priors may be compared. They may be informative or, more usually, noninformative.

In clinical trials sceptical priors express scepticism about large treatment effects and have been put forward as a reasonable expression of doubt.The sceptical prior formalises the belief that large treatment differences are unlikely. This is usually set up, see Ashby (2000), as having a mean of no treatment effect and only a small probability of the effect achieving a clinically relevant effect. Alternatively, subjective clinical opinion may form the basis of a prior.

Tessella plc, see Tessella (2009/11), believe that the most obvious, but also contentious use of the Bayesian approach in clinical trials is to include a prior belief for the effect of the treatment in a clinical trial. Normally they would have to include a sceptical prior in order that the posterior results are convincing to a regulatory body such as the FDA. The US Department of Health and Human Services, Food and Drug Administration (FDA) is the US regulatory body that forms part of the ICH. In its Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials, FDA (2010), it has recommended that prior distributions be based on "good" information such as pilot studies, which should be presented and discussed with FDA reviewers before the study begins. This will then lead to informative priors but the FDA have suggested that noninformative priors will still need to be used in certain circumstances.

An important question is whose prior information. This information may be possessed by the developing company. In general the regulatory bodies start from a position of scepticism and may interpret this as meaning that they should have a prior prejudice against new treatments. The choice of prior beliefs is not a simple matter. It has been stated in O'Hagan & Stevens (2001) that an appropriate prior should incorporate scepticism about the cost-effectiveness of a new treatment.

The comments above are motivated by the effectiveness leg of cost-effectiveness but are equally applicable to the cost leg.

So to return to the question of the numerical values that we will give to the parameters of the prior distributions, we wish to choose a sceptical prior belief that is noninformative. In Congdon (2001), Example 2.4 takes a G(0.0001,0.0001) prior for the precision and a N(0,10000) for the mean. This specification is even more diffuse than that which we will use in Section 3.4.2 and shows how far researchers may be prepared to go to capture noninformative prior beliefs.

We believe that the numerical values that we will give to the parameters of the prior distributions in Section 3.4.2 have been chosen to include any values that we might expect to see for the mean and precision whereas more restrictive choices for the parameter values may lead to conflict with future data sets.

WinBUGS can only produce true values for the posterior predictive mean that are finite. However, the instability of the results produced by WinBUGS, when using customary very weak priors for a logNormal data model, as the number of samples is increased do indicate that this model produces infinite values.

A theoretical analysis will establish that this is indeed the case when using customary noninformative prior beliefs for a logNormal data model.

The main focus of our research will begin in Chapter 4 where we will develop a novel approach to modelling independent prior beliefs for the shape and scale parameters for this logNormal data model. We will utilise some of the properties of the logNormal distribution that have been introduced in the Glossary to be able to restructure our prior beliefs in the observation space and work with the Median (a function of the scale parameter alone) and the Quantile Ratio (a function of the shape parameter alone).

We will then establish analytically the existence of finite posterior predictive moments. In particular, we will show that posterior predictive moments for an unobserved member of the population exist for a wide choice of prior beliefs for the Median but care is required to ensure convergence for our choice of prior belief for the Quantile Ratio.

We will introduce in Chapter 4 an alternative representation and notation for "the posterior predictive mean of an unobserved member of the population for our Bayesian model" equivalently "the posterior expectation of the population mean" as the "Bayesian posterior expectation", which is $\mathbb{E}\{\exp(\mu+\sigma^2/2)|\mathbf{y}\}$, or the Bpe. We will work with whichever definition is most convenient in the future.

In Chapter 5 we will develop default values for the prior beliefs introduced in Chapter 4 when they are trained by models introduced in Briggs *et al* (2005). We will then make comparisons between the estimators that they introduced and the posterior predictive mean of an unobserved member of the population for our Bayesian model for their data sets. We will extend their analysis by considering other data generating models as well as other observed data sets.

We will also, in contrast to the Briggs *et al* (2005) simulation study, be able to present a number of theoretical results for their type of comparisons and also be able to compare shrinkage estimators using theoretical results for versions of their estimators with our Bpe. Finally we will make comparisons between the estimators for the mean of logNormal distributions considered in Zhou (1998) and our Bpe.

In Chapter 6 we will develop the principles behind the procedure to elicit the prior beliefs that have been proposed in Chapter 4. The details of The elicitation will be shown in the Appendix. Two case studies applying this procedure, where in each case we have elicited an expert's prior beliefs about salary distribution, will also be presented.

We will summarise the achievements of this thesis in Chapter 7 and present opportunities for further research.

Chapter 2

Data model selection

2.1 Introduction

This chapter will explore some possible candidate models that will capture the non-negative positive skew shape of the distribution of the random variable Y, observed as y, in the pMDI+ data set.

Bayesian statistics recognises two kinds of uncertainty - which is alluded to in The elicitation in the Appendix in the section headed Uncertainty. Epistemic uncertainty arises because of lack of knowledge, such as the values of parameters. Aleatory uncertainty arises because of randomness.

Whenever we have observed data then we will assume that this data has arisen from some "true" data-generating process which we would be able to specify if we had complete knowledge. Uncertainty about this true data-generating process exists because we don't have complete knowledge and this uncertainty is typically expressed through a statistical model, which is the set of possible data-generating processes.

To decide the statistical model that is most appropriate for a given data set we use model selection techniques. There are three main Bayesian methods, namely the full probability model that we will adopt here, the Bayes information criterion or BIC and the Deviance information criterion or DIC.

The Schwarz criterion, also known as the BIC, was introduced by Schwarz (1978) and is a common approximation to the log of the Bayes factor. It favours more strongly the model with fewer parameters. An equivalent result was obtained by Poskitt (1987) for quite general models.

For complex hierarchical models, where the number of parameters may not be a well-defined quantity, Spiegelhalter *et al* (2002) adopted a semiformal approach to introduce the DIC.

The full probability model leads to comparisons of pairs of models using the ratio of the marginal (or integrated) densities for the models being considered, known as the Bayes factor. Analytical evaluation of these marginal densities is possible in certain situations, see DeGroot (1970) and Zellner (1971a), and this is the approach that we will follow here.

When comparing pairs of models in a Bayesian framework then hypothesis testing is undertaken by determining the posterior odds in favour of one of the models which can be expressed, as will be shown in Section 2.2, as the prior odds in favour of that model multiplied by the Bayes factor in favour of that model. The Bayes factor is derived after observing the data and, if the Bayes factor is greater than one, then that model fits the data better than the alternative model.

As we will show in Section 2.2 the Bayes factor for comparing model M_i against M_j for the observed data $\mathbf{y}$, $B_{ij}(\mathbf{y})$, is

$$B_{ij}(\mathbf{y}) = \frac{m_i(\mathbf{y})}{m_j(\mathbf{y})} \tag{2.1}$$

where

$$m_i(\mathbf{y}) = \int_{\Omega_i} \pi_i(\boldsymbol{\theta}_i) f_i(\mathbf{y}|\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$$

is the marginal density of $\mathbf{y}$ under M_i, whose parameters are defined over the parameter space $\boldsymbol{\Omega}_i$, with prior density $\pi_i(\boldsymbol{\theta}_i)$ and likelihood $f_i(\mathbf{y}|\boldsymbol{\theta}_i)$. The marginal densities are obtained by integrating over the parameter space for the model being considered where, for Equation 2.1 to be defined, it is necessary that the marginal density of each $\mathbf{y}$ under M_i is proper.

The concepts underlying Bayes factors are introduced in O'Hagan and Forster (2004) and Gelman *et al* (1995) with a more detailed discussion given in Kass and Raftery (1995). Amongst the limitations of Bayes factors are their sensitivity to the choice of the data model and prior beliefs. When there is not any available information from which to construct prior beliefs then the invariance principle proposed by Jeffreys for noninformative priors, see Jeffreys (1961), is frequently adopted. We will follow Jeffreys' rule in Section 2.5.2 for our Gamma data model and in Section 2.4.2 for our Normal data model, although we will follow Jeffreys himself and rather than using $\pi(\mu, \sigma^2) = (\sigma^3)^{-1}$, which the rule suggests, we will specify $\pi(\mu, \sigma^2) = (\sigma^2)^{-1}$, as he did.

General noninformative prior beliefs which are also improper will be introduced in Section 2.2.1 where we will not impose the restriction that the marginal density of each $\mathbf{y}$ under M_i is proper. To deal with the difficulty that this causes Smith and Spiegelhalter (1980) and Spiegelhalter and Smith (1982) introduced the concept of an "imaginary training sample device".

Another approach relies on the concept of the "training sample" introduced by Lempers (1971) from which partial Bayes factors will be developed in Section 2.2.1. Berger and Pericchi (1996) developed, what they called, intrinsic Bayes factors by averaging the partial Bayes factors arising from all the possible training samples arising from some fixed training sample size. We will follow here the fractional Bayes factor approach which we will introduce in Section 2.2.2 and so allow Bayes factors to be calculated without having to specify which data points are present in the training sample.

We will simplify the notation used from Section 2.3 onwards for the rest of Chapter 2 and refer to, for example, $f_X(x)$ rather than the strictly correct $f_X(x|\boldsymbol{\theta})$.

The "composite models" that will be developed here are formulated by firstly considering a range of possible power transformations, as introduced in Section 2.3, to reduce the positive skew shape. In Sections 2.4 and 2.5 we will develop the Bayesian analysis for the two parametric models that will be considered after applying an appropriate power transformation to Y.

In Section 2.6 the theory will be produced to allow the fBf to be defined for the comparison of a range of rootNormal vs logNormal and logNormal vs rootGamma models. In Section 2.8 will be the numerical results to show these comparisons for the pMDI+ data set.

Section 2.7 will discuss the choice of training fraction to be used while in Section 2.8.3 we will produce numerical results to support the choice of training fraction. This latter section will close with the justification for proceeding with the use of the logNormal data model using non-Bayesian examples.

There are of course many Bayesian examples, for example Padgett and Johnson (1983) considered applications in reliability and life testing, Chen (2002) gave an application in fish stock-recruitment, Zellner (1971b) explored both the logNormal distribution and logNormal regression models where his particular field of interest is econometrics whilst Khan *et al* (2005) have explored Bayesian prediction under censoring as used in biological, industrial and medical sciences.

2.1.1 Acknowledgement of my source material

Section 2.2 is based on Chapter 7 from O'Hagan and Forster (2004) and also on O'Hagan (1995) and (1997). Sections 2.4, 2.5, 2.6 and 2.8 are built around hand written notes by Prof O'Hagan.

2.2 Bayes factors for model comparisons

We will develop the analysis for the case of m models under consideration for the data set $\mathbf{y} = (y_1, y_2 \ldots y_n)$ which are n observations of the independent identically distributed random variables $(Y_1, Y_2 \ldots Y_n)$.

Let model M_i have parameters $\boldsymbol{\theta}$ defined over a parameter space $\boldsymbol{\Omega}_i$ with prior density $\pi_i(\boldsymbol{\theta}_i)$ and likelihood $f_i(\mathbf{y}|\boldsymbol{\theta}_i)$.

Hence the posterior distribution for $\boldsymbol{\theta}_i$, conditional on M_i, becomes

$$p_i(\boldsymbol{\theta}_i|\mathbf{y}) = \frac{\pi_i(\boldsymbol{\theta}_i)f_i(\mathbf{y}|\boldsymbol{\theta}_i)}{m_i(\mathbf{y})}$$

where

$$m_i(\mathbf{y}) = p\{\mathbf{y}|\boldsymbol{\theta}_0 \in \boldsymbol{\Omega}_i\} = \int_{\boldsymbol{\Omega}_i} \pi_i(\boldsymbol{\theta}_i)f_i(\mathbf{y}|\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \tag{2.2}$$

is the marginal density of $\mathbf{y}$ under M_i.

If we let the prior probability for model M_i be

$$p_i = p\{\boldsymbol{\theta}_0 \in \boldsymbol{\Omega}_i\} = p\{M_i\}$$

then we can define a unified model M_0, with parameter $\boldsymbol{\theta}_0$, whose parameter space is $\boldsymbol{\Omega}_0 = \bigcup_{i=1}^{m} \boldsymbol{\Omega}_i$. Model M_i is chosen whenever $\boldsymbol{\theta}_0 \in \boldsymbol{\Omega}_i$.

So the prior density becomes

$$\pi_0(\boldsymbol{\theta}_0) = p_i\,\pi_i(\boldsymbol{\theta}_0) \qquad \text{if} \qquad \boldsymbol{\theta}_0 \in \boldsymbol{\Omega}_i$$

and hence the posterior density is

$$p_0(\boldsymbol{\theta}_0|\mathbf{y}) = p_i^*\,p_i(\boldsymbol{\theta}_0|\mathbf{y}) \qquad \text{if} \qquad \boldsymbol{\theta}_0 \in \boldsymbol{\Omega}_i$$

where

$$p_i^* = p\{\boldsymbol{\theta}_0 \in \boldsymbol{\Omega}_i|\mathbf{y}\} = p\{M_i|\mathbf{y}\}$$

may be found using Bayes theorem as

$$\begin{aligned} p_i^* &= \frac{p\{\mathbf{y}|\boldsymbol{\theta}_0 \in \boldsymbol{\Omega}_i\}p\{\boldsymbol{\theta}_0 \in \boldsymbol{\Omega}_i\}}{p\{\mathbf{y}\}} \\ &= p_i \frac{m_i(\mathbf{y})}{m_0(\mathbf{y})} \end{aligned}$$

where

$$m_0(\mathbf{y}) = \sum_{i=1}^{m} p_i m_i(\mathbf{y})$$

is the marginal density of $\mathbf{y}$ under M_0.

To compare model M_i against model M_j we will determine the ratio of their posterior probabilities as

$$\frac{p_i^*}{p_j^*} = \frac{p_i}{p_j}\frac{m_i(\mathbf{y})}{m_j(\mathbf{y})} = \frac{p_i}{p_j}B_{ij}(\mathbf{y})$$

where $B_{ij}(\mathbf{y})$ is called the Bayes factor for comparing model M_i against M_j for the observed data $\mathbf{y}$.

The ratio p_i/p_j represents the prior odds for comparing model M_i against model M_j. The corresponding ratio p_i^*/p_j^*, as the posterior odds, is expressed as the prior odds multiplied by the Bayes factor.

The Bayes factor is a summary of (the strength of) the evidence provided by the data in favour of one statistical model when compared with an alternative. If model M_j fits the data $\mathbf{y}$ better than model M_i then the Bayes factor will be less than 1 and the posterior odds will be less than the prior odds. Conversely, if model M_i fits the data $\mathbf{y}$ better than model M_j then the Bayes factor will be greater than 1 and the posterior odds will be greater than the prior odds.

We obtained the marginal density $m_i(\mathbf{y})$ by integrating over the parameter space $\boldsymbol{\Omega}_i$ as the parameter $\boldsymbol{\theta}_i$ is unknown. This relationship, see (2.2), requires the prior density $\pi_i(\boldsymbol{\theta}_i)$ and may therefore be sensitive to the prior that is chosen.

2.2.1 Partial Bayes factors for improper priors

Consider now comparing model 1 against model 2 using the Bayes factor

$$B_{12}(\mathbf{y}) = \frac{m_1(\mathbf{y})}{m_2(\mathbf{y})} \tag{2.3}$$

where

$$m_i(\mathbf{y}) = \int_{\Omega_i} \pi_i(\boldsymbol{\theta}_i) f_i(\mathbf{y}|\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i.$$

If we wish to represent weak prior information about, say, the parameter $\boldsymbol{\theta}_1$ as an noninformative prior distribution then this will generally be improper, which is defined as

$$\pi_1(\boldsymbol{\theta}_1) \propto k_1(\boldsymbol{\theta}_1)$$

where $\int_{\Omega_1} k_1(\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1$ does not converge.

It is possible to write

$$\pi_1(\boldsymbol{\theta}_1) = c_1 k_1(\boldsymbol{\theta}_1)$$

where c_1 is an undefined (and strictly non-existent) constant.

The Bayes factor then becomes

$$B_{12}(\mathbf{y}) = c_1 \frac{\int_{\Omega_1} k_1(\boldsymbol{\theta}_1) f_1(\mathbf{y}|\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1}{\int_{\Omega_2} \pi_2(\boldsymbol{\theta}_2) f_2(\mathbf{y}|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2}$$

and depends on the unspecified c_1.

Similarly, if an improper prior for model 2 is expressed as $\pi_2(\boldsymbol{\theta}_2) = c_2 k_2(\boldsymbol{\theta}_2)$ then

$$B_{12}(\mathbf{y}) = \frac{c_1}{c_2} \frac{\int_{\Omega_1} k_1(\boldsymbol{\theta}_1) f_1(\mathbf{y}|\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1}{\int_{\Omega_2} k_2(\boldsymbol{\theta}_2) f_2(\mathbf{y}|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2}$$

which depends on the unspecified ratio c_1/c_2.

To resolve this problem we will introduce the concept of a "training sample" as first proposed by Lempers (1971).

We will partition the data $\mathbf{y}$ as $\mathbf{y} = (\mathbf{w}, \mathbf{z})$, where $\mathbf{w}$ denotes the training sample which will be used to provide improved prior information and $\mathbf{z}$ is the comparison sample. There are not any assumptions made about which observations comprise $\mathbf{w}$ from within $\mathbf{y}$ here, where we will resolve the selection of $\mathbf{w}$ in Section 2.2.2.

Using $\mathbf{w}$ to derive posterior distributions gives

$$p_i(\theta_i|\mathbf{w}) = \frac{\pi_i(\theta_i)f_i(\mathbf{w}|\theta_i)}{m_i(\mathbf{w})} \tag{2.4}$$

where

$$m_i(\mathbf{w}) = \int_{\Omega_i} \pi_i(\theta_i)f_i(\mathbf{w}|\theta_i)d\theta_i. \tag{2.5}$$

If $\pi_i(\theta_i) = c_i k_i(\theta_i)$ then $p_i(\theta_i|\mathbf{w})$, from (2.4) above, contains c_i in its numerator and denominator but as long as the integral (2.5) for $m_i(\mathbf{w})$ converges then the c_i's may be cancelled and $p_i(\theta_i|\mathbf{w})$ does not contain c_i.

So when the training data $\mathbf{w}$ produces proper posterior distributions we are able to produce the Bayes factor using the comparison sample $\mathbf{z}$ as

$$B_{12}(\mathbf{z}|\mathbf{w}) = \frac{m_1(\mathbf{z}|\mathbf{w})}{m_2(\mathbf{z}|\mathbf{w})}$$

where

$$m_i(\mathbf{z}|\mathbf{w}) = \int_{\Omega_i} p_i(\theta_i|\mathbf{w})f_i(\mathbf{z}|\mathbf{w}, \theta_i)d\theta_i. \tag{2.6}$$

$B_{12}(\mathbf{z}|\mathbf{w})$ is called a partial Bayes factor and is now properly defined because any unspecified constants in the prior distributions $\pi_i(\theta_i)$ will have been cancelled out when calculating the posterior distributions shown in (2.4) above and so will not be present in $m_i(\mathbf{z}|\mathbf{w})$. Substituting (2.4) into (2.6) shows that

$$m_i(\mathbf{z}|\mathbf{w}) = \frac{m_i(\mathbf{y})}{m_i(\mathbf{w})}$$

and hence

$$B_{12}(\mathbf{z}|\mathbf{w}) = \frac{B_{12}(\mathbf{y})}{B_{12}(\mathbf{w})}. \tag{2.7}$$

2.2.2 Fractional Bayes factors

Whilst partial Bayes factors allow the comparison of models based on improper prior distributions, they introduce the partition of the full data $\mathbf{y}$ into the two parts $\mathbf{w}$ and $\mathbf{z}$ where there are not any assumptions made about which observations comprise $\mathbf{w}$ and $\mathbf{z}$.

If the full data $\mathbf{y}$ contains n observations and the training sample $\mathbf{w}$ contains g observations then let $t = g/n$ define the training fraction, the proportion of the observations used to provide improved prior information.

O'Hagan (1991) proposed the use of a training fraction and showed in O'Hagan (1995) that, asymptotically, the likelihood $f_i(\mathbf{w}|\boldsymbol{\theta}_i)$, based only on the training sample $\mathbf{w}$, behaves as the full likelihood $f_i(\mathbf{y}|\boldsymbol{\theta}_i)$ raised to the power t.

Equation (2.7) may be rearranged as

$$\begin{aligned} B_{12}(\mathbf{z}|\mathbf{w}) &= \frac{B_{12}(\mathbf{y})}{B_{12}(\mathbf{w})} \\ &= \frac{m_1(\mathbf{y})}{m_2(\mathbf{y})} \Big/ \frac{m_1(\mathbf{w})}{m_2(\mathbf{w})} \\ &= \frac{m_1(\mathbf{y})}{m_1(\mathbf{w})} \Big/ \frac{m_2(\mathbf{y})}{m_2(\mathbf{w})} \\ &= \frac{m_1^t(\mathbf{y})}{m_2^t(\mathbf{y})} \end{aligned}$$

where

$$m_i^t(\mathbf{y}) = \frac{m_i(\mathbf{y})}{m_i(\mathbf{w})} = \frac{\int_{\Omega_i} \pi_i(\boldsymbol{\theta}_i) f_i(\mathbf{y}|\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i}{\int_{\Omega_i} \pi_i(\boldsymbol{\theta}_i)[f_i(\mathbf{y}|\boldsymbol{\theta}_i)]^t d\boldsymbol{\theta}_i}$$

and we are then able to make an alternative definition of the Bayes factor (2.3) for comparing model M_1 against M_2 for the data $\mathbf{y}$ when using improper priors as

$$B_{12}^t(\mathbf{y}) = \frac{m_1^t(\mathbf{y})}{m_2^t(\mathbf{y})}$$

where $B_{12}^t(\mathbf{y})$ was designated the fractional Bayes factor (fBf) in O'Hagan (1995).

2.3 Power transformations

2.3.1 The General Power Transformation

The General Power Transformation (GPT) was introduced by Box and Cox (1964) as

$$X = \begin{cases} \lambda^{-1}(Y^{\lambda} - 1) & : \quad \lambda \neq 0 \\ \log Y & : \quad \lambda = 0 \end{cases}$$

for any strictly positive Y, where they worked in a non-Bayesian framework using maximum likelihood techniques to determine λ.

2.3.2 The Simple Power Transformation

The Simple Power Transformation (SPT) is defined as

$$X = Y^{\lambda}$$

where we will work with any strictly positive value of Y.

2.3.3 Choice of range for λ

In Section 1.6 we introduced the pMDI+ data set. This data set contains observed values in the range 48 to 26201 and is positively skewed. We will confine our analysis to positively skewed data sets.

So for any $0 \leq \lambda \leq 1$, for a SPT we can transform $Y \in (0, \infty)$ to $X \in (0, \infty)$ whereas for a GPT we can transform $Y \in (0, \infty)$ to $X \in (-1/\lambda, \infty)$, where the choice of λ is influenced by the degree of skewness. As $\lambda \longrightarrow 0$, increasing degrees of positive skewness are reduced.

After initially taking a power transformation, where we will work with λ in the range $\lambda \in [0, 1]$, we will now formulate a Bayesian analysis using Bayes factors to compare candidate models.

2.4 The Normal model

2.4.1 The model

To look for a realistic model for the random variable Y, observed as y, we will make a General Power Transformation from Y to the normally distributed random variable X as

$$X = \lambda^{-1}(Y^{\lambda} - 1)$$

where this transformation is continuous at $\lambda = 0$.

As X is defined on $X \in (-1/\lambda, \infty)$, unless $\lambda = 0$ this transformation will only be an approximation as a normally distributed random variable is defined over the range $(-\infty, \infty)$. Particular care is needed when $1/\lambda$ takes values close to 1, say less than 5, as the approximation will be at its most crude at best.

If $X \sim \mathrm{N}(\mu, \sigma^2)$ then

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

and when $Y = g(X)$ then

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}[g^{-1}(y)]\, f_X[g^{-1}(y)] \\ &= \frac{d}{dy}[\lambda^{-1}(y^{\lambda} - 1)] f_X(x). \end{aligned}$$

So

$$f_\lambda(y) = y^{\lambda-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

and

$$\begin{aligned} f_\lambda(\mathbf{y}) &= \prod f_\lambda(y_i) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \left(\prod y_i\right)^{\lambda-1} \exp\left[-\frac{1}{2\sigma^2}\sum(x_i - \mu)^2\right] \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \left(\prod y_i\right)^{\lambda-1} \exp\left[-\frac{1}{2\sigma^2}\{S_\lambda + n(\bar{x}_\lambda - \mu)^2\}\right] \end{aligned}$$

where $x_i = \lambda^{-1}(y_i^\lambda - 1)$ depends on λ although the index λ has been omitted for simplicity and $\sum(x_i - \mu)^2 = \sum((x_i - \bar{x}_\lambda) + (\bar{x}_\lambda - \mu))^2$, with $S_\lambda = \sum(x_i - \bar{x}_\lambda)^2$ and $\bar{x}_\lambda = \dfrac{1}{n}\sum x_i$.

To use this model for the population mean problem we need to be able to find $\mathbb{E}\{Y\} = \mathbb{E}\{(1+\lambda X)^{\frac{1}{\lambda}}\}$.

Whilst we are unable in general to find the expectation μ_λ of the $\frac{1}{\lambda}^{th}$ power of the random variable $(1+\lambda X)$ we are able to do it for the following special cases where Y is described as belonging to the rootNormal family of distributions indexed by the parameter λ

(i) $\lambda = 1 : Y \sim$ Normal and $Y = 1 + X$
with $\mu_1 = 1 + \mu$

(ii) $\lambda = \frac{1}{2} : Y \sim$ squarerootNormal and $Y = (1+\frac{1}{2}X)^2 = 1 + X + \frac{1}{4}X^2$
with $\mu_{\frac{1}{2}} = 1 + (\mu + \frac{1}{4}\sigma^2) + \frac{1}{4}\mu^2$

(iii) $\lambda = \frac{1}{3} : Y \sim$ cuberootNormal and $Y = (1+\frac{1}{3}X)^3 = 1 + X + \frac{1}{3}X^2 + \frac{1}{27}X^3$.
with $\mu_{\frac{1}{3}} = 1 + (\mu + \frac{1}{3}\sigma^2) + (\frac{1}{3}\mu^2 + \frac{1}{9}\mu\sigma^2) + \frac{1}{27}\mu^3$

(iv) For succeeding values of $\lambda = \frac{1}{m}$, where m is a positive integer, then it is always possible to find $\mathbb{E}\{Y\}$ by using the moment generating function of X, namely $\exp(\mu s + \frac{1}{2}\sigma^2 s^2)$.

However, as m increases, then Y will increasingly behave like the logNormal random variable that arises at the limiting value of $\lambda = 0$

(v) $\lambda = 0 : Y \sim$ logNormal and $Y = \exp(X)$
with $\mu_0 = \exp(\mu + \frac{1}{2}\sigma^2)$.

2.4.2 Bayesian analysis

The marginal density for a training sample $\mathbf{w}$ that represents a training fraction t of the data $\mathbf{y}$, whose General Power Transformation is indexed by λ, when the customary improper prior $\pi(\mu, \sigma^2) = (\sigma^2)^{-1}$ is specified, is

$$\begin{aligned} m_\lambda(\mathbf{w}) &= \int_0^\infty \int_{-\infty}^\infty \pi(\mu, \sigma^2) \left[f_\lambda(\mathbf{y})\right]^t d\mu d\sigma^2 = (2\pi)^{-\frac{nt}{2}} \left(\prod y_i\right)^{(\lambda-1)t} \\ &\times \int_0^\infty \int_{-\infty}^\infty (\sigma^2)^{-\frac{nt}{2}-1} \exp\left[-\frac{t}{2\sigma^2}\left\{S_\lambda + n(\bar{x}_\lambda - \mu)^2\right\}\right] d\mu d\sigma^2 \end{aligned}$$

and noting that

$$\begin{aligned} \int_{-\infty}^\infty \exp\left[-\frac{nt}{2\sigma^2}(\bar{x}_\lambda - \mu)^2\right] d\mu &= \frac{\sqrt{2\pi\sigma^2/nt}}{\sqrt{2\pi\sigma^2/nt}} \int_{-\infty}^\infty \exp\left[-\frac{(\bar{x}_\lambda - \mu)^2}{2\sigma^2/nt}\right] d\mu \\ &= \sqrt{2\pi/nt}(\sigma^2)^{\frac{1}{2}} \end{aligned}$$

we then obtain

$$\begin{aligned} m_\lambda(\mathbf{w}) &= (2\pi)^{-\frac{nt}{2}} \left(\prod y_i\right)^{(\lambda-1)t} \sqrt{2\pi/nt} \\ &\times \int_0^\infty (\sigma^2)^{-\frac{(nt-1)}{2}-1} \exp\left(-\frac{S_\lambda t}{2}\frac{1}{\sigma^2}\right) d\sigma^2 \qquad (2.8) \end{aligned}$$

where the integral part of (2.8) may be evaluated after making the transformation

$$z = \frac{S_\lambda t}{2}\frac{1}{\sigma^2}$$

as

$$\begin{aligned} \int_0^\infty (\sigma^2)^{-\frac{(nt-1)}{2}-1} \exp\left(-\frac{S_\lambda t}{2}\frac{1}{\sigma^2}\right) d\sigma^2 &= \left(\frac{S_\lambda t}{2}\right)^{-\frac{(nt-1)}{2}} \int_0^\infty z^{\frac{(nt-1)}{2}-1} \exp(-z) dz \\ &= \left(\frac{S_\lambda t}{2}\right)^{-\frac{(nt-1)}{2}} \Gamma\left(\frac{nt-1}{2}\right) \end{aligned}$$

and it follows that

$$\begin{aligned} m_\lambda(\mathbf{w}) &= (2\pi)^{-\frac{(nt-1)}{2}} \left(\prod y_i\right)^{(\lambda-1)t} \sqrt{\frac{1}{nt}}\Gamma\left(\frac{nt-1}{2}\right) \left(\frac{S_\lambda t}{2}\right)^{-\frac{(nt-1)}{2}} \\ &= (\pi S_\lambda)^{-\frac{(nt-1)}{2}} \left(\prod y_i\right)^{(\lambda-1)t} \sqrt{\frac{1}{n}}\Gamma\left(\frac{nt-1}{2}\right) t^{-\frac{nt}{2}}. \qquad (2.9) \end{aligned}$$

For $m_\lambda(\mathbf{w})$ to be finite it is necessary that $\Gamma\left(\frac{nt-1}{2}\right)$ exists and so, from the Glossary, this means that $nt - 1 > 0$ or the training fraction $t > 1/n$.

Therefore

$$\begin{aligned} m_\lambda^t(\mathbf{y}) = \frac{m_\lambda(\mathbf{y})}{m_\lambda(\mathbf{w})} &= \frac{(\pi S_\lambda)^{-\frac{(n-1)}{2}} \left(\prod y_i\right)^{\lambda-1} \sqrt{\frac{1}{n}}\Gamma\left(\frac{n-1}{2}\right) 1^{-\frac{n}{2}}}{(\pi S_\lambda)^{-\frac{(nt-1)}{2}} \left(\prod y_i\right)^{(\lambda-1)t} \sqrt{\frac{1}{n}}\Gamma\left(\frac{nt-1}{2}\right) t^{-\frac{nt}{2}}} \\ &= (\pi S_\lambda)^{-\frac{n(1-t)}{2}} \left(\prod y_i\right)^{(\lambda-1)(1-t)} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{nt-1}{2})} t^{\frac{nt}{2}}. \qquad (2.10) \end{aligned}$$

Hence, when the General Power Transformation indexed by λ of the random variable Y has a Normal distribution then, when making a model selection from within this rootNormal family, the fBf for comparing model 1 (λ_1) vs model 2 (λ_2) for data $\mathbf{y}$ becomes

$$\begin{aligned} B_{\lambda_1,\lambda_2}^t(\mathbf{y}) &= \frac{(\pi S_{\lambda_1})^{-\frac{n(1-t)}{2}} \left(\prod y_i\right)^{(\lambda_1-1)(1-t)}}{(\pi S_{\lambda_2})^{-\frac{n(1-t)}{2}} \left(\prod y_i\right)^{(\lambda_2-1)(1-t)}} \\ &= \left(\frac{S_{\lambda_1}}{S_{\lambda_2}}\right)^{-\frac{n(1-t)}{2}} \left(\prod y_i\right)^{(\lambda_1-\lambda_2)(1-t)} \\ &= \left[\left(\frac{S_{\lambda_1}}{S_{\lambda_2}}\right)^{-\frac{n}{2}} \left(\prod y_i\right)^{(\lambda_1-\lambda_2)}\right]^{(1-t)} \end{aligned}$$

and in particular when comparing model 1 against the logNormal for model 2, $\lambda_2 = 0$, then

$$B_{\lambda_1,0}^t(\mathbf{y}) = \left[\left(\frac{S_{\lambda_1}}{S_0}\right)^{-\frac{n}{2}} \left(\prod y_i\right)^{\lambda_1}\right]^{(1-t)} \qquad (2.11)$$

where, when $\lambda = 0$, $x_i = \log y_i$ and $\bar{x}_0 = \dfrac{1}{n}\sum \log y_i$ with $S_0 = \sum(\log y_i - \bar{x}_0)^2$.

2.5 The Gamma model

2.5.1 The model

To look for a realistic model for the random variable Y, observed as y, we only need to make a Simple Power Transformation from Y to X, where X is a Gamma random variable, as

$$X = Y^{\lambda}$$

because X has a strictly positive range.

Suppose $X \sim \mathrm{G}(\alpha, \beta)$ then

$$f_X(x) = \frac{\alpha^{\beta}}{\Gamma(\beta)} x^{\beta-1} \exp(-\alpha x)$$

and when $Y = g(X)$ then

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}[g^{-1}(y)]\, f_X[g^{-1}(y)] \\ &= \frac{d}{dy}(y^{\lambda}) f_X(y^{\lambda}). \end{aligned}$$

So

$$\begin{aligned} f_{\lambda}(y) &= \lambda y^{\lambda-1} \frac{\alpha^{\beta}}{\Gamma(\beta)} (y^{\lambda})^{\beta-1} \exp[-\alpha(y^{\lambda})] \\ &= \frac{\lambda \alpha^{\beta}}{\Gamma(\beta)} y^{\lambda\beta-1} \exp(-\alpha y^{\lambda}) \end{aligned}$$

and

$$\begin{aligned} f_{\lambda}(\mathrm{y}) &= \prod f_{\lambda}(y_i) \\ &= \frac{\lambda^n \alpha^{\beta n}}{[\Gamma(\beta)]^n} \left(\prod y_i\right)^{\lambda\beta-1} \exp\left(-\alpha \sum y_i^{\lambda}\right). \end{aligned}$$

To use this for the population mean problem we need to be able to determine $\mathbb{E}\{Y\} = \mathbb{E}\{X^{\frac{1}{\lambda}}\}$ which we certainly can do for the special cases $\lambda = \frac{1}{m}$, where m is a positive integer, using the moment generating function, $(1-s/\alpha)^{-\beta}$, of X and noting that

$$\mathbb{E}\{X^m\} = \frac{\beta(\beta+1)\cdots(\beta+[m-1])}{\alpha^m}$$

(i) $\lambda = 1 : Y \sim$ Gamma and $Y = X$ with $\mu_1 = \frac{\beta}{\alpha}$

(ii) $\lambda = \frac{1}{2} : Y \sim$ squarerootGamma and $Y = X^2$ with $\mu_{\frac{1}{2}} = \frac{\beta(\beta+1)}{\alpha^2}$

(iii) $\lambda = \frac{1}{3} : Y \sim$ cuberootGamma and $Y = X^3$ with $\mu_{\frac{1}{3}} = \frac{\beta(\beta+1)(\beta+2)}{\alpha^3}$, etc etc.

2.5.2 Bayesian analysis

The marginal density for a training sample **w** that represents a training fraction t of the data **y**, whose Simple Power Transformation is indexed by λ, when the improper prior $\pi(\alpha,\beta) = \alpha^{-1}\beta^{-1}$ is specified, is

$$\begin{aligned}
m_\lambda(\mathbf{w}) &= \int_0^\infty \int_0^\infty \pi(\alpha,\beta)\,[f_\lambda(\mathbf{y})]^t\, d\alpha d\beta \\
&= \int_0^\infty \int_0^\infty \alpha^{-1}\beta^{-1}\frac{\lambda^{nt}\alpha^{nt\beta}}{[\Gamma(\beta)]^{nt}}\left(\prod y_i\right)^{t(\lambda\beta-1)} \exp\left(-\alpha t \sum y_i^\lambda\right) d\alpha d\beta \\
&= \int_0^\infty \int_0^\infty \frac{\lambda^{nt}\beta^{-1}}{[\Gamma(\beta)]^{nt}}\left(\prod y_i\right)^{t(\lambda\beta-1)} \alpha^{nt\beta-1}\exp\left(-\left(t\sum y_i^\lambda\right)\alpha\right) d\alpha d\beta
\end{aligned}$$

where the integral with respect to α above may be evaluated after making the transformation

$$z = \left(t\sum y_i^\lambda\right)\alpha$$

as

$$\begin{aligned}\int_0^\infty \alpha^{nt\beta-1}\exp\left(-\left(t\sum y_i^\lambda\right)\alpha\right)d\alpha &= \int_0^\infty \frac{z^{nt\beta-1}}{\left(t\sum y_i^\lambda\right)^{nt\beta-1}}\exp(-z)\frac{dz}{\left(t\sum y_i^\lambda\right)} \\ &= \frac{\Gamma(nt\beta)}{\left(t\sum y_i^\lambda\right)^{nt\beta}}\end{aligned}$$

and it follows that

$$m_\lambda(\mathbf{w}) = \lambda^{nt}\int_0^\infty \frac{(\prod y_i)^{t(\lambda\beta-1)}}{\left(t\sum y_i^\lambda\right)^{nt\beta}}\frac{\Gamma(nt\beta)}{\beta[\Gamma(\beta)]^{nt}}d\beta. \tag{2.12}$$

To establish the conditions under which this integral converges we will initially look at small β, remembering from Section 2.2.2 that $nt = g$ is the number of observations in the training sample $\mathbf{w}$ and we can certainly do this, as we will show below, when $nt \geq 1$ is a positive integer.

Looking at the integrand in (2.12) above, considering this as a function of β, then as $\beta \longrightarrow 0$

$(\prod y_i)$ is a constant and so $(\prod y_i)^{t(\lambda\beta-1)} \longrightarrow (\prod y_i)^{-t}$ which is a constant
$\left(t\sum y_i^\lambda\right)$ is a constant and so $\left(t\sum y_i^\lambda\right)^{nt\beta} \longrightarrow 1$.

Hence to investigate the conditions under which this integral converges we have to examine how

$$\frac{\Gamma(nt\beta)}{\beta[\Gamma(\beta)]^{nt}}$$

behaves as $\beta \longrightarrow 0$.

When $nt = 1$

$$\frac{\Gamma(nt\beta)}{\beta[\Gamma(\beta)]^{nt}} = \frac{\Gamma(\beta)}{\beta\Gamma(\beta)} = \frac{\Gamma(\beta)}{\Gamma(\beta+1)} \to \infty \qquad \text{as } \beta \to 0.$$

If we use the Gauss multiplication theorem, defined in the Glossary, when $b = \beta$ and $m = nt$, then

$$\Gamma(\beta)\Gamma\left(\beta+\frac{1}{nt}\right)\Gamma\left(\beta+\frac{2}{nt}\right)\cdots\Gamma\left(\beta+\frac{nt-1}{nt}\right) = (2\pi)^{\frac{nt-1}{2}}(nt)^{\frac{1}{2}-nt\beta}\Gamma(nt\beta)$$

and so

$$\Gamma(nt\beta) = \Gamma(\beta)\Gamma\left(\beta+\frac{1}{nt}\right)\Gamma\left(\beta+\frac{2}{nt}\right)\cdots\Gamma\left(\beta+\frac{nt-1}{nt}\right)\frac{1}{(2\pi)^{\frac{nt-1}{2}}(nt)^{\frac{1}{2}-nt\beta}}$$

and hence

$$\frac{\Gamma(nt\beta)}{\beta[\Gamma(\beta)]^{nt}} = \frac{1}{\beta}\frac{\Gamma\left(\beta+\frac{1}{nt}\right)}{\Gamma(\beta)}\frac{\Gamma\left(\beta+\frac{2}{nt}\right)}{\Gamma(\beta)}\frac{\cdots}{\cdots}\frac{\Gamma\left(\beta+\frac{nt-1}{nt}\right)}{\Gamma(\beta)}\frac{1}{(2\pi)^{\frac{nt-1}{2}}(nt)^{\frac{1}{2}-nt\beta}}$$

and when $nt = 2$

$$\begin{aligned}\frac{\Gamma(nt\beta)}{\beta[\Gamma(\beta)]^{nt}} &= \frac{\Gamma(2\beta)}{\beta[\Gamma(\beta)]^{2}} = \frac{1}{\beta}\frac{\Gamma\left(\beta+\frac{1}{2}\right)}{\Gamma(\beta)}\frac{1}{(2\pi)^{\frac{2-1}{2}}2^{\frac{1}{2}-2\beta}} \\ &= \frac{\Gamma\left(\beta+\frac{1}{2}\right)}{\Gamma(\beta+1)}\frac{1}{\pi^{\frac{1}{2}}2^{1-2\beta}} \to \frac{1}{2} \qquad \text{as } \beta \to 0,\end{aligned}$$

but when $nt = 3$

$$\begin{aligned}\frac{\Gamma(nt\beta)}{\beta[\Gamma(\beta)]^{nt}} &= \frac{\Gamma(3\beta)}{\beta[\Gamma(\beta)]^{3}} = \frac{1}{\beta}\frac{\Gamma\left(\beta+\frac{1}{3}\right)}{\Gamma(\beta)}\frac{\Gamma\left(\beta+\frac{2}{3}\right)}{\Gamma(\beta)}\frac{1}{(2\pi)^{\frac{3-1}{2}}(3)^{\frac{1}{2}-3\beta}} \\ &= \frac{\Gamma\left(\beta+\frac{1}{3}\right)}{\Gamma(\beta+1)}\frac{\Gamma\left(\beta+\frac{2}{3}\right)}{\Gamma(\beta)}\frac{1}{2\pi 3^{\frac{1}{2}-3\beta}} \\ &\to \frac{\Gamma(\frac{1}{3})\Gamma(\frac{2}{3})}{\Gamma(\beta)}\frac{1}{2\pi 3^{\frac{1}{2}}} \to 0 \qquad \text{as } \beta \to 0,\end{aligned}$$

and for values of $nt \geq 4$

$$\frac{\Gamma(nt\beta)}{\beta[\Gamma(\beta)]^{nt}} \to 0 \qquad \text{as } \beta \to 0.$$

Hence we have shown that, for integer valued $nt > 1$, the integral (2.12) will converge when β is small.

We will now examine the conditions under which (2.12) converges for large β and appeal to Stirling's formula as

$$\begin{aligned}
\Gamma(b) &\simeq \sqrt{2\pi}\exp(-b)b^{b-\frac{1}{2}}\left(1+\frac{1}{12b}+\frac{1}{288b^2}+\cdots\right) \\
&\simeq \sqrt{2\pi}b^{b-\frac{1}{2}}\exp\left(-b+\frac{1}{12b}\right)
\end{aligned}$$

and so

$$\begin{aligned}
\frac{\Gamma(nt\beta)}{[\Gamma(\beta)]^{nt}} &\simeq \frac{\sqrt{2\pi}(nt\beta)^{nt\beta-\frac{1}{2}}\exp(-nt\beta+\frac{1}{12nt\beta})}{(\sqrt{2\pi})^{nt}(\beta^{\beta-\frac{1}{2}})^{nt}\exp(-nt\beta+\frac{nt}{12\beta})} \\
&= (2\pi)^{\frac{1-nt}{2}}(nt)^{nt\beta-\frac{1}{2}}\beta^{\frac{nt-1}{2}}\exp\left(\frac{1}{12nt\beta}-\frac{nt}{12\beta}\right). \qquad (2.13)
\end{aligned}$$

Hence (2.12) becomes approximately

$$\begin{aligned}
m_\lambda(\mathbf{w}) &\simeq \frac{\lambda^{nt}(2\pi)^{\frac{1-nt}{2}}}{\left(\prod y_i\right)^t(nt)^{\frac{1}{2}}} \\
&\times \int_0^\infty \beta^{\frac{nt-3}{2}}\left[\left(\prod y_i\right)^{t\lambda}n^{nt}/\left(\sum y_i^\lambda\right)^{nt}\right]^\beta \exp\left(\frac{1}{12nt\beta}-\frac{nt}{12\beta}\right)d\beta
\end{aligned}$$

and using the simplification

$$s_\lambda = \left(\prod y_i\right)^{t\lambda}n^{nt}/\left(\sum y_i^\lambda\right)^{nt}$$

the integral part of $m_\lambda(\mathbf{w})$ becomes

$$\int_0^\infty \beta^{\frac{nt-3}{2}}s_\lambda^\beta\exp\left(\frac{1}{12nt\beta}-\frac{nt}{12\beta}\right)d\beta$$

where for large β the integrand behaves like

$$\beta^{\frac{nt-3}{2}}s_\lambda^\beta.$$

Now

$$s_\lambda = \frac{\left[\left(\prod y_i^\lambda\right)^{\frac{1}{n}}\right]^{nt}}{\left[\sum y_i^\lambda/n\right]^{nt}} = \left[\frac{gm}{am}\right]^{nt}$$

where

$$gm = \left(\prod y_i^\lambda\right)^{\frac{1}{n}}$$

and

$$am = \sum y_i^\lambda / n.$$

Therefore

$$\begin{aligned} \beta^{\frac{nt-3}{2}} s_\lambda^\beta &= \beta^{\frac{nt-1}{2}-1} \exp[-(-\beta \log s_\lambda)] \\ &= \beta^{\frac{nt-1}{2}-1} \exp\{-[nt(\log am - \log gm)]\beta\} \\ &= \beta^{\frac{nt-1}{2}-1} \exp(-nts_\lambda^* \beta) \end{aligned}$$

where

$$s_\lambda^* = \log am - \log gm > 0 \qquad (2.14)$$

because $gm < am$ and s_λ^* also depends on $\mathbf{y}$ and n although this indexation has been omitted for simplicity.

So for large β the integral part of $m_\lambda(\mathbf{w})$ behaves like

$$\int_0^\infty \beta^{\frac{nt-1}{2}-1} \exp(-nts_\lambda^* \beta) d\beta$$

which converges for $nt > 1$.

So the results for both large and small β tell us that the integral on the right hand side of (2.12) for $m_\lambda(\mathbf{w})$ converges for $nt > 1$ or the training fraction $t \geq 2/n$.

We can then say, after using the substitution (2.14), that (2.12) will become approximately

$$\begin{aligned}
m_\lambda(\mathrm{w}) &\simeq \frac{\lambda^{nt}(2\pi)^{-\frac{nt-1}{2}}}{\left(\prod y_i\right)^t (nt)^{\frac{1}{2}}} \int_0^\infty \beta^{\frac{nt-3}{2}} \exp(-nts^*_\lambda \beta) \exp\left(\frac{1}{12nt\beta} - \frac{nt}{12\beta}\right) d\beta \\
&\simeq \frac{\lambda^{nt}(2\pi)^{-\frac{nt-1}{2}}}{\left(\prod y_i\right)^t (nt)^{\frac{1}{2}}} \int_0^\infty \beta^{\frac{nt-3}{2}} \exp(-nts^*_\lambda \beta) \frac{\Gamma(nt\beta)}{[\Gamma(\beta)]^{nt}} \frac{(2\pi)^{\frac{nt-1}{2}}}{(nt)^{nt\beta-\frac{1}{2}} \beta^{\frac{nt-1}{2}}} d\beta \\
&= \tilde{m}_\lambda(\mathrm{w}) \int_0^\infty \frac{(2\pi)^{\frac{nt-1}{2}} \Gamma(nt\beta)}{(nt)^{nt\beta-\frac{1}{2}} \beta^{\frac{nt-1}{2}} [\Gamma(\beta)]^{nt}} \frac{(nts^*_\lambda)^{\frac{nt-1}{2}} \beta^{\frac{nt-1}{2}-1} \exp(-nts^*_\lambda \beta)}{\Gamma(\frac{nt-1}{2})} d\beta \\
&= \tilde{m}_\lambda(\mathrm{w}) \mathbb{E}^t_\lambda\{g_t(\beta)\}
\end{aligned}$$

where

$$\tilde{m}_\lambda(\mathrm{w}) = \frac{\lambda^{nt}(2\pi)^{-\frac{nt-1}{2}} \Gamma(\frac{nt-1}{2})}{\left(\prod y_i\right)^t (nt)^{\frac{1}{2}} (nts^*_\lambda)^{\frac{nt-1}{2}}}$$

and

$$g_t(\beta) = \frac{(2\pi)^{\frac{nt-1}{2}} \Gamma(nt\beta)}{(nt)^{nt\beta-\frac{1}{2}} \beta^{\frac{nt-1}{2}} [\Gamma(\beta)]^{nt}}$$

where the expectation of $g_t(\beta)$ has been evaluated with respect to the Gamma distribution $\mathrm{G}(nts^*_\lambda, \frac{nt-1}{2})$. This suggests that a way of evaluating this integral is to use importance sampling from the $\mathrm{G}(nt(\log am - \log gm), \frac{nt-1}{2})$ distribution.

It also suggests (when $t = 1$) that this Gamma distribution is approximately the posterior distribution of β and hence that the posterior mean of the shape parameter β is approximately

$$\frac{1}{2s^*_\lambda} = \frac{1}{2(\log am - \log gm)}.$$

which becomes larger as λ reduces in value.

This in turn makes sense as s^*_λ is clearly some kind of measure of dispersion because, as λ reduces in value, the transformed data set has a smaller range with its reducing skewness and s^*_λ reflects this transformation with its reducing value.

Our theory suggests that if the posterior probability is concentrated on values of β that are reasonably large (ie s_λ^* is small enough and hence the scale factor nts_λ^* of the Gamma distribution for β) then, using (2.13)

$$g_t(\beta) \simeq \exp\left(\frac{1}{12nt\beta} - \frac{nt}{12\beta}\right) \simeq 1$$

and $m_\lambda(\mathbf{w}) \simeq \tilde{m}_\lambda(\mathbf{w})$.

The choice of the value of the training fraction t is discussed in Section 2.7, but if we let $t = 2/n$ then, by using Stirling's formula as developed in (2.13)

$$g_{\frac{2}{n}}(\beta) \simeq \exp\left(-\frac{1}{8\beta}\right)$$

and we can see that $g_{\frac{2}{n}}(0.2) \simeq 0.51$, $g_{\frac{2}{n}}(1.2) \simeq 0.90$ and $g_{\frac{2}{n}}(7.2) \simeq 0.98$.

Hence we would like our $\mathrm{G}(2s_\lambda^*, \frac{1}{2})$ to be concentrated on values of β that are mainly above 1.2. For $p\{\beta \leq 1.2\} \leq 0.1$ this occurs when $2s_\lambda^* \leq 0.0066$ or whenever $1/\lambda \geq 19$, for $p\{\beta \leq 1.2\} \leq 0.25$ this occurs when $2s_\lambda^* \leq 0.0424$ or whenever $1/\lambda \geq 8$.

For larger values of t then the requirement that $g_t(\beta) \simeq 1$ does becomes less restrictive as, for example, when $t = 4/n$ then $g_{\frac{4}{n}}(3.0) \simeq 0.90$ and $p\{\beta \leq 3\} \leq 0.1$ occurs when $4s_\lambda^* \leq 0.0975$ or whenever $1/\lambda \geq 7$.

We will show in Section 2.8.3, for the model comparisons undertaken and the pMDI+ data set, that the choice of the training fraction t does not alter the model choice - only the strength of that preference.

Although particular care is needed when $1/\lambda$ takes small values, say less than 7, when the approximation will be at its most crude, we can feel confident that the approximation for $m_\lambda^t(\mathbf{y})$ below does provide reasonable results over most of the range for $1/\lambda$ and we will be able to use it in Section 2.6.2

$$\frac{m_\lambda(\mathbf{y})}{m_\lambda(\mathbf{w})} \simeq \frac{\tilde{m}_\lambda(\mathbf{y})}{\tilde{m}_\lambda(\mathbf{w})} = \lambda^{n(1-t)}(2\pi s_\lambda^* n)^{-\frac{n(1-t)}{2}} t^{\frac{nt}{2}} \left(\prod y_i\right)^{-(1-t)} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{nt-1}{2})}. \qquad (2.15)$$

2.6 Model comparisons : theory

2.6.1 rootNormal vs logNormal

From (2.11) in Section 2.4.2, we know that for comparison of models within the rootNormal family as model 1 against the logNormal for model 2, $\lambda_2 = 0$, then the fBf is

$$B^t_{\lambda_1,0}(\mathbf{y}) = \left[\left(\frac{S_{\lambda_1}}{S_0}\right)^{-\frac{n}{2}} \left(\prod y_i\right)^{\lambda_1}\right]^{(1-t)}$$

which for the comparison of Normal, $\lambda_1 = 1$, vs logNormal becomes

$$B^t_{1,0}(\mathbf{y}) = \left[\left(\frac{S_1}{S_0}\right)^{-\frac{n}{2}} \left(\prod y_i\right)^{1}\right]^{(1-t)} \qquad (2.16)$$

for squarerootNormal, $\lambda_1 = \frac{1}{2}$, vs logNormal

$$B^t_{\frac{1}{2},0}(\mathbf{y}) = \left[\left(\frac{S_{\frac{1}{2}}}{S_0}\right)^{-\frac{n}{2}} \left(\prod y_i\right)^{\frac{1}{2}}\right]^{(1-t)} \qquad (2.17)$$

for cuberootNormal, $\lambda_1 = \frac{1}{3}$, vs logNormal becomes

$$B^t_{\frac{1}{3},0}(\mathbf{y}) = \left[\left(\frac{S_{\frac{1}{3}}}{S_0}\right)^{-\frac{n}{2}} \left(\prod y_i\right)^{\frac{1}{3}}\right]^{(1-t)} \qquad (2.18)$$

where the form of this relationship continues for values of $\lambda_1 = \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \ldots$.

2.6.2 logNormal vs rootGamma

From (2.10) in Section 2.4.2 we know that for a member of the rootNormal family

$$m^t_{\lambda_1}(\mathbf{y}) = \frac{m_{\lambda_1}(\mathbf{y})}{m_{\lambda_1}(\mathbf{w})} = (\pi S_{\lambda_1})^{-\frac{n(1-t)}{2}} \left(\prod y_i\right)^{(\lambda_1-1)(1-t)} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{nt-1}{2})} t^{\frac{nt}{2}}$$

which for the logNormal distribution, when $\lambda_1 = 0$, becomes

$$\frac{m_0(\mathbf{y})}{m_0(\mathbf{w})} = (\pi S_0)^{-\frac{n(1-t)}{2}} \left(\prod y_i\right)^{-(1-t)} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{nt-1}{2})} t^{\frac{nt}{2}}$$

whereas from (2.15) in Section 2.5.2 we have the approximation for a member of the rootGamma family

$$m^t_{\lambda_2}(\mathbf{y}) = \frac{m_{\lambda_2}(\mathbf{y})}{m_{\lambda_2}(\mathbf{w})} \simeq \frac{\tilde{m}_{\lambda_2}(\mathbf{y})}{\tilde{m}_{\lambda_2}(\mathbf{w})} = {\lambda_2}^{n(1-t)}(2\pi s^*_{\lambda_2} n)^{-\frac{n(1-t)}{2}} t^{\frac{nt}{2}} \left(\prod y_i\right)^{-(1-t)} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{nt-1}{2})}.$$

Hence the approximate fBf for the comparison of logNormal, model 1, vs rootGamma, model 2, is

$$\begin{aligned}
\tilde{B}^t_{0,\lambda_2}(\mathbf{y}) &= \frac{m_0(\mathbf{y})}{m_0(\mathbf{w})} \Big/ \frac{\tilde{m}_{\lambda_2}(\mathbf{y})}{\tilde{m}_{\lambda_2}(\mathbf{w})} \\
&= \frac{(\pi S_0)^{-\frac{n(1-t)}{2}} (\prod y_i)^{-(1-t)} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{nt-1}{2})} t^{\frac{nt}{2}}}{\lambda_2^{n(1-t)} (2\pi s^*_{\lambda_2} n)^{-\frac{n(1-t)}{2}} t^{\frac{nt}{2}} (\prod y_i)^{-(1-t)} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{nt-1}{2})}} \\
&= \left(\frac{S_0}{2s^*_{\lambda_2} n}\right)^{-\frac{n(1-t)}{2}} \lambda_2^{-n(1-t)} \\
&= \left[\left(\frac{S_0}{2s^*_{\lambda_2} n}\right)^{-\frac{n}{2}} \lambda_2^{-n}\right]^{(1-t)} \qquad (2.19)
\end{aligned}$$

which for the comparison of logNormal vs Gamma, $\lambda_2 = 1$, becomes

$$\tilde{B}^t_{0,1}(\mathbf{y}) = \left[\left(\frac{S_0}{2s^*_1 n}\right)^{-\frac{n}{2}} 1^{-n}\right]^{(1-t)} \qquad (2.20)$$

for logNormal vs squarerootGamma, $\lambda_2 = \frac{1}{2}$,becomes

$$\tilde{B}^0_{0,\frac{1}{2}}(\mathbf{y}) = \left[\left(\frac{S_0}{2s^*_{\frac{1}{2}} n}\right)^{-\frac{n}{2}} \left(\frac{1}{2}\right)^{-n}\right]^{(1-t)} \qquad (2.21)$$

and for logNormal vs cuberootGamma, $\lambda_2 = \frac{1}{3}$, becomes

$$\tilde{B}^0_{0,\frac{1}{3}}(\mathbf{y}) = \left[\left(\frac{S_0}{2s^*_{\frac{1}{3}} n}\right)^{-\frac{n}{2}} \left(\frac{1}{3}\right)^{-n}\right]^{(1-t)} \qquad (2.22)$$

where the form of this relationship continues for values of $\lambda_1 = \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \ldots$.

However note that the true value of the fBf, $B^t_{0,\lambda_2}(\mathbf{y})$, is the approximation (2.19) divided by

$$\frac{\mathbb{E}_{\lambda_2}\{g(\beta)\}}{\mathbb{E}^t_{\lambda_2}\{g_t(\beta)\}}$$

where this true value, considering the expectation of $g_t(\beta)$, is with respect to the Gamma distribution $\mathrm{G}(nts^*_{\lambda_2}, \frac{nt-1}{2})$ and is only defined when $t > \frac{1}{n}$.

2.7 Choice of training fraction

The choice of the size of the training fraction was discussed in O'Hagan (1995) and also in O'Hagan (1997) where the concept of a minimal training sample, n_0, was introduced and it was proposed that the minimal training fraction takes the value $t_0 = n_0/n$.

In Section 2.4 we introduced the Normal model and we know from Section 2.4.2 that the marginal density $m_\lambda(\mathrm{w})$ for the training sample w, defined in (2.9), will converge whenever the training fraction $t > 1/n$. Hence we can identify $n_0 = 2$ for this Normal model.

Similarly in Section 2.5 we introduced the Gamma model and from Section 2.5.2 we know that the marginal density $m_\lambda(\mathrm{w})$ for the training sample w, defined in (2.12), will converge whenever the training fraction $t > 1/n$. So once again we can identify $n_0 = 2$ for this Gamma model.

We will adopt here the minimal training fraction for t which becomes $t_0 = 2/n$ and, because $n = 26$ for the pMDI+ data set, $t_0 = 2/26 = 1/13$.

If robustness to misspecification of the prior or models is a concern then two other ways to set t were proposed in O'Hagan (1995), $t = n^{-1}\max\{n_0, \sqrt{n}\}$ or $n^{-1}\max\{n_0, \log n\}$, to generate values of $t = 5.099/26$ or $3.26/26$ respectively and while they alter the strength of the model preference we will show in Section 2.8.3 that they do not alter the choice.

2.8 Model comparisons : using the pMDI+ data

2.8.1 rootNormal vs logNormal

From (2.16) in Section 2.6.1 to compare the Normal, $\lambda_1 = 1$, vs logNormal, $\lambda_2 = 0$, for data $\mathbf{y}$, when $n = 26$ and $t = 1/13$, then the fBf is

$$B_{1,0}^{\frac{1}{13}}(\mathrm{y}) = \left[\left(\frac{S_1}{S_0}\right)^{-13} \left(\prod y_i\right)^{1}\right]^{\frac{12}{13}} \simeq 1.5 \times 10^{-26}$$

and so there is overwhelming evidence in favour of the logNormal over the Normal.

From Equation 2.17 in Section 2.6.1 to compare the squarerootNormal, $\lambda_1 = \frac{1}{2}$, vs logNormal for data $\mathbf{y}$, the fBf may be calculated as

$$B_{\frac{1}{2},0}^{\frac{1}{13}}(\mathrm{y}) = \left[\left(\frac{S_{\frac{1}{2}}}{S_0}\right)^{-13} \left(\prod y_i\right)^{\frac{1}{2}}\right]^{\frac{12}{13}} \simeq 4.8 \times 10^{-11}.$$

So there is very strong evidence in favour of the logNormal model over the squarerootNormal.

From Equation 2.18 in Section 2.6.1 to compare the cuberootNormal, $\lambda_1 = \frac{1}{3}$, vs logNormal for data $\mathbf{y}$, the fBf may be calculated as

$$B_{\frac{1}{3},0}^{\frac{1}{13}}(\mathrm{y}) = \left[\left(\frac{S_{\frac{1}{3}}}{S_0}\right)^{-13} \left(\prod y_i\right)^{\frac{1}{3}}\right]^{\frac{12}{13}} \simeq 8.4 \times 10^{-7}.$$

So there is strong evidence in favour of the logNormal over the cuberootNormal.

These calculations may be continued for other values of $\lambda = \frac{1}{n}$, where n is a positive integer, but are best represented by the figure below

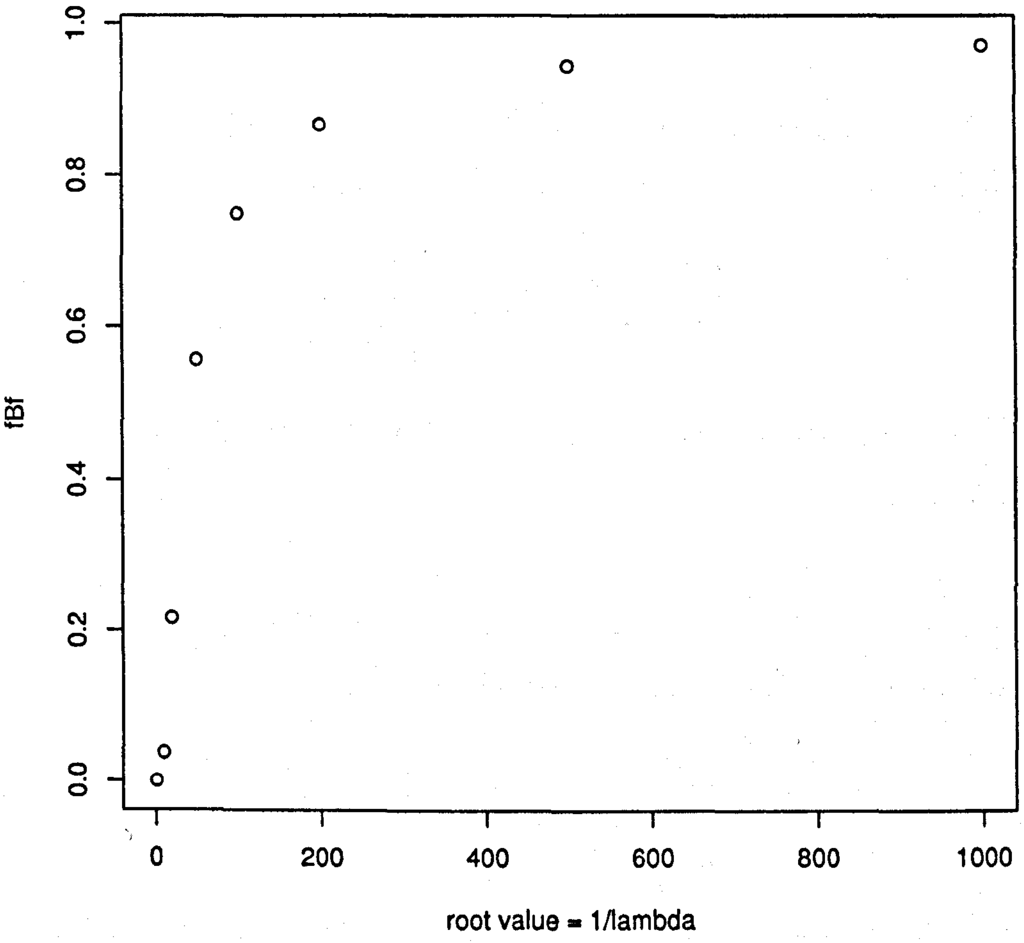


Figure 2.1: Plot showing the fBf for comparing the rootN vs logN models

where the value of the fBf → 1 as $\lambda \to 0$.

2.8.2 logNormal vs rootGamma

Equation 2.20 in Section 2.6.2 shows that for the comparison of logNormal vs Gamma, $\lambda_2 = 1$ for data $\mathbf{y}$, when $n = 26$ and $t = 1/13$, then the approximate fBf becomes

$$\tilde{B}_{0,1}^{\frac{1}{13}}(\mathbf{y}) = \left[\left(\frac{S_0}{52s_1^*}\right)^{-13} 1^{-26}\right]^{\frac{12}{13}} \simeq 627.$$

Hence there is overwhelming evidence in favour of the logNormal over the Gamma.

Using Equation 2.21 in Section 2.6.2 to compare the logNormal model vs squarerootGamma, $\lambda_2 = \frac{1}{2}$, for data $\mathbf{y}$, the approximate fBf becomes

$$\tilde{B}^{\frac{1}{13}}_{0,\frac{1}{2}}(\mathrm{y}) = \left[\left(\frac{S_0}{52 s^*_{\frac{1}{2}}}\right)^{-13}\left(\frac{1}{2}\right)^{-26}\right]^{\frac{12}{13}} \simeq 100.$$

So there is very strong evidence in favour of the logNormal model over the squarerootGamma.

Using Equation 2.22 in Section 2.6.2 to then compare the logNormal model vs cuberootGamma, $\lambda_3 = \frac{1}{3}$, for data $\mathbf{y}$, the approximate fBf becomes

$$\tilde{B}^{\frac{1}{13}}_{0,\frac{1}{3}}(\mathrm{y}) = \left[\left(\frac{S_0}{52 s^*_{\frac{1}{3}}}\right)^{-13}\left(\frac{1}{3}\right)^{-26}\right]^{\frac{12}{13}} \simeq 25.$$

Therefore there is strong evidence in favour of the logNormal model over the cuberootGamma.

Also to compare the logNormal model vs fourthrootGamma, $\lambda_4 = \frac{1}{4}$, for data $\mathbf{y}$, the approximate fBf becomes

$$\tilde{B}^{\frac{1}{13}}_{0,\frac{1}{4}}(\mathrm{y}) = \left[\left(\frac{S_0}{52 s^*_{\frac{1}{4}}}\right)^{-13}\left(\frac{1}{4}\right)^{-26}\right]^{\frac{12}{13}} \simeq 11.$$

Therefore there is some evidence in favour of the logNormal model over the fourthrootGamma.

These calculations may be continued for other values of $\lambda = \frac{1}{n}$, where n is a positive integer, but are best represented by the figure below

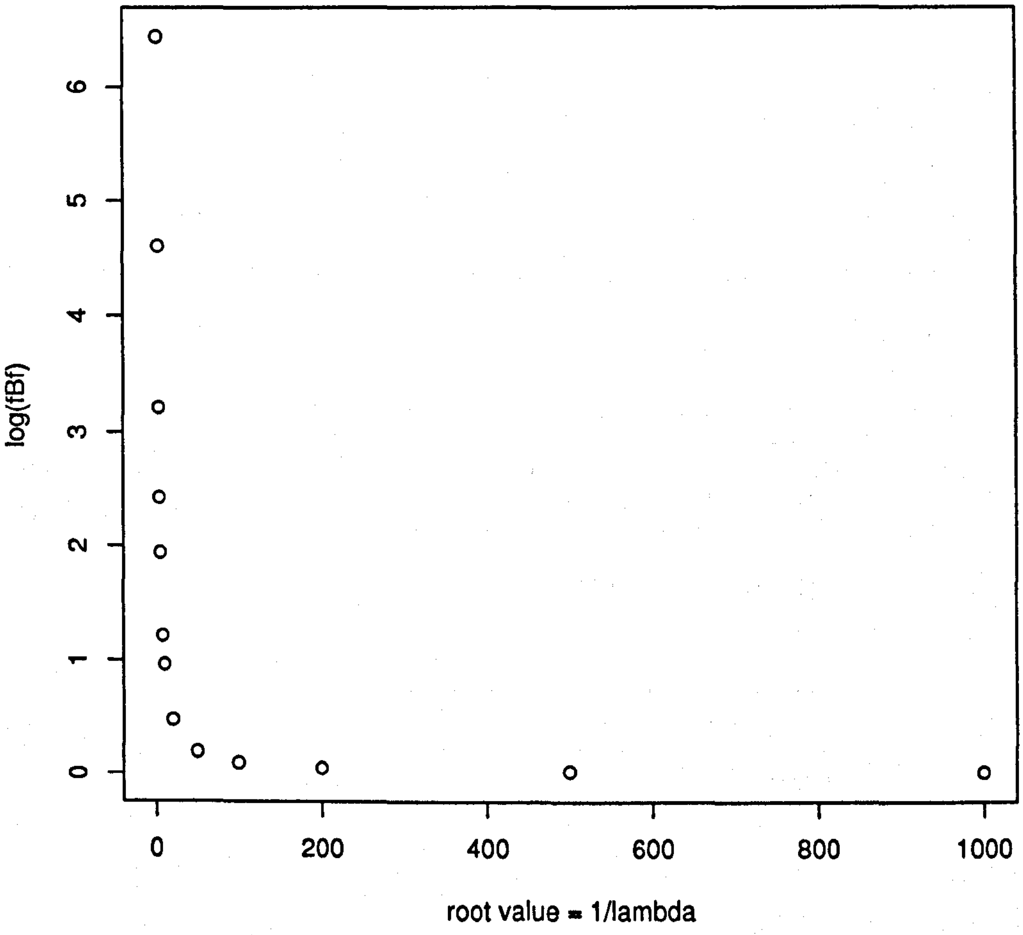


Figure 2.2: Plot showing log(fBf) for comparing the logN vs rootG models

where the value of the fBf → 1 as $\lambda \rightarrow 0$.

2.8.3 Conclusions

Composite models have been constructed using a power transformation followed by a parametric model which, when improper prior beliefs are specified, can be compared using fractional Bayes factors.

These comparisons have been undertaken using the minimal training fraction $t_0 = n_0/n$ which in this case was $t = 2/26$. The choice of size of training fraction influences the strength of model preference but it does not influence the model choice as is shown in the sample results below

	rootN vs logN	logN vs rootG
tn	$\lambda = 0.1$	$\lambda = 0.2$
(0)	(0.0288)	(8.1976)
2	0.0378	6.9727
4	0.0497	5.9309
6	0.0653	5.0447
8	0.0858	4.2909
10	0.1127	3.6498
12	0.1480	3.1045
14	0.1945	2.6406
16	0.2555	2.2460
18	0.3357	1.9104
20	0.4410	1.6250
22	0.5794	1.3822
24	0.7612	1.1757
(26)	(1)	(1)

Table 2.1: fractional Bayes factors for two model comparisons

although the minimal training fraction yields stronger model preferences.

The logNormal was the favoured model over rootNormal and also rootGamma models for the pMDI+ data set. However, as $\lambda \rightarrow 0$, then the fBf could barely distinguish between the logNormal and rootGamma models.

The Simple Power Transformation, introduced in Section 2.3.2 and then used for the rootGamma models, is only a linear transformation of the General Power Transformation introduced in Section 2.3.1 and used for the rootNormal models. The value $\lambda = 0$ gives the same interpretation of a log transformation to both models.

The Normal family of distributions has a fixed (symmetric) shape and has one parameter that controls its location and another that controls its scale. However the Gamma distribution has one parameter that controls its scale and another that controls its shape and as its shape parameter increases, say for values above 5, while the resulting Gamma distribution is still positively skewed it does begin to approach the shape of a Normal distribution.

The logNormal model will be adopted in this thesis for two reasons. It is the dominant preference for the pMDI+ data set when undertaking the model comparisons above. It is a statistical model that has been widely chosen in many other applications and so, although our original motivation is the medical cost data set pMDI+, the wide applicability of the logNormal model gives the techniques of this thesis applications in fields far removed from medicine, such as astrophysics, see Kawahara *et al* (2008), transport, see Graham *et al* (2005), language, see Novotny and Drozd (2000), physics, see Nöllmann and Etchegoin (2001), reliability, see Steele (2008), agriculture, see Korpalski *et al* (2005) and finance, see Al-Eideh *et al* (2004). Further applications within medicine include survival analysis, see Mould *et al* (2002), surgical procedure times, see May *et al* (2000) and radiography, see Neti and Howell (2006).

Chapter 3

The logNormal data model with noninformative prior beliefs

3.1 Introduction

Any Health Care Provider has responsibility for a potentially very large but finite population. So for the scenario described here, when some results are available following a clinical trial, then these trial data represent a sample from the finite population. To be able to budget for costs in future years the Health Care Provider needs to know the expected costs, or mean value, for the other members of their population who were not part of the clinical trial. For our Bayesian solution to this problem in Sections 3.2 and 3.3 we will use super-population theory, which was first proposed by classical statisticians, to model the population structure.

Initially a numerical solution was obtained as will be described in Section 3.4.2 but the results appeared somewhat surprising and did not show the stability for the mean that was expected as the number of samples was increased. This prompted a closer look at an analytical approach which will be described in Section 3.4.3.

There are few models for which an analytical approach will yield a posterior predictive distribution that has a recognisable parametric form. In this case not only was this possible but we will be able to show that all posterior predictive moments for one unobserved member of the population were infinite.

3.2 Super-population approach

3.2.1 Classical sampling theory

Classical sampling theory, in the main, is concerned with sampling from finite populations. The characteristics, or variables, that are measured (where it is assumed that this can be undertaken without error) are commonly nominal or ordinal, as well as metric (interval or ratio). Classical sampling theory is also known as *design based* because it concentrates on sampling design.

A particular difficulty for inference in classical sampling theory arises because the sampling mechanism is ancillary.

Cassel *et al* (1993) contains a useful theoretical introduction to the two main classical approaches to inference when sampling from finite populations, namely the fixed population approach and the super-population approach.

In the fixed population approach, which was developed first, the sampling design introduces the only source of randomness.

The super-population approach is a more recent innovation although an early reference to the use of the concept, although not the actual term, is to be found in Cochrane (1939). The difficulties with the use of super-population theory for classical inference are explored in O'Hagan and Forster (2004) and typically entail abandoning the classical statistician's strict reliance on frequency probability.

The super-population approach is also known within survey sampling as *model based design.*

3.2.2 Super-population approach

When the question of inference for finite populations from samples was considered in more detail then attention became focused on the structure of the population. The concept of the super-population supposes that the finite population of interest has itself been generated as a random sample of N units from some underlying infinite population, known as a super-population. If the variable of interest can be assumed to follow a parametric model in the super-population then the values associated with each unit are the observed outcome of the random variable. The values of the finite population not in the sample can be related to those in the sample using the assumed super-population distribution.

3.3 Bayesian super-population approach

3.3.1 Introduction

We will consider here a finite population comprising N units where we wish to examine the value of a variable $Y_i : i = 1, 2, \ldots, N$, for each unit or member of the population. The approach assumes a model where the Y_i's are exchangeable.

So the Y_i's are N members of the super-population whose members follow a parametric distribution with unknown parameter $\boldsymbol{\theta}$. To be even more specific we will make the stronger assumption that the Y_i's are independent and identically distributed with prior distribution $\pi(\boldsymbol{\theta})$ to represent the uncertainty about the parameter $\boldsymbol{\theta}$. The observed y_i values may be considered the first n members of the finite population because it is not generally relevant whether the units have been chosen at random in Bayesian inference.

When the first n Y_i's have been observed as y_i's the Bayesian super-population approach permits the $N - n$ unobserved members of the finite population to be estimated from their posterior predictive distribution.

3.3.2 Predicting future observations in general

We have decided in Chapter 2 to adopt the logNormal distribution as our data model for the random variable Y, observed as y, where $\log Y = X \sim \mathrm{N}(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\mu, \sigma^2)$ and can therefore simplify our notation by replacing f_0 by f, S_0 by S and $\bar{x}_0$ by $\bar{x}$.

So when data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ have been observed for the n independent identically distributed random variables $(Y_1, Y_2, \ldots, Y_n)$, which we will denote as $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ then

$$f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$$

and if prior beliefs about the uncertainty in the value for the parameter $\boldsymbol{\theta}$ are expressed as $\pi(\boldsymbol{\theta})$ then the posterior distribution of $\boldsymbol{\theta}$ becomes

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta}) \times f(\mathbf{y}|\boldsymbol{\theta})$$

where to predict future observations $\mathbf{y}_N = (y_{n+1}, y_{n+2}, \ldots, y_N)$ of the random variables $(Y_{n+1}, Y_{n+2}, \ldots, Y_N)$, denoted as $\mathbf{Y}_N = (Y_{n+1}, Y_{n+2}, \ldots, Y_N)$, when data $\mathbf{y}$ have been observed, we will use the posterior predictive distribution of $\mathbf{Y}_N$ which is $h(\mathbf{y}_N|\mathbf{y})$.

This is obtained from the joint posterior density of $\mathbf{Y}_N$ and $\boldsymbol{\theta}$, $\mathrm{h}(\mathbf{y}_N, \boldsymbol{\theta}|\mathbf{y})$ by integrating out $\boldsymbol{\theta}$, as

$$\begin{aligned} h(\mathbf{y}_N|\mathbf{y}) &= \int h(\mathbf{y}_N, \boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &= \int f(\mathbf{y}_N|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &= \int f(\mathbf{y}_N|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \qquad (3.1) \end{aligned}$$

where we have made the assumption that $\mathbf{Y}_N$ is conditionally independent of $\mathbf{Y}$ given $\boldsymbol{\theta}$.

3.3.3 Predicting future observations for Health Care costs

For the medical cost problem described here we wish to make predictions about the cost of treating an illness for our finite population where N is known. Let T denote the total cost, where $T = Y_1 + Y_2 + \cdots + Y_N$.

We have already observed $= (y_1, y_2, \ldots, y_n)$ and so we wish to determine the properties of

$$\begin{aligned} T &= y_1 + y_2 + \cdots + y_n + Y_{n+1} + Y_{n+2} + \cdots + Y_N \\ &= \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} Y_i. \end{aligned} \tag{3.2}$$

The posterior predictive distribution of $\sum_{i=n+1}^{N} Y_i$ was determined, following Equation (3.1), as

$$h(y_{n+1} + y_{n+2} + \cdots + y_N|\mathbf{y}) = \int f(y_{n+1} + y_{n+2} + \cdots + y_N|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}.$$

To enable posterior predictions for the sum of the unobserved members of the population to be made it is necessary to know the distribution of the sum of those unobserved values, as well as the posterior distribution of $\boldsymbol{\theta}$.

The distribution of the sum of the random variables, $Y_{n+1} + Y_{n+2} + \cdots + Y_N|\boldsymbol{\theta}$, is straightforward for some distributions like the Normal but is not available for many other distributions including the logNormal.

In Chapter 1 we introduced the typical medical problem under discussion here when a Health Care Provider needs to budget for the cost of an intervention for a finite population suffering from the disease. Hence we are interested in the mean population cost and we will concentrate on determining the posterior predictive mean of an unobserved member of the population and so, from Equation 3.2

$$\mathbb{E}\{T|\mathbf{y}\} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \mathbb{E}\{Y_i|\mathbf{y}\}.$$

3.4 Data model

To take account of the potentially positively skewed nature of the data naturally leads to examination of those models that allow the representation of this positive skew. In Chapter 2 we used fractional Bayes factors to compare candidate models and the logNormal distribution was strongly favoured and will be used here.

We will continue to choose the same logNormal model for both the observed and unobserved members of the population.

3.4.1 Prior beliefs

When considering the choice of prior, due consideration has to be given to the sceptical nature of a Health Care Provider, or regulatory body like NICE, and a noninformative prior may well (need to) be chosen.

Frequentist inference may be characterised as concerned only with what the data say, whereas Bayesian inference is concerned with using all of the relevant evidence.

The basic principle of Bayesian inference is that all inferences are derived from the posterior distribution. Noninformative prior distributions, however, represent no prior information. The basis of this is that if we have a completely flat prior distribution that is a constant then the posterior density is proportional to the likelihood and all the information in the posterior comes from the data.

The motivation for noninformative priors has been to produce an objective Bayesian analysis by using prior distributions that have been determined by some rule. There is continuing difficulty in defining what is meant by noninformative and a lack of agreed noninformative priors in all but simple situations. The rationale behind noninformative priors is to let the data determine the inference that can be made.

The ideas behind noninformative priors are introduced in most Bayesian text books, such as Gelman *et al* (1995), O'Hagan and Forster (2004) and Spiegelhalter *et al* (2004). Kass and Wasserman (1996) have produced a survey of methods to select noninformative priors constructed by some formal rule. They assert that the fundamental ideas and methods were laid down in Jeffreys (1961) and whilst it is possible to make a number of objections to the Jeffreys prior it has provided the improper prior $\pi(\boldsymbol{\theta}) \propto 1/\sigma^2$ used in Section 2.4.2 as well as this Chapter.

There are a number of other formal rules that have followed Jeffreys' initial work, see, for example, the data-translated likelihood of Box and Tiao (1973) and the Berger-Bernado Method introduced by Bernado (1979). In a subsequent series of papers, mostly by Berger and Bernado, it was refined and applied to various problems, see, for example, Berger and Bernado (1989).

For the analytical approach developed in Chapter 2, to compare candidate models for costs, customary improper priors were chosen for the models. These improper prior beliefs are noninformative because they are so weak.

For numerical analysis noninformative prior beliefs have to be proper no matter how weak they may be. Furthermore, there may not be a direct analogy, as will be seen to be the case here, between the priors used for an analytical approach and those that are used for a comparable numerical analysis.

3.4.2 Numerical analysis

WinBUGS is a software package freely available from the World Wide Web that is relatively easy to use. These advantages positively support the appeal of the Bayesian approach to problems - which does usually require numerical solutions because analytical solutions are intractable except in particular cases.

However, it also requires a prior understanding of both Bayesian analysis and MCMC, Markov Chain Monte Carlo, techniques to produce useful and realistic results. Using that knowledge, WinBUGS will produce estimates of the mean, standard deviation, median and other percentiles for those posterior distributions of interest.

WinBUGS was used to examine this logNormal model which is specified in terms of its mean, μ, and precision, τ, (the inverse of its variance). Customary prior beliefs would be a Normal distribution for the mean and an Inverse-Gamma distribution for the variance, or Gamma distribution for the precision. So for the pMDI+ cost data set we have $n = 26$ observations, $\mathbf{y}$, of the random variable, Y, with data model $Y \sim \text{logN}(\mu, \sigma^2) \equiv \text{logN}(\mu, 1/\tau)$ and prior beliefs $\mu \sim \text{N}(a, b)$ and $\tau \sim \text{G}(c, d)$.

The noninformative priors for the mean and precision respectively that we will use are N(0,1000) and G(0.001,0.001), which while proper are extremely weak and are customarily used. The range of values that are most likely for μ and τ with this choice of priors will include any values that we might expect to see. The values of the parameters for these noninformative priors are chosen to be of the order that a sceptical regulatory body might want to see.

We will show later in this Section that Gamma priors for τ are not a good choice because they allow τ to take values that are very small and so σ^2 may take values that are very large. Hence G(0.001,0.001) is a particularly poor choice and developing appropriate priors will commence in Chapter 4.

The code that was used, in WinBUGS version 1.3, for the numerical analysis of this logNormal model to determine the properties of of one unobserved member of the population from its posterior predictive distribution is shown in the Appendix.

The model that has been outlined earlier leads to moments for its posterior predictive distributions as shown in this table

Num/1000			Percentiles		
of Samples	mean	sd	2.5	50	97.5
1	1499	8852	16.39	339.5	7004
10	1536	26730	15.80	349.3	7749
100	2028	222900	16.62	354.4	7708
1000	1401	71880	16.22	355.4	7797
5000	1338	38920	16.33	356.5	7787

Table 3.1: logNormal model for 1 predictive value

WinBUGS is using proper distributions and a finite number of samples to produce its output which will in turn always be finite. If the mean, for example, of the posterior distribution of interest is infinite according to an analytical analysis, then this may not be immediately obvious from the WinBUGS output.

The graphical output ("density") from WinBUGS shows a highly positively skewed posterior predictive distribution with occasional very large values. The percentile values in the output are relatively robust for such a distribution but that is not true for the moments. The values that will be obtained may vary significantly as the posterior predictive values themselves will vary significantly.

The results in Table 3.1 above do not indicate any stability in the values of the posterior predictive moments shown as the number of samples is increased and question whether finite posterior predictive moments do exist.

It will not be at all easy in general, however, to determine from the WinBUGS output whether the posterior predictive distribution does possess finite moments. The logNormal model is a very specific case where an analytical solution is readily available to confirm that the numerical results do not indicate a finite first moment.

3.4.3 Analytical approach

Noninformative (improper) prior beliefs

If we assume that the prior beliefs for $\boldsymbol{\theta} = (\mu, \sigma^2)$ are independent for μ and σ^2 then the customary choice for noninformative prior beliefs will be the improper prior $\pi(\boldsymbol{\theta}) \propto 1/\sigma^2$, as used in Chapter 2, where the posterior distribution of $\boldsymbol{\theta}$ will be

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})}{\int\int \pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})d\mu d\sigma^2}$$

and, after cancelling the term $(\Pi y_i)^{-1}$, the denominator becomes

$$\begin{aligned}\int\int \pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})d\mu d\sigma^2 \;\propto\; & \int_0^\infty \int_{-\infty}^\infty (\sigma^2)^{-1}(2\pi\sigma^2)^{-\frac{n}{2}} \\ & \times\; \exp\left[-\frac{1}{2\sigma^2}\left\{S + n(\bar{x}-\mu)^2\right\}\right] d\mu d\sigma^2.\end{aligned}$$

If we consider the integral with respect to μ first, then

$$\int_{-\infty}^\infty \exp\left[-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}\right] d\mu = \sqrt{(2\pi\sigma^2/n)}$$

and the denominator becomes

$$\int_0^\infty (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-1-\frac{n}{2}+\frac{1}{2}}(2\pi/n)^{\frac{1}{2}} \exp\left(-\frac{S}{2}\frac{1}{\sigma^2}\right) d\sigma^2$$

which, after using the substitution $t = \frac{S}{2}\frac{1}{\sigma^2}$, becomes

$$(2\pi)^{-\frac{n}{2}}(2\pi/n)^{\frac{1}{2}}\left(\frac{S}{2}\right)^{-\frac{n}{2}+\frac{1}{2}} \int_0^\infty t^{\frac{n}{2}-\frac{1}{2}-1}\exp(-t)dt = (\pi S)^{-\frac{(n-1)}{2}}\frac{1}{\sqrt{n}}\Gamma\left(\frac{n-1}{2}\right).$$

So the posterior distribution of $\boldsymbol{\theta}$ is

$$\begin{aligned}p(\boldsymbol{\theta}|\mathbf{y}) = p(\mu, \sigma^2|\mathbf{y}) \;&=\; \frac{(\sigma^2)^{-1}(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\left\{S + n(\bar{x}-\mu)^2\right\}\right]}{(\pi S)^{-\frac{(n-1)}{2}}\frac{1}{\sqrt{n}}\Gamma\left(\frac{n-1}{2}\right)} \\ &=\; \frac{(\pi S)^{-\frac{1}{2}}}{(2/S)^{\frac{n}{2}}}\frac{\sqrt{n}}{\Gamma(\frac{n-1}{2})}(\sigma^2)^{-\frac{(n+2)}{2}}\exp\left[-\frac{1}{2\sigma^2}\left\{S + n(\bar{x}-\mu)^2\right\}\right].\end{aligned}$$

Hence the posterior predictive distribution for one unobserved member Y of the population becomes, from Equation (3.1) where y_N is y

$$\begin{aligned}
h(y|\mathbf{y}) &= \int f(y|\theta)p(\theta|\mathbf{y})d\theta \\
&= \int\int (2\pi\sigma^2)^{-\frac{1}{2}} y^{-1} \exp\left[-\frac{(\log y-\mu)^2}{2\sigma^2}\right] \\
&\times \frac{(\pi S)^{-\frac{1}{2}}}{(2/S)^{\frac{n}{2}}} \frac{\sqrt{n}}{\Gamma(\frac{n-1}{2})} (\sigma^2)^{-\frac{(n+2)}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{S+n(\bar{x}-\mu)^2\right\}\right] d\mu d\sigma^2 \\
&\propto \int\int (\sigma^2)^{-\frac{(n+3)}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{S+n(\bar{x}-\mu)^2+(\log y-\mu)^2\right\}\right] d\mu d\sigma^2 \\
&= \int\int (\sigma^2)^{-\frac{(n+3)}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{S+\frac{n}{n+1}(\log y-\bar{x})^2\right\}\right] \\
&\times \exp\left[-\frac{\left(\frac{n\bar{x}+\log y}{n+1}-\mu\right)^2}{2\sigma^2/(n+1)}\right] d\mu d\sigma^2 \\
&= \int (\sigma^2)^{-\frac{(n+3)}{2}} \sqrt{2\pi\sigma^2/(n+1)} \\
&\times \exp\left[-\frac{1}{2\sigma^2}\left\{S+\frac{n}{n+1}(\log y-\bar{x})^2\right\}\right] d\sigma^2 \\
&\propto \int (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left[-\left\{\frac{S+\frac{n}{n+1}(\log y-\bar{x})^2}{2}\right\}\frac{1}{\sigma^2}\right] d\sigma^2 \\
&= \frac{\Gamma(\frac{n}{2})2^{\frac{n}{2}}}{[S+\frac{n}{n+1}(\log y-\bar{x})^2]^{\frac{n}{2}}}.
\end{aligned}$$

So finally, after collecting together all the constants of proportionality

$$\begin{aligned}
h(y|\mathbf{y}) &= \frac{(\pi S)^{-\frac{1}{2}}}{(2/S)^{\frac{n}{2}}} \frac{\sqrt{n}}{\Gamma(\frac{n-1}{2})} (2\pi)^{-\frac{1}{2}} \sqrt{2\pi/(n+1)}\,\Gamma\left(\frac{n}{2}\right)(S/2)^{-\frac{n}{2}} \\
&\times y^{-1}\left[1+\frac{(\log y-\bar{x})^2}{S(n+1)/n}\right]^{-\frac{n}{2}} \\
&= \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi S(n+1)/n}} y^{-1}\left[1+\frac{(\log y-\bar{x})^2}{S(n+1)/n}\right]^{-\frac{n}{2}}.
\end{aligned}$$

Hence we can see from the Glossary that the posterior predictive distribution of Y follows a $\log t$ distribution, where $\nu = n-1$, $\mu = \bar{x}$ and $\sigma^2 = \frac{S(n+1)}{n(n-1)}$. We note also that S is only defined when integer valued $n \geq 2$ which tells us that ν is a positive integer as required.

To examine the expected value of the moments of this $\log t_{n-1}$ distribution we will use the substitution $t = \frac{\log y - \bar{x}}{\sqrt{S(n+1)/n}}$ to show that

$$\begin{aligned}
\mathbb{E}\{Y^r|\mathbf{y}\} &= \int_0^\infty y^r \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi S(n+1)/n}} y^{-1}\left[1+\frac{(\log y-\bar{x})^2}{S(n+1)/n}\right]^{-\frac{n}{2}} dy \\
&= \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi}} \int_{-\infty}^\infty \exp[r\{\bar{x}+\sqrt{S(n+1)/n}\,t\}]\left(1+t^2\right)^{-\frac{n}{2}} dt \\
&= \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi}} \exp(r\bar{x}) \int_{-\infty}^\infty \frac{\exp[r\sqrt{S(n+1)/n}\,t]}{(1+t^2)^{\frac{n}{2}}} dt.
\end{aligned}$$

As the denominator is $O(t^n)$ it is clear that there will be some value $L > 0$ such that for any $r \geq 1$, $n \geq 2$ and $S > 0$

$$\exp[r\sqrt{S(n+1)/n}\,t] > \left(1+t^2\right)^{\frac{n}{2}} \quad \text{for all} \quad t > L$$

and so $\mathbb{E}\{Y^r|\mathbf{y}\} = \infty$ for all values of r.

Natural conjugate prior beliefs

The likelihood for $\mathbf{y}|\boldsymbol{\theta}$ is

$$(\Pi y_i)^{-1}(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{S+n(\bar{x}-\mu)^2\right\}\right]$$

and the natural conjugate family of joint prior distributions for μ and σ^2 has a similar form to the likelihood as

$$\pi(\boldsymbol{\theta}) \propto \exp\left[-\frac{(a-\mu)^2}{2b\sigma^2}\right](\sigma^2)^{-d-\frac{3}{2}} \exp\left(-\frac{c}{\sigma^2}\right)$$

where $\mu|\sigma^2 \sim \mathrm{N}(a, b\sigma^2)$, $\sigma^2 \sim \mathrm{IG}(c, d)$ and

$$\pi(\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi b\sigma^2}} \exp\left[-\frac{(a-\mu)^2}{2b\sigma^2}\right] \quad \times \quad \frac{c^d(\sigma^2)^{-d-1}\exp(-\frac{c}{\sigma^2})}{\Gamma(d)}.$$

After following algebraic manipulation that is similar to the noninformative (improper) prior beliefs case above we are able to show that

(i) the posterior distribution of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}|\mathbf{y}) = p(\mu, \sigma^2|\mathbf{y}) \quad = \quad \frac{\exp\left[-\frac{\left(\frac{nb\bar{x}+a}{nb+1}-\mu\right)^2}{2[b/(nb+1)]\sigma^2}\right](\sigma^2)^{-d-\frac{n}{2}-\frac{3}{2}}\exp\left(-\frac{K}{\sigma^2}\right)}{[2\pi b/(nb+1)]^{\frac{1}{2}}\Gamma(d+\frac{n}{2})K^{-d-\frac{n}{2}}}$$

where $K = c + \frac{S}{2} + \frac{(a-\bar{x})^2}{2(nb+1)/n}$

(ii) the posterior predictive distribution for one unobserved member, Y, of the population is

$$\begin{aligned} h(y|\mathbf{y}) \quad &= \quad \frac{\Gamma(\frac{2d+n+1}{2})}{\Gamma(\frac{2d+n}{2})\sqrt{2\pi[(nb+1+b)/(nb+1)]\left[c+\frac{S}{2}+\frac{(a-\bar{x})^2}{2(nb+1)/n}\right]}} \\ &\times \quad y^{-1}\left[1+\frac{(\log y - \frac{nb\bar{x}+1}{nb+1})^2}{2[(nb+1+b)/(nb+1)]\left[c+\frac{S}{2}+\frac{(a-\bar{x})^2}{2(nb+1)/n}\right]}\right]^{-\frac{(2d+n+1)}{2}} \end{aligned}$$

and again we can see from the Glossary that the posterior predictive distribution of Y follows a $\log t$ distribution, with $\nu = 2d+n$ where $2d$ must be an integer, and

$$\mu = \frac{nb\bar{x}+1}{nb+1}$$

$$\sigma^2 = 2[(nb+1+b)/(nb+1)]\left[c+\frac{S}{2}+\frac{(a-\bar{x})^2}{2(nb+1)/n}\right]/(2d+n).$$

Hence we can show that the expected value of the moments of this $\log t_{2d+n}$ distribution are

$$\mathbb{E}\{Y^r|\mathbf{y}\} \propto \int_{-\infty}^{\infty} \frac{\exp\left(r\sqrt{2[(nb+1+b)/(nb+1)]\left[c+\frac{S}{2}+\frac{(a-\bar{x})^2}{2(nb+1)/n}\right]}t\right)}{(1+t^2)^{\frac{(2d+n+1)}{2}}}dt$$

and $\mathbb{E}\{Y^r|\mathbf{y}\} = \infty$ for all values of r.

This result holds for all logical values of the hyperparameters a, b, c and d (b, c and $d > 0$, with $2d$ an integer) where the hyperparameters may have values that make the prior beliefs extremely informative.

Numerical output from WinBUGS

Hence the customary noninformative, and indeed improper, prior cannot be used when finite posterior predictive moments are required. In fact the natural conjugate prior also yields posterior predictive moments that are not finite even if the prior beliefs are extremely informative.

This serves as a warning when using WinBUGS with this logNormal data model. Sampling techniques within WinBUGS will always allow numerical values to be obtained but careful examination of the output will be required to determine whether the true posterior predictive moments are finite.

Chapter 4

Developing prior beliefs for the logNormal data model

4.1 Introduction

In Chapter 2 we have adopted the logNormal distribution as the data model. The posterior predictive distribution was introduced in Chapter 3, using weak priors for the mean and variance of the logNormal distribution, and produced posterior predictive expected values that were infinite.

We will now need to consider how to represent the prior beliefs for the joint probability distribution of $\boldsymbol{\theta} = \mu, \sigma^2$ that ensure that the mean of the posterior predictive distribution exists, will allow numerical evaluation and also that we can elicit prior beliefs from an expert to determine the values of the hyperparameters in the joint distribution.

We will look to formulate the joint distribution in terms of two independent marginal distributions each of which will only involve (a function of) one of the parameters. These univariate distributions may be considerably easier to elicit than their joint distribution.

It is recognised as generally good practise to elicit beliefs about quantities that are directly observable. We will follow this practise wherever possible but will need a different approach when eliciting beliefs about the shape parameter σ^2.

In Section 4.2 we will consider three possible formulations for prior beliefs. The three models that will be examined are similar but each represents a different prior formulation for the same logNormal data model and bring their individual pro's and con's.

We will commence our study of the existence of posterior predictive moments in Section 4.3 by developing an analytically useful manipulation of the problem and in Section 4.4 we will examine a number of possible choices of models for our prior beliefs for the scale parameter of our data model.

In Section 4.5 we will examine the existence of posterior predictive moments when neither the scale parameter, μ, nor shape parameter, σ^2, for our data model are known, but we will be unable to obtain analytical solutions to the integrals involved. In Section 4.6 we will explore the value of the posterior expectation of the population mean for a particular class of prior beliefs choosing computational techniques that, by reducing the dimension of the problem, considerably improve the speed of computation.

4.2 Three possible models

4.2.1 Introduction

To enable us to consider possible models we note a number of useful properties of the logNormal distribution $Y \sim \text{logN}(\mu, \sigma^2)$

Quantity	Formula
mean	$\exp(\mu + \sigma^2/2)$
median	$\exp(\mu)$
mode	$\exp(\mu)/\exp(\sigma^2)$
variance	$\exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]$
coefficient of variation	$[\exp(\sigma^2) - 1]^{\frac{1}{2}}$
Φ(q) quantile	$\exp(\mu + q\sigma)$
t^{th} moment	$\exp(t\mu + t^2\sigma^2/2)$

Table 4.1: Quantities and their formulae for a logNormal distribution

whilst the coefficient of kurtosis and the coefficient of skewness are functions of σ^2 alone, but their formulae are even more complex.

The considerations that will motivate our exploration of possible models are the need to be able to elicit prior beliefs from an expert and to discover prior distributions that enable posterior moments to exist.

We will seek to work with two quantities, or possibly functions of quantities, that are independent. If the one of the quantities is only a function of the scale parameter μ and the other only the shape parameter σ^2 then considerable technical advantages become available when examining moments of the posterior predictive distribution and enable us to only have to elicit two univariate distributions.

So we can see immediately that the median is a (relatively simple) function of μ alone but to obtain a function of σ^2 alone may require the ratio of two quantities.

4.2.2 Mean and variance unknown

In this approach we will work with the mean $E_Y(\boldsymbol{\theta}) = \mathbb{E}\{Y|\boldsymbol{\theta}\}$ and variance $V_Y(\boldsymbol{\theta}) = \mathcal{V}\{Y|\boldsymbol{\theta}\}$ of the logNormal random variable Y, where to simplify the nomenclature we will now represent $E_Y(\boldsymbol{\theta})$ as E_Y and $V_Y(\boldsymbol{\theta})$ as V_Y .

We will model our prior beliefs for E_Y and V_Y with $\pi_{E_Y,V_Y}(e,v)$ where

$$e = \exp(\mu + \sigma^2/2) \qquad \text{and} \qquad v = e^2[\exp(\sigma^2) - 1]$$
$$\text{hence} \qquad \sigma^2 = \log\left(1 + \frac{v}{e^2}\right) \qquad \text{and}$$
$$\mu = \log e - \frac{1}{2}\log\left(1 + \frac{v}{e^2}\right) = \frac{1}{2}\log\left(\frac{e^2}{1 + v/e^2}\right) = \log\left(\frac{e}{\sqrt{1 + v/e^2}}\right).$$

We will transform our beliefs in the observation space, Y, into beliefs in the X space using the following usual methods

$$\begin{aligned}\pi(\mu, \sigma^2) &= \pi_{E_Y,V_Y}(\exp(\mu + \sigma^2/2), \exp(2\mu)[\exp(2\sigma^2) - \exp(\sigma^2)]) \times |J| \\ &= \exp(3\mu + 5\sigma^2/2)\pi_{E_Y,V_Y}(\exp(\mu + \sigma^2/2), \exp(2\mu)[\exp(2\sigma^2) - \exp(\sigma^2)])\end{aligned}$$

where

$$J = \det\begin{pmatrix} \frac{\partial e}{\partial \mu} & \frac{\partial e}{\partial \sigma^2} \\ \\ \frac{\partial v}{\partial \mu} & \frac{\partial v}{\partial \sigma^2} \end{pmatrix}.$$

Even if the prior $\pi_{E_Y,V_Y}(\cdot,\cdot)$ is assumed to factorize as $\pi_{E_Y}(\cdot)\pi_{V_Y}(\cdot)$ the function $\pi(\mu, \sigma^2)$ is complex because both of the terms $\pi_{E_Y}(\cdot)$ and $\pi_{V_Y}(\cdot)$ are functions of μ and of σ^2.

Elicitation of the mean (for a skewed distribution) and variance are generally considered difficult with unreliable answers, whilst eliciting their distributions has rarely been considered.

So although this is the customary model it will not be pursued here because of the elicitation difficulties.

4.2.3 Median and coefficient of variation unknown

For this approach we will work with the median and coefficient of variation of the logNormal random variable Y as

$$\begin{aligned} M_Y(\theta) &= \text{median } \{Y|\theta\} = \exp(\mu) \\ C_Y(\theta) &= \text{coefficient of variation } \{Y|\theta\} = [\exp(\sigma^2) - 1]^{\frac{1}{2}} \end{aligned}$$

where the median is a function of (only) the scale parameter, μ, of the logNormal distribution and the coefficient of variation is a function of (only) σ its shape parameter.

To simplify the nomenclature we will represent $M_Y(\theta)$ as M_Y and $C_Y(\theta)$ as C_Y .

We will model our prior beliefs for M_Y and C_Y with $\pi_{M_Y,C_Y}(m, r)$ where

$$m = \exp(\mu) \quad \text{and} \quad c = [\exp(\sigma^2) - 1]^{\frac{1}{2}}.$$

We will transform our beliefs in the observation space, Y, into beliefs in the X space using the following usual methods

$$\begin{aligned} \pi(\mu, \sigma^2) &= \pi_{M_Y,C_Y}(\exp(\mu), [\exp(\sigma^2) - 1]^{\frac{1}{2}}) \times |J| \\ &= \exp(\mu)\exp(\sigma^2)\frac{1}{2}[\exp(\sigma^2) - 1]^{-\frac{1}{2}}\pi_{M_Y,C_Y}(\exp(\mu), [\exp(\sigma^2) - 1]^{\frac{1}{2}}). \end{aligned}$$

If we assume that M_Y and R_Y are independent then we could now proceed using the considerable technical advantage that $\pi(\mu, \sigma^2)$ factorises into a function of μ alone and a function of σ^2 alone. However, the coefficient of variation is not directly observable and is a ratio of two different quantities, neither of which individually may be elicited easily.

This approach will not be pursued here.

4.2.4 Median and quantile ratio unknown

We will work with the median and quantile ratio of the logNormal random variable Y defined as

$$\begin{aligned}
M_Y(\boldsymbol{\theta}) &= \text{median } \{Y|\boldsymbol{\theta}\} = \exp(\mu) \\
R_Y(\boldsymbol{\theta}) &= \frac{\Phi(q) \text{ quantile } \{Y|\boldsymbol{\theta}\}}{\text{median } \{Y|\boldsymbol{\theta}\}} \\
&= \frac{\exp(\mu + q\sigma)}{\exp(\mu)} = \exp(q\sigma)
\end{aligned}$$

where the quantile ratio, $R_Y(\boldsymbol{\theta})$, is a function of (only) the shape parameter σ for a fixed value of q.

To simplify the nomenclature we will represent $M_Y(\boldsymbol{\theta})$ as M_Y and $R_Y(\boldsymbol{\theta})$ as R_Y .

We will model our prior beliefs for M_Y and R_Y with $\pi_{M_Y,R_Y}(m,r)$ where

$$m = \exp(\mu) \quad \text{and} \quad r = \exp(q\sigma).$$

We will transform our beliefs in the observation space Y into beliefs in the X space using the following usual methods

$$\begin{aligned}
\pi(\mu, \sigma^2) &= \pi_{M_Y,R_Y}(\exp(\mu), \exp(q\sigma)) \times |J| \\
&= \exp(\mu)\frac{q\exp(q\sigma)}{2\sigma}\pi_{M_Y,R_Y}(\exp(\mu), \exp(q\sigma)).
\end{aligned}$$

We will again assume that M_Y and R_Y are independent when we could proceed using the considerable technical advantage that $\pi(\mu, \sigma^2)$ factorises into a function of μ alone and a function of σ^2 alone.

This approach will be pursued in Section 4.5 where we will consider the case $q > 0$, where the choice of $q > 0$ will be discussed in more detail in Section 6.1, which in turn implies that $r \geq 1$ and we need to choose distributions to model our prior beliefs for R_Y whose support is $[1, \infty)$.

4.3 Posterior prediction

We proceed to evaluate the existence of posterior predictive moments as follows.

The posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ may be expressed as proportional to the product of our prior beliefs $\pi(\boldsymbol{\theta})$ and the likelihood of the observations $f(\mathbf{y}|\boldsymbol{\theta})$ as

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}).$$

The posterior predictive distribution of a future observation y, $h(y|\mathbf{y})$, will be obtained from the joint posterior distribution of y and $\boldsymbol{\theta}$, $h(y, \boldsymbol{\theta}|\mathbf{y})$, as

$$\begin{aligned} h(y|\mathbf{y}) &= \int h(y, \boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &= \int h(y|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &= \int f(y|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \end{aligned}$$

where we have made the assumption that Y is conditionally independent of $\mathbf{Y}$ given $\boldsymbol{\theta}$.

So to examine the existence of the tth posterior predictive moment of Y we evaluate

$$\begin{aligned} \mathbb{E}\{Y^t|\mathbf{y}\} &= \int y^t h(y|\mathbf{y}) dy \\ &= \int y^t \left[\int f(y|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \right] dy \\ &= \int \left[\int y^t f(y|\boldsymbol{\theta}) dy \right] p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &= \int j_t(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &\propto \int j_t(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} \end{aligned}$$

where $j_t(\boldsymbol{\theta}) = \int y^t f(y|\boldsymbol{\theta}) dy$ is the tth moment of Y and Fubini's Theorem, see Malliavin (1995) or Körner (1988), has been invoked to justify interchanging the order of integration.

To ensure the conditions for Fubini to be applicable are satisfied we restrict our interest to a logNormal model for the likelihood to ensure $j_t(\boldsymbol{\theta})$ exists (which may not be true for a Student's t distribution for example). Also to prior beliefs $\pi(\boldsymbol{\theta})$ such that $\mathbb{E}\{Y^t|\mathrm{y}\} < \infty$, which precludes such choices as the improper prior $\pi(\boldsymbol{\theta}) \propto 1/\sigma^2$ from Section 3.4.3.

We have been able to establish now that determining the 1^{st} posterior predictive moment of Y, $\mathbb{E}\{Y|\mathrm{y}\}$, is equivalent to determining the posterior population mean, $\mathbb{E}\{\exp(\mu + \sigma^2/2)|\mathrm{y}\}$, and we will work with whichever is most convenient in the future.

4.4 Prior beliefs when the median is unknown

4.4.1 Introduction

We will look at a range of commonly chosen models for prior beliefs to examine the influence that the tail thickness of the prior beliefs exert on the posterior predictive values.

We will start with the case that the quantile ratio is known and only the median is unknown where we know from Section 4.2.4 that

$$\pi(\mu) \propto \exp(\mu)\pi_{M_Y}(\exp(\mu))$$

and that

$$j_t(\boldsymbol{\theta}) = \exp(t\mu).$$

4.4.2 Gamma prior beliefs for the median

If $M_Y \sim \mathrm{G}(a,b)$ then

$$\pi(\mu) \propto \exp(\mu)[\exp(\mu)]^{(b-1)} \exp[-a\exp(\mu)] \qquad \text{and}$$

$$\begin{aligned}
\mathbb{E}\{Y^t|\mathbf{y}\} &\propto \int_{-\infty}^{\infty} \exp(t\mu)\exp(\mu)[\exp(\mu)]^{(b-1)}\exp[-a\exp(\mu)]\exp\left[-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}\right] d\mu \\
&= \int_0^{\infty} w^{b+t}\exp(-aw)\exp\left[-\frac{(\bar{x}-\log w)^2}{2\sigma^2/n}\right]\frac{dw}{w} \qquad \text{if w} = \exp(\mu) \\
&\leq \int w^{(b+t)-1}\exp(-aw)dw = \frac{\Gamma(b+t)}{a^{(b+t)}}.
\end{aligned}$$

This result establishes an upper bound for $E\{Y^t|\mathbf{y}\}$, a constant for fixed a, b and t. Hence the t^{th} posterior predictive moment always exists when $M_Y \sim \mathrm{G}(a,b)$.

4.4.3 logNormal prior beliefs for the median

This prior belief is chosen to illustrate a prior belief that possesses a thicker tail than the Gamma distribution, where if $M_Y \sim \mathrm{logN}(a,b)$ then

$$\pi(\mu) \propto \exp\left[-\frac{(a-\mu)^2}{2b}\right] \qquad \text{and}$$

$$\begin{aligned}
\mathbb{E}\{Y^t|\mathbf{y}\} &\propto \int_{-\infty}^{\infty} \exp(t\mu)\exp\left[-\frac{(a-\mu)^2}{2b}\right]\exp\left[-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}\right] d\mu \\
&= \exp\left[-\frac{(a-\bar{x})^2}{2(b+\sigma^2/n)}\right]\int \exp(t\mu)\exp\left[-\frac{\left(\frac{b\bar{x}+a\sigma^2/n}{b+\sigma^2/n}-\mu\right)^2}{2b\sigma^2/n(b+\sigma^2/n)}\right] d\mu \\
&= \exp\left[-\frac{(a-\bar{x})^2}{2(b+\sigma^2/n)}\right]\exp\left[t\frac{\left\{2\left(\frac{b\bar{x}+a\sigma^2/n}{b+\sigma^2/n}\right)+t\left(\frac{b\sigma^2/n}{b+\sigma^2/n}\right)\right\}}{2}\right] \\
&\times \int \exp\left[-\frac{\left\{\left(\frac{b\bar{x}+a\sigma^2/n}{b+\sigma^2/n}+t\frac{b\sigma^2/n}{b+\sigma^2/n}\right)-\mu\right\}^2}{2b\sigma^2/n(b+\sigma^2/n)}\right] d\mu
\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{\frac{2\pi b\sigma^2/n}{(b+\sigma^2/n)}} \exp\left[-\frac{(a-\bar{x})^2}{2(b+\sigma^2/n)}\right] \\
&\times \exp\left[t\frac{\{2(b\bar{x}+a\sigma^2/n)+bt\sigma^2/n\}}{2(b+\sigma^2/n)}\right] \qquad (4.1) \\
&\leq \sqrt{2\pi b} \times 1 \times K
\end{aligned}
$$

where the term $\exp\left[t\frac{\{2(b\bar{x}+a\sigma^2/n)+bt\sigma^2/n\}}{2(b+\sigma^2/n)}\right]$ is bounded by $\exp(t\bar{x})$ and $\exp(t\frac{2a+bt}{2})$. Which bound is upper and which bound is lower is determined by the values of $\bar{x}$ and $\frac{2a+bt}{2}$ where K will be the larger of $\exp(t\bar{x})$ and $\exp(t\frac{2a+bt}{2})$.

This result also establishes an upper bound for $E\{Y^t|\mathrm{y}\}$, a constant for fixed a, b, t and $\bar{x}$. Hence the t^{th} posterior predictive moment always exists whenever $M_Y \sim \mathrm{logN}(a, b)$.

4.4.4 First improper prior beliefs for the median

If we represent our prior beliefs about the median as $\pi_{M_Y}(m) \propto 1/m$ then

$$\pi(\mu) \propto 1 \qquad \text{and}$$

$$
\begin{aligned}
\mathbb{E}\{Y^t|\mathrm{y}\} &\propto \int_{-\infty}^{\infty} \exp(t\mu)\exp\left[-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}\right] d\mu \\
&= \exp\left[\frac{t(2\bar{x}+t\sigma^2/n)}{2}\right] \int \exp\left[-\frac{\{(\bar{x}+t\sigma^2/n)-\mu\}^2}{2\sigma^2/n}\right] d\mu \\
&\propto \exp\left(\frac{t^2}{2n}\sigma^2\right)\left(\frac{1}{n}\sigma^2\right)^{\frac{1}{2}}
\end{aligned}
$$

which establishes that the value of $E\{Y^t|\mathrm{y}\}$ always exists, albeit as a function of t, n and σ^2.

4.4.5 Second improper prior beliefs for the median

If we now represent our prior beliefs about the median as $\pi_{M_Y}(m) \propto 1$ then

$$\pi(\mu) \propto \exp(\mu) \qquad \text{and}$$

$$\begin{aligned}
\mathbb{E}\{Y^t|\mathrm{y}\} &\propto \int_{-\infty}^{\infty} \exp(t\mu)\exp(\mu)\exp\left[-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}\right] d\mu \\
&= \exp\left[\frac{2(t+1)\bar{x}+(t+1)^2\sigma^2/n}{2}\right] \\
&\times \int \exp\left[-\frac{\{(\bar{x}+\sigma^2/n+t\sigma^2/n)-\mu\}^2}{2\sigma^2/n}\right] d\mu \\
&\propto \exp\left[\frac{(t+1)^2}{2n}\sigma^2\right]\left(\frac{1}{n}\sigma^2\right)^{\frac{1}{2}}
\end{aligned}$$

which again establishes that the value of $E\{Y^t|\mathrm{y}\}$ always exists, although again as a function of t, n and σ^2.

4.4.6 Summary

The proper priors were chosen to illustrate a range of tail thicknesses and, along with the two particular choices of improper priors, all show results that we will demonstrate allow prior beliefs about the quantile ratio to be chosen that also ensure $\mathbb{E}\{Y^t|\mathrm{y}\}$ is finite.

4.5 Prior beliefs when the median and also the quantile ratio are unknown

4.5.1 Introduction

When we model uncertainty in both the median and quantile ratio then

$$j_t(\theta) = \exp(t\mu)\exp\left(\frac{t^2}{2}\sigma^2\right)$$

and hence we can say that

$$\begin{aligned}\mathbb{E}\{Y^t|\mathrm{y}\} &\propto \int_0^\infty\int_{-\infty}^\infty \exp(t\mu)\exp\left(\frac{t^2}{2}\sigma^2\right)\pi(\mu,\sigma^2) \\ &\times (2\pi\sigma^2)^{-\frac{n}{2}}\left(\prod y_i\right)^{-1}\exp\left[-\frac{\{S+n(\bar{x}-\mu)^2\}}{2\sigma^2}\right]d\mu d\sigma^2\end{aligned}$$

where $\pi(\mu,\sigma^2) = \exp(\mu)\pi_{M_Y}(exp(\mu)) \times \pi(\sigma^2)$.

If we let

$$I_\mu = \int_{-\infty}^\infty \exp(t\mu)\exp(\mu)\pi_{M_Y}(exp(\mu))\exp\left[-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}\right]d\mu$$

then, in Section 4.4, we have examined the convergence of I_μ for a range of prior beliefs when only the median is unknown.

We can then say that

$$\mathbb{E}\{Y^t|\mathrm{y}\} \propto \int_0^\infty I_\mu \exp\left(\frac{t^2}{2}\sigma^2\right)\pi(\sigma^2)(\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{S}{2}\frac{1}{\sigma^2}\right)d\sigma^2$$

If we now return to the case of logNormal prior beliefs for the median where, after a little re-arrangement of the results in Equation 4.1, we can say that

$$\begin{aligned}\mathbb{E}\{Y^t|\mathrm{y}\} &\propto \int_0^\infty \frac{(\sigma^2)^{\frac{1}{2}}}{(bn+\sigma^2)^{\frac{1}{2}}}\exp\left[\frac{2bnt\bar{x}-n(a-\bar{x})^2+t(2a+bt)\sigma^2}{2(bn+\sigma^2)}\right] \\ &\times \exp\left(\frac{t^2}{2}\sigma^2\right)\pi(\sigma^2)(\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{S}{2}\frac{1}{\sigma^2}\right)d\sigma^2\end{aligned}$$

and after noting that $t(2a+bt)\sigma^2 = t(2a+bt)(bn+\sigma^2) - bnt(2a+bt)$ we can say

$$\begin{aligned} \mathbb{E}\{Y^t|\mathrm{y}\} &\propto \int_0^\infty \frac{1}{(bn+\sigma^2)^{\frac{1}{2}}} \exp\left[\frac{2bnt\bar{x} - n(a-\bar{x})^2 - bnt(2a+bt)}{2(bn+\sigma^2)}\right] \\ &\times \exp\left(\frac{t^2}{2}\sigma^2\right) \pi(\sigma^2)(\sigma^2)^{-\frac{(n-1)}{2}} \exp\left(-\frac{S}{2}\frac{1}{\sigma^2}\right) d\sigma^2. \end{aligned}$$

If we now let $H = 2bnt\bar{x} - n(a-\bar{x})^2 - bnt(2a+bt)$ then we can observe over the stated range of integration that

$\frac{1}{(bn+\sigma^2)^{\frac{1}{2}}}$ is bounded above by $\frac{1}{(nb)^{\frac{1}{2}}}$ and that

$\exp\left[\frac{H}{2(bn+\sigma^2)}\right]$ is bounded above by $\exp\left(\frac{H}{2bn}\right)$ if $H > 0$ or 1 if $H \leq 0$.

Hence

$$\begin{aligned} \mathbb{E}\{Y^t|\mathrm{y}\} &\propto \int_0^\infty \frac{1}{(bn+\sigma^2)^{\frac{1}{2}}} \exp\left[\frac{2bnt\bar{x} - n(a-\bar{x})^2 - bnt(2a+bt)}{2(bn+\sigma^2)}\right] \\ &\times \exp\left(\frac{t^2}{2}\sigma^2\right) \pi(\sigma^2)(\sigma^2)^{-\frac{(n-1)}{2}} \exp\left(-\frac{S}{2}\frac{1}{\sigma^2}\right) d\sigma^2 \\ &\leq A\int_0^\infty \exp\left(\frac{t^2}{2}\sigma^2\right) \pi(\sigma^2)(\sigma^2)^{-\frac{(n-1)}{2}} \exp\left(-\frac{S}{2}\frac{1}{\sigma^2}\right) d\sigma^2 \end{aligned}$$

where the value of A depends on the sign of $2bnt\bar{x} - n(a-\bar{x})^2 - bnt(2a+bt)$.

Looking at the results in Sections 4.4.2, 4.4.3, 4.4.4 and 4.4.5 and after noting the result above, we will continue to examine the convergence of

$$I_{0,\infty} = \int_0^\infty \exp\left(\tfrac{T}{2}\sigma^2\right) \pi(\sigma^2)(\sigma^2)^{-\frac{N}{2}} \exp\left(-\tfrac{S}{2}\tfrac{1}{\sigma^2}\right) d\sigma^2$$

where $T = t^2$ or $t^2 + t^2/n$ or $t^2 + (t+1)^2/n$, $N = n$ or $n-1$ with $N, T, S > 0$ for any $n > 1$.

We will examine the convergence of this integral by initially concentrating on

$$\begin{aligned} I_{1,\infty} &= \int_1^\infty \exp\left(\frac{T}{2}\sigma^2\right) \pi(\sigma^2)(\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{S}{2}\frac{1}{\sigma^2}\right) d\sigma^2 \\ &\leq \int_1^\infty \exp\left(\frac{T}{2}\sigma^2\right) \pi(\sigma^2)(\sigma^2)^{-\frac{N}{2}} d\sigma^2 = I'_{1,\infty} \end{aligned}$$

when once the convergence of $I'_{1,\infty}$ has been established, the convergence of $I_{0,1}$, defined in the natural way, should follow from the dominating rate of convergence of $\exp\left(-\frac{S}{2}\frac{1}{\sigma^2}\right)$ in $(0,1)$ and hence the convergence of $I_{0,\infty}$ is obtained.

The convergence of $I_{0,1}$ will of course need to be verified in each case.

4.5.2 1+Gamma prior beliefs for the quantile ratio

Looking at the quantile ratio, where if $R_Y \sim$ 1+G(c,d) then

$$\pi(\sigma^2) \propto \frac{\exp(q\sigma)}{\sigma}[\exp(q\sigma)-1]^{(d-1)}\exp[-c\{\exp(q\sigma)-1\}] \qquad \text{and hence}$$

$$\begin{aligned} I'_{1,\infty} &\propto \int_1^\infty \exp\left(\frac{T}{2}\sigma^2\right)\exp(q\sigma)[\exp(q\sigma)-1]^{(d-1)} \\ &\times \exp[-c\{\exp(q\sigma)-1\}](\sigma^2)^{-\frac{(N+1)}{2}}d\sigma^2 \end{aligned}$$

and $I'_{1,\infty}$ is finite because of the dominating effect of the term

$$\exp[-c\{\exp(q\sigma)-1\}]$$

in the integrand as σ^2 becomes large, remembering that $q > 0$.

$I_{0,1}$ converges because

$$I_{0,1} \leq B\int_0^1 (\sigma^2)^{-\frac{(N+1)}{2}}\exp\left(-\frac{S}{2}\frac{1}{\sigma^2}\right)d\sigma^2$$

which is finite because of the dominant effect of $\exp\left(-\frac{S}{2}\frac{1}{\sigma^2}\right)$ in the integrand.

Hence the convergence of $I_{0,\infty}$ has been established.

4.5.3 logNormal prior beliefs for the quantile ratio

Looking at the quantile ratio, where if $R_Y \sim$ 1+logN(c,d) then

$$\pi(\sigma^2) \propto \frac{\exp(q\sigma)}{\sigma}\frac{1}{[\exp(q\sigma)-1]}\exp\left[-\frac{\{\log[\exp(q\sigma)-1]-c\}^2}{2d}\right] \qquad \text{hence}$$

$$
\begin{aligned}
I'_{1,\infty} \;\propto\; & \int_1^\infty \exp\left(\frac{T}{2}\sigma^2\right) \frac{\exp(q\sigma)}{[\exp(q\sigma)-1]} \\
\times\; & \exp\left[-\frac{\{\log[\exp(q\sigma)-1]-c\}^2}{2d}\right] (\sigma^2)^{-\frac{(N+1)}{2}} d\sigma^2
\end{aligned}
$$

and $I'_{1,\infty}$ will only converge if

$$\log[\exp(q\sigma)-1] > (dT\sigma^2)^{\frac{1}{2}}$$

where for large values of σ^2

$$\log[\exp(q\sigma)-1] \approx q\sigma$$

and we obtain convergence if $d < \frac{q^2}{T}$.

$I_{0,1}$ will converge because

$$I_{0,1} \leq C \int_0^1 \frac{\exp(q\sigma)}{[\exp(q\sigma)-1]} (\sigma^2)^{-\frac{(N+1)}{2}} \exp\left(-\frac{K}{2\sigma^2}\right) d\sigma^2$$

which is finite because of the dominant effect of $\exp\left(-\frac{K}{2\sigma^2}\right)$ in the integrand.

Hence the convergence of $I_{0,\infty}$ has been established for the quantile ratio when $d < \frac{q^2}{T}$ or when the shape parameter $d^{\frac{1}{2}} < \frac{q}{T^{\frac{1}{2}}}$.

The logNormal prior beliefs, when transformed into prior beliefs for σ^2, includes the term

$$\exp\left[-\frac{\{\log[\exp(q\sigma)-1]-c\}^2}{2d}\right]$$

which is of lower order for σ^2 than

$$\exp[-c\{\exp(q\sigma)-1\}]$$

the comparable term when Gamma prior beliefs are modeled.

Thus 1+Gamma prior beliefs for the quantile ratio do allow posterior predictive moments to exist unconditionally, whereas for logNormal prior beliefs we have only been able to establish the existence of posterior predictive moments when d is constrained below an upper bound, where we note that for small values of d $\mathrm{logN}(c,d) \longrightarrow \mathrm{N}(c,d)$ with thin tails.

4.6 Numerical integration

4.6.1 Introduction

We have shown in Sections 4.4.2 to 4.4.5 that the existence of posterior moments is available for a wide range of prior beliefs for the median.

To enable evaluation of posterior moments it will be extremely useful, from a computational point of view, to reduce the dimension of the problem from two to one and then compute its value by quadrature.

The general theory is developed in O'Hagan and Forster (2004) but we will concentrate here on a logNormal model for the data with logNormal prior beliefs for the median.

4.6.2 Reducing the dimension

We want to evaluate the posterior expectation of a function of our parameter $\boldsymbol{\theta}$ as

$$\mathbb{E}\{k(\boldsymbol{\theta})|\mathbf{y}\} = \int k(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}.$$

where we are particularly interested in the population mean $k(\boldsymbol{\theta}) = \exp(\mu+\sigma^2/2)$.

Although we can say, in general, that

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}$$

we are usually unable to analytically determine the denominator and this is certainly true in general for our model.

As we have $\boldsymbol{\theta} = (\mu, \sigma^2)$ our posterior expectation becomes

$$\mathbb{E}\{k(\mu, \sigma^2)|\mathbf{y}\} = \frac{\int\int k(\mu, \sigma^2)\pi(\mu, \sigma^2)f(\mathbf{y}|\mu, \sigma^2)d\mu d\sigma^2}{\int\int \pi(\mu, \sigma^2)f(\mathbf{y}|\mu, \sigma^2)d\mu d\sigma^2} \tag{4.2}$$

where both the numerator and denominator are double integrals.

If we can find a way to reduce these integrals to one variable only then they will become significantly easier to evaluate. It is most expedient to work with an inner integral with respect to μ initially as we will now demonstrate.

We have chosen to work with the median and quantile ratio which, when we can express our prior beliefs independently as $\pi_{M_Y}(m)$ and $\pi_{R_Y}(r)$, then allows us to state that

$$\pi(\mu, \sigma^2) = \pi(\mu)\pi(\sigma^2)$$

and if we are able to choose $\pi_{M_Y}(m)$ so that $p(\mu|\sigma^2, \mathbf{y})$ is a member of the same family of distributions as $\pi(\mu)$ where

$$\pi(\mu, \sigma^2)f(\mathbf{y}|\mu, \sigma^2) \propto p_1(\sigma^2|\mathbf{y})p(\mu|\sigma^2, \mathbf{y})$$

then this conditional conjugacy gives us the potential to integrate out μ and reduce the dimension of our problem as required.

If we choose a logNormal prior belief for the median then we will demonstrate in Section 4.6.3 that both $\pi(\mu)$ and $p(\mu|\sigma^2, \mathbf{y})$ are members of the Normal family. The inner integral in the denominator of (4.2) with respect to μ yields 1 when $p(\mu|\sigma^2, \mathbf{y})$ is expressed as a probability density function, complete with its constant of integration. The inner integral in the numerator may also be evaluated with respect to μ for our form of $k(\boldsymbol{\theta})$ for a Normal distribution as we will demonstrate in Section 4.6.3.

For an appropriate choice of $p_{M_Y}(m)$ we may then say that

$$k_1(\sigma^2) = \int k(\mu, \sigma^2)p(\mu|\sigma^2, \mathbf{y})d\mu$$

and hence that

$$\mathbb{E}\{k(\mu, \sigma^2)|\mathbf{y}\} = \frac{\int k_1(\sigma^2)p_1(\sigma^2|\mathbf{y})d\sigma^2}{\int p_1(\sigma^2|\mathbf{y})d\sigma^2}$$

which has reduced our posterior expectation to evaluating the ratio of two one dimensional integrals.

We are then able to evaluate each integral separately using the simplest of quadrature rules where we divide the common range of integration [0,U] into the same w equal parts and then calculating the value of the integrand at the midpoint of each of the w intervals and applying equal weights.

So, using the denominator as an example, if

$$\mathrm{I} = \int_0^U p_1(\sigma^2|\mathrm{y})d\sigma^2$$

we may compute the approximate value as

$$\hat{\mathrm{I}} \approx \frac{U}{w}\sum_{i=1}^{w} p_1\left[\frac{U}{2w}(2i-1)\right]$$

where the following two points need careful consideration.

The range of integration for our problem extends to ∞ and to be able to ensure convergence of $\hat{\mathrm{I}}$ we need to choose U so that the value of $p_1\left[\frac{U}{2w}(2i-1)\right]$ is so small as to be unimportant whenever $\frac{U}{2w}(2i-1) \geq v$ for some $v < U$.

The error inherent in the approximation $\hat{\mathrm{I}}$ will reduce as the value of w increases and w may be chosen as the minimum value to achieve the desired level of accuracy.

4.6.3 Application when the median and the quantile ratio are unknown

We know from Section 4.2.4 that our prior beliefs in the X space may be expressed in general as

$$\pi(\mu,\sigma^2) = \exp(\mu)\frac{q\exp(q\sigma)}{2\sigma}\pi_{M_Y,R_Y}(\exp(\mu),\exp(q\sigma)).$$

If we assume that M_Y and R_Y are independent then we proceed using the considerable technical advantage that $\pi(\mu,\sigma^2)$ factorises into a function of μ alone and a function of σ^2 alone.

Then using the results from Section 4.6.2 we invoke the concept of conditional conjugacy by only considering the case when $M_Y \sim \text{logN}(a, b)$. To consider a model for prior beliefs for R_Y that allows flexibility in its choice with posterior predictive moments that exist unconditionally, after noting that we want to consider the case $q > 0$, we will use $R_Y \sim 1+\text{G}(c, d)$.

We are now able to say that

$$\pi(\mu, \sigma^2) \propto \exp\left[-\frac{(a-\mu)^2}{2b}\right] (\sigma^2)^{-\frac{1}{2}} \exp(q\sigma)[\exp(q\sigma) - 1]^{(d-1)} \exp[-c\{\exp(q\sigma) - 1\}]$$

and so

$$\begin{aligned}
\pi(\mu, \sigma^2) f(\mathbf{y}|\mu, \sigma^2) &\propto \exp\left[-\frac{(a-\mu)^2}{2b}\right] \exp\left[-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}\right] (\sigma^2)^{-\frac{(n+1)}{2}} \\
&\times \exp\left(-\frac{S}{2}\frac{1}{\sigma^2}\right) \exp(q\sigma)[\exp(q\sigma) - 1]^{(d-1)} \exp[-c\{\exp(q\sigma) - 1\}] \\
&= \frac{1}{\sqrt{2\pi(b^{-1} + n\sigma^{-2})^{-1}}} \exp\left[-\frac{\left(\frac{b\bar{x}+a\sigma^2/n}{b+\sigma^2/n} - \mu\right)^2}{2(b^{-1} + n\sigma^{-2})^{-1}}\right] \\
&\times \sqrt{2\pi(b^{-1} + n\sigma^{-2})^{-1}} \exp\left[-\frac{(a-\bar{x})^2}{2(b + \sigma^2/n)}\right] (\sigma^2)^{-\frac{(n+1)}{2}} \\
&\times \exp\left(-\frac{S}{2}\frac{1}{\sigma^2}\right) \exp(q\sigma)[\exp(q\sigma) - 1]^{(d-1)} \exp[-c\{\exp(q\sigma) - 1\}]
\end{aligned}$$

where we can identify

$$p(\mu|\sigma^2, \mathbf{y}) = \frac{1}{\sqrt{2\pi(b^{-1} + n\sigma^{-2})^{-1}}} \exp\left[-\frac{\left(\frac{b\bar{x}+a\sigma^2/n}{b+\sigma^2/n} - \mu\right)^2}{2(b^{-1} + n\sigma^{-2})^{-1}}\right]$$

and

$$\begin{aligned}
p_1(\sigma^2|\mathbf{y}) &= \sqrt{2\pi}(b^{-1} + n\sigma^{-2})^{-\frac{1}{2}} \exp\left[-\frac{(a-\bar{x})^2}{2(b + \sigma^2/n)}\right] (\sigma^2)^{-\frac{(n+1)}{2}} \\
&\times \exp\left(-\frac{S}{2}\frac{1}{\sigma^2}\right) \exp(q\sigma)[\exp(q\sigma) - 1]^{(d-1)} \exp[-c\{\exp(q\sigma) - 1\}].
\end{aligned}$$

So following Section 4.6.2 when

$$k(\mu, \sigma^2) = \exp(\mu + \sigma^2/2)$$

we can say that

$$
\begin{aligned}
k_1(\sigma^2) &= \exp\left(\frac{1}{2}\sigma^2\right)\int \exp(\mu)p(\mu|\sigma^2,\mathbf{y})d\mu \\
&= \exp\left(\frac{1}{2}\sigma^2\right)\int \exp(\mu)\frac{1}{\sqrt{2\pi(b^{-1}+n\sigma^{-2})^{-1}}}\exp\left[-\frac{\left(\frac{b\bar{x}+a\sigma^2/n}{b+\sigma^2/n}-\mu\right)^2}{2(b^{-1}+n\sigma^{-2})^{-1}}\right]d\mu \\
&= \exp\left(\frac{1}{2}\sigma^2\right)\exp\left[\frac{\left\{2\left(\frac{b\bar{x}+a\sigma^2/n}{b+\sigma^2/n}\right)+(b^{-1}+n\sigma^{-2})^{-1}\right\}}{2}\right] \\
&\times \int \frac{1}{\sqrt{2\pi(b^{-1}+n\sigma^{-2})^{-1}}}\exp\left[-\frac{\left(\mu-\left\{\frac{b\bar{x}+a\sigma^2/n}{b+\sigma^2/n}+(b^{-1}+n\sigma^{-2})^{-1}\right\}\right)^2}{2(b^{-1}+n\sigma^{-2})^{-1}}\right]d\mu \\
&= \exp\left(\frac{1}{2}\sigma^2\right)\exp\left[\frac{1+2ab^{-1}+2n\bar{x}\sigma^{-2}}{2(b^{-1}+n\sigma^{-2})}\right]
\end{aligned}
$$

and we are now able to determine the posterior population mean as

$$\mathbb{E}\{\exp(\mu+\sigma^2/2)|\mathbf{y}\} = \frac{\int_0^\infty k_1(\sigma^2)p_1(\sigma^2|\mathbf{y})d\sigma^2}{\int_0^\infty p_1(\sigma^2|\mathbf{y})d\sigma^2}$$

where we will use the midpoint rule to evaluate the integral as

$$\mathbb{E}\{\exp(\mu+\sigma^2/2)|\mathbf{y}\} \approx \frac{\sum_{i=1}^{w} k_1\left[\frac{U}{2w}(2i-1)\right]p_1\left[\frac{U}{2w}(2i-1)\right]}{\sum_{i=1}^{w} p_1\left[\frac{U}{2w}(2i-1)\right]}.$$

4.6.4 Determining the range and number of intervals for our Numerical integration

In Section 4.6.2 we stated that we will evaluate the integral in the numerator and the denominator separately using the simplest of quadrature rules by dividing the common range of integration [0,U] into the same w equal parts and calculating the value of the integrand at the midpoint of each of the w intervals and applying equal weights.

The three considerations that need to be weighed up and balanced are

1. the range of integration for our problem extends to ∞ and to ensure convergence of $\hat{\text{I}}$ we need to choose U so that the value of $p_1\left[\frac{U}{2w}(2i-1)\right]$ is so small as to be unimportant whenever $\frac{U}{2w}(2i-1) \geq v$ for some $v < U$

2. the error inherent in the approximation $\hat{\text{I}}$ will reduce as the value of w increases and w may be chosen as the minimum value to achieve the desired level of accuracy

3. the computing time to undertake these evaluations.

The initial investigations were conducted using Excel which allows each of the w lines of calculation to be viewed and the values of $p_1\left[\frac{U}{2w}(2i-1)\right]$ to be observed. The simulations using large numbers of draws were undertaken in R and the values of U and w were chosen to minimise computing time without compromising the error in $\hat{\text{I}}$ whilst ensuring that the value of $p_1\left[\frac{U}{2w}(2i-1)\right]$ becomes so small as to be unimportant.

From the investigations made the value of U=40 was chosen because if U=30 was chosen the value of $p_1\left[\frac{U}{2w}(2i-1)\right]$ did not always become so small as to be unimportant within the w lines of calculation. If U=50 was chosen then this was not computationally efficient because the value of $p_1\left[\frac{U}{2w}(2i-1)\right]$ became zero after approximately 80 of the w lines of calculation. The final 20% of the calculations did not contribute to the evaluation of the integral as the values of the numerator and the denominator were 0 for each of these lines.

The Bayesian posterior expectation (Bpe) is mentioned first in Chapter 1 where Bpe is simply shorthand for E$\{\exp(\mu + \sigma^2/2)|\text{y}\}$ when $M_Y \sim$ logN(scale,shape) and also $R_Y \sim$ 1+G(scale,shape).

Although it will be in Chapter 6 that we will develop the theory behind the choice of the Quantile Ratio, R_Y, as the ratio of the Third Quartile to Second Quartile (or Median) we do need to recognise now, to be able to perform our Bpe calculations, that we will work with the value of q = 0.6745 throughout this thesis.

The data enters the calculations shown in Section 4.6.3 as $\bar{x}$ and S, which would be recognised in classical statistics as sufficient statistics. The table that follows shows, for a range of values of w, and $\bar{x}$ and S, the value of the Bpe that was obtained

				w						
Distn	CoV	$\bar{x}$	S	100	200	400	500	800	1000	10000
Pareto	0.25	6.8	0.2	997	946	922	917	910	907	906
Gamma	0.25	8.2	0.8	4041	3836	3741	3730	3782	3774	3761
Weibull	0.25	6.8	4.3	1000	1059	1036	1036	1036	1036	1036
logN	2.00	6.0	51.0	948	947	947	947	947	947	947
Gamma	2.00	1.5	660.0	117	117	117	117	117	117	117

Table 4.2: Bpe when n=20, U=40 ; $M_Y \sim$ logN(0,100) and $R_Y \sim$ 1+G(11.5,6)

where for the distribution and CoV indicated the value of $\bar{x}$ and S that were quoted represent values at the high or low end of the range for S, for that distribution and CoV, for an individual sample of size 20. Values of $\bar{x}$ and S for other values of the CoV for these distributions will typically fall within the envelope of values quoted.

It can also be observed that as the size of S increases the calculated Bpe is stable across the range of values of w.

The value of S only enters the calculation of the Bpe in the term $\exp(-\frac{S}{2}\frac{1}{\sigma^2})$ in $p_1(\sigma^2|\mathrm{y})$. When S is small then $\exp(-\frac{S}{2}\frac{1}{\sigma^2})$ is relatively large for all values of σ^2 and quickly approaches its limiting value of 1 (reaches 0.90 when σ^2=0.95 for S=0.2). The values of the numerator and denominator calculated for small values of σ^2 are very large and if their values were plotted against σ^2 then each would be J-shaped. The value of w influences the values of σ^2 chosen to evaluate the Bpe and the values of the numerator and denominator are very sensitive to the values of small σ^2.

For large values of S the term $\exp(-\frac{S}{2}\frac{1}{\sigma^2})$ is very small for small σ^2 and only increases slowly (reaches 0.70 when σ^2=925.3 for S=660). A plot of the values of the numerator and denominator against σ^2 would be unimodal with both the numerator and denominator only showing much smaller values but with smooth behaviour around the mode (where values of the numerator and denominator both contribute to the value of the Bpe). Hence the calculated Bpe is stable across the range of values of w.

From the investigations made the value chosen was w=500 to provide as small value for w as was compatible with providing representative results.

O'Hagan and Forster (2004) have stated that "for one-dimensional integrals, acceptable accuracy can generally be achieved with w=100 evaluations". Table 4.2 provides evidence that, at least in this case, w, the number of divisions of the range of U, should be chosen as larger than 100 to obtain representative results. The results in Table 4.2 were obtained by taking the sum of the first w terms for the numerator and the denominator. If however, each Bpe was evaluated by taking the sum of the first 100 terms then the same answers were obtained, reflecting the fast convergence of the terms.

In the exploration of numerical integration other integrals have been evaluated where convergence was not so fast and so the conservative approach of w=500 was adopted for both the number of divisions of the range of integration as well as the number of evaluations (terms in the sums).

The numerical integration techniques developed here will be used extensively in Chapter 5, Applying the Bayesian model to practical situations. To evaluate moments for the posterior predictive distribution they offer considerable reduction in computational time compared with using MCMC techniques implemented in WinBUGS with comparable accuracy.

Chapter 5

Applying the Bayesian model

5.1 Introduction

The paper by Briggs *et al* (2005) examined clinical trials cost data modelling where an estimate of the population mean value is of interest. Cost data is non-negative and typically positively skewed and the customary parametric data models, which they used, were the Gamma and logNormal distributions. The estimators of the population mean that they used were the sample mean (sm) and exp(lm + lv/2), respectively, where lm and lv were the log scale sample mean and variance. The estimator exp(lm + lv/2) is one of a number possible for the population mean of a logNormal random variable as will be expanded in Section 5.10.

Briggs *et al* (2005) compared these two estimators by calculating the root mean square error (RMSE) for a simulation experiment with 10,000 replications as the square root of the mean (estimate - population mean)^2. They simulated from two parametric distributions, with five different values for the coefficient of variation (CoV) and samples of five different sizes but a constant population mean of 1000.

We will consider here estimates of population mean costs using a Bayesian method as well as the two Briggs *et al* (2005) methods. Our Bayesian approach entails calculating the posterior mean at each replication under the assumptions about the prior beliefs being considered.

Their results, presented in their Table 1 and reproduced here as our Table 5.1 overleaf, show that both of their estimators performed worst with respect to sample size when a simulation sample size of 20 (their smallest value) is taken and it is this value that we will use here. To enable us to make direct comparisons with their results we will simulate from the same two parametric distributions using RMSE's to compare the three estimators.

		RMSE for sm estimator Simulation sample sizes					RMSE for exp(lm +lv/2) estimator Simulation sample sizes				
Distribution	CoV	20	50	200	500	2000	20	50	200	500	2000
Gamma	0.25	56	35	18	11	6	56	35	18	11	6
	0.50	112	71	35	22	11	114	73	38	25	16
	1.00	221	141	70	44	22	400	304	241	226	218
	1.50	333	214	105	67	34	1388	1097	925	896	878
	2.00	440	284	141	89	45	2663	1914	1510	1420	1378
logNormal	0.25	56	36	18	11	6	56	36	18	11	6
	0.50	112	71	35	22	11	112	71	35	22	11
	1.00	224	141	72	45	23	221	137	69	43	22
	1.50	336	214	109	67	34	328	197	99	61	31
	2.00	450	288	143	63	45	419	250	122	54	38

Table 5.1: Simulated root mean square error by parametric distribution and also estimator for different sample sizes and coefficients of variation, 10000 replications (from Briggs *et al* (2005))

In Section 5.2 we will introduce our concerns that arise because we are drawing samples from positively skewed distributions. Whenever we conduct simulations we will draw from a sequence of random numbers that is specified by the number of replications and, wherever possible, by the seed value. We will seek to keep the number of replications to a minimum consistent with drawing samples that adequately represent the distributions from which they have been drawn in the sense of the sample first and second moments compared to the population values.

Specifying the sequence of random numbers by fixing the seed value enables simulations to be repeated to check results and enables direct comparisons between distributions to be made because the sampling variations are comparable.

The Bayesian model will be introduced in Sections 5.3 where we will use our Bayesian posterior expectation (Bpe) as the comparable estimator to the sample mean (sm) and exp(lm + lv/2) of Briggs *et al* (2005). We will continue to use the RMSE loss function introduced in Briggs *et al* (2005) throughout Section 5, except for Section 5.10, to enable direct comparisons to be made with their results.

We have established in Chapter 3 that using customary weak prior beliefs for our logNormal data model does not lead to a posterior predictive mean that is finite. However, it is not always possible to elicit an expert's prior beliefs but we do need to establish some prior beliefs that can be used in those situations. These objective prior beliefs will provide a reference against which elicited prior beliefs can be compared.

A default prior is a means of having objective prior beliefs readily available that can be used in a wide range of circumstances.

The default prior for the quantile ratio will be determined by the methods described in Section 5.4 where it will be noted that, when CoV = 2, the average value of the Bpe is below the sample mean value because of the need to constrain the large values of the Bpe that may arise when optimising this default prior.

In Section 5.5 we will apply our Bayesian model to the three observed data sets of Briggs *et al* (2005). I would like to thank Prof Simon Dixon of ScHARR, the University of Sheffield, for making these data sets available to me. Some of the reasons why the Bpe performs best, in the sense of lowest RMSE, will be explored here and will be developed further in Section 5.6.

Briggs *et al* (2005) relied solely on simulation to produce their four sets of numerical results shown in their Table 1 although it is possible, as will be shown in Section 5.7, to obtain theoretical results for three of their sets of parametric simulations.

In Section 5.8 we will be able to obtain approximate theoretical results when applying the logNormal estimator $\exp(\bar{X} + S^2/2)$, or exp(lm + lv/2), to Gamma distributions.

Shrinkage estimators will be explored in Section 5.9 to determine whether, in a classical framework, lower RMSE can be achieved at the expense of bias.

In Section 5.10 we will compare, for logNormal distributions, the conditionally minimal MSE estimator of Zhou (1998) with the sample mean, exp(lm + lv/2) and Bpe estimators to determine the range of values of σ^2, the variance of the log-scale, where the Bpe achieves lower Relative MSE than Zhou's CMMSE.

5.2 Drawing samples

Distributions with support on the positive part of the real line are generally skewed to the right and all the distributions considered here are of this type.

A distribution is said to possess a heavy, or thick, right tail if

$$p\{Y > y\} \rightarrow 0 \text{ as } y \rightarrow \infty \text{ slower than an exponential function}$$

and the probability of exceeding some large value is greater the heavier the tail.

This can be illustrated by examining, for an increasing sequence of p values, the p-percentiles for the random variable Y, following a range of distributions that possess increasingly heavier right tails and whose mean = 1000 and CoV = 2, to obtain

	Distribution			
p values	Gamma	Weibull	logNormal	Pareto
90	3002	2675	2273	1566
99	9736	9594	8555	4643
99.9	17506	20252	22548	13770
99.99	25713	34411	50067	40838
99.999	34158	51912	100070	121116
99.9999	42751	72639	185981	359201
99.99999	51445	96501	327454	1065304
99.999999	60212	123421	552729	3159437
99.9999999	69034	153337	901730	9370130
99.99999999	77900	186192	1430110	27789552

Table 5.2: p-percentiles

where we will formally introduce the Weibull and Pareto distributions later.

So if we look at $p\{Y > 75000\}$ then following the table above we obtain

$$\begin{aligned}\text{Gamma} &= 0.000000001\\ \text{Weibull} &= 0.000000785\\ \text{logNormal} &= 0.000027004\\ \text{Pareto} &= 0.000027605\end{aligned}$$

where the Gamma distribution is light tailed and the Weibull (for this value of the CoV), logNormal and Pareto are increasingly heavy tailed.

The consequence of this observation is that when drawing samples from these right skewed distributions using random numbers to obtain the very large values possible for the heavier tailed distributions is indeed a rare event. According to the cycle of the congruential generator used to provide the random numbers a very large sample size may be necessary for a very large value to occur in the sample. The size of sample required increases as the distribution becomes heavier tailed.

The way that we have chosen to draw samples that adequately represent the first and second order moments of the population from which it has been drawn is by judicious choice of seed value and number of replications, although the minimum number of replications will be used for computational efficiency.

In Briggs *et al* (2005) each simulation experiment had a maximum of 2000 draws with 10000 replications. We would expect that sampling from positively skewed distributions will produce many replications that do not reflect the occasional extremely large values that may be possible and sample mean values, and more pertinently sample variances, will be lower than the corresponding values for the populations from which they were sampled.

We expect to obtain a value 552729 or larger from the logNormal distribution once in 100,000,000 sampled values. The largest number of sampled values that Briggs *et al* (2005) obtain for any single simulation experiment was 20,000,000 and so they would not have expected to see a sample value of this magnitude.

5.3 Bayesian model

To formulate our Bayesian model we will follow our work in Chapter 2 and use a logNormal data model and following Chapter 4 we will specify prior beliefs for the median and quantile ratio of the logNormal random variable. We know from the results in Chapter 4 that we have a wide choice of distributions available to represent our prior beliefs for the median but, as we want to undertake many replications, using numerical integration as introduced in Section 4.6 will give us considerable computational advantages over MCMC techniques such as WinBUGS.

As explained in Section 4.6 if we represent our prior beliefs for the median as a logNormal distribution then the Bayesian posterior expectation (Bpe) becomes a ratio of two one dimensional integrals which may then be easily evaluated using quadrature techniques. To complete our specification we will represent our prior beliefs for the quantile ratio as a 1+Gamma distribution, where the choice of the Gamma distribution allows a range of possible prior beliefs to be captured.

To conduct the Bayesian analysis it is helpful to be able to specify default priors that can be used no matter what the specification of the data model and without the need to conduct elicitations. Although we will use the simulations from the customary parametric data models to determine, or train, our choice of default prior, when observed cost data sets are analysed then we do not know the actual population distribution from which they have been sampled and we will look to determine a default prior that is robust against misspecification.

The RMSE used in Briggs *et al* (2005) is a particular form of a loss function as used in statistical decision theory. The mean square error (MSE) is known as the quadratic loss function whose properties may be summarised as giving equal weight to values that are equally above or below the target value whilst the quadratic nature of the calculations give larger weight to values that are further from the target value.

In the Bayesian model below our estimator is the Bpe and we then define the Total RMSE (TRMSE) as

$$\text{TRMSE } = \Sigma \text{ RMSE}$$

which gives equal weight to each of the four parametric models considered, namely logNormal and Gamma each for CoV $= 0.25$ and 2, when searching amongst candidate prior distributions to determine the default prior that will minimise the TRMSE.

The likelihood model that we are using is the logNormal distribution. So the range of possible values for the posterior expectation is non-negative and this range may be divided, in this case with known population mean, into 0 to population mean and then population mean to infinity. When we are searching to minimise the TRMSE the process is driven by not allowing values for the posterior expectation above the population mean to become too large - where the range is unbounded. Values below the population mean are not subject to such large changes as the choice of candidate default prior changes.

The median, M_Y, of the $\log\text{N}(\mu, \sigma^2)$ distribution is $\exp(\mu)$ and so the prior belief chosen for the median as $\log\text{N}(a, b)$ in general, with $\log\text{N}(0, 100)$ in this case, translates to a prior belief for the mean on the log scale of N(0,100) which is not atypical if an noninformative prior is to be considered. The value $b = 100$ was chosen to be sufficiently weak to allow a wide range of possible values for the mean to be considered but the value is within a range of values for b where the value chosen has very little impact on the calculated value of the posterior mean. Similar comments apply to the choice of $a = 0$ as will now be demonstrated in Table 5.3 overleaf where 10K replications have been taken with an 0710 seed value, $\mu_Y = 1000, n = 20$ and R $\sim$ 1+G(11.5,6).

		a , b					
Distribution	CoV	0 , 10^2	10 , 10^2	-10 , 10^2	0 , 5^2	0 , 20^2	50 , 5^2
Gamma	0.25	60	61	60	60	61	65
logN	0.25	62	63	62	62	62	67
Gamma	2.00	552	552	553	551	552	703
logN	2.00	313	312	313	313	312	326

Table 5.3: Simulated root mean square error for the Bpe for a range of logN(a,b) prior beliefs for the median

The first column shows the simulated RMSE values when $M_Y \sim \text{logN}(0, 100)$, which is a not atypical choice if we wish to model M_Y with fairly weak prior beliefs.

In the next four columns we look at variations around $M_Y \sim \text{logN}(0, 100)$ to determine if the values of the estimated RMSE are robust to other specifications that represent ranges of values that may be possible for the Median (from data that has been observed). We can observe that this is indeed the case.

In the last column we have looked at $M_Y \sim \text{logN}(50, 25)$ to see the effect of a stronger prior belief, which does not represent a range of values that observed data suggest would be possible for the Median. We can observe that, with the sole exception of the parametric data model Gamma CoV = 2, the values of the estimated RMSE's are also reasonably robust to other specifications.

Hence, we will use $M_Y \sim \text{logN}(0, 100)$ to represent prior beliefs for the Median in all the subsequent analysis, unless specified to the contrary.

5.4 Default prior for the Quantile Ratio

To determine a default prior the TRMSE was used to make comparisons between candidate priors for the two parametric distributions using a simulation sample size of 20. In Briggs *et al* (2005) for these two parametric distributions a population mean of 1000 was taken with a range of coefficient of variation (CoV) from 0.25 to 2.00.

It is worth noting that the smaller the CoV the less skewed both distributions become and they then approach the same approximate Normal distribution. The largest values of RMSE's, with respect to CoV, arise in Table 5.1 for CoV = 2 and this end of the chosen range for the CoV's clearly shows that the misapplication of the exp(lm + lv/2) estimator to Gamma distributed data produces poor results.

The value of the CoV is used to determine the numerical values drawn from the distribution at random where to undertake the numerical integration only the log scale sample mean and log scale sample sum of squares are required for each replication.

To calculate the Bpe using numerical integration we also need to specify the candidate prior parameter values and can then use the Bpe as the estimate when calculating the (estimate - population mean)^2 and hence the RMSE. It was in the worst case, in the sense of the largest values of the RMSE, which arose for both distributions when n = 20 and CoV = 2, that the choice of prior for the Quantile Ratio, R_Y, influenced the value of the TRMSE.

Briggs *et al* (2005) used 10,000 (10K) replications to generate their results. When the Bpe was calculated by numerical integration using the R programming package the value of the RMSE produced was influenced by the initial seed value that had been used for the random number generator, which was particularly noticeable for the Gamma distribution with CoV = 2, as indicated in Table 5.4 on the following page, with comparable variations for other seed values

	number of replications		
seed	10K	100K	1000K
3.21E08	565.32	554.34	553.43
0710	551.90	551.22	553.15

Table 5.4: Simulated root mean square error for Bpe with sample size 20 for a Gamma distribution with CoV = 2 when $R_Y \sim$ 1+G(11.5,6) and an increasing number of replications from the same seed value

To reduce sampling errors when determining the default prior the seed value of 0710 was used with 1000K replications.

When samples were drawn from the Gamma and logNormal distributions they more than adequately represented the population first and second moments, as shown in the table below

		population		sample values				number of
Distribution	CoV	mean	sd	mean	sd	CoV	seed	replications
Gamma	0.25	1000	250	1000	250	0.250	0710	1000K
logN	0.25	1000	250	1000	250	0.250	0710	1000K
Gamma	2.00	1000	2000	1001	2004	2.001	0710	1000K
logN	2.00	1000	2000	1000	2001	2.000	0710	1000K

Table 5.5: Parametric distributions, mean = 1000, sample size = 20

After comparing candidate priors we find that the TRMSE was minimised for the Bpe when $R_Y \sim$ 1+G(11.5,6) which give the following comparative results

		sm		exp(lm + lv/2)		Bpe	
Distribution	CoV	RMSE	average	RMSE	average	RMSE	average
Gamma	0.25	56	1000	56	1002	61	1019
logN	0.25	56	1000	56	1002	62	1017
Gamma	2.00	448	1001	4.89E+48	4.89E+45	553	668
logN	2.00	447	1000	459	1079	309	862
TRMSE	-	1007	-	4.89E+48	-	985	-

Table 5.6: Estimated root mean square error for sample size 20 by underlying distribution and estimator for different coefficients of variation

where "average" denotes the average of the estimated mean values.

We will now use $R_Y \sim 1+G(11.5,6)$ as the default prior in all the Bayesian analysis that follows.

We note that, for small values of CoV, the Bpe performs quite well whilst predicting average values that are above the sample mean. For larger values of CoV, when the degree of skewness is greater, the Bpe performs better than the sample mean for the logNormal distribution but worse for the Gamma distribution and in both cases predicts an average value that is below the sample mean.

This reflects the need, when optimising the default prior for the quantile ratio, to constrain the large values of the Bpe that arise as a result of the occasional large values in these distributions which would consequently have a large influence on the value of the RMSE. This optimisation will however be dominated by the calculations for CoV = 2 and in particular for the Gamma distribution as this RMSE is the largest for the Bpe.

We are now able to display a plot of the probability density function of the default prior for the Quantile Ratio to illustrate how we have determined a prior that, while it does not have an upper limit on possible values of the Quantile Ratio, constrains values of the Quantile Ratio to generally lie in the range 1 to 3.

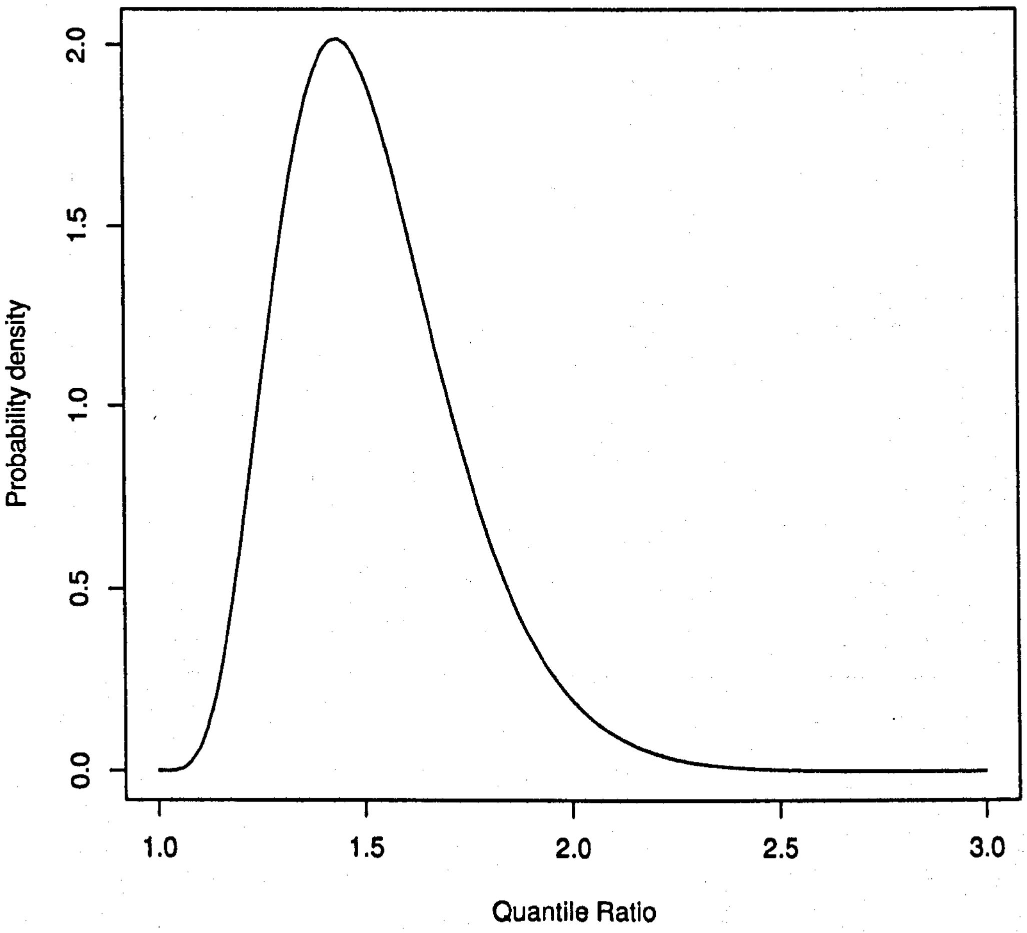


Figure 5.1: Plot to show the probability density function for the default prior for the Quantile Ratio

Values of the Quantile Ratio larger than 3 occur with very small probability.

We have now shown a plot of the probability density function of the variance of the log-scale, σ^2, to illustrate how the choice of prior for the Quantile Ratio transforms into the prior for σ^2.

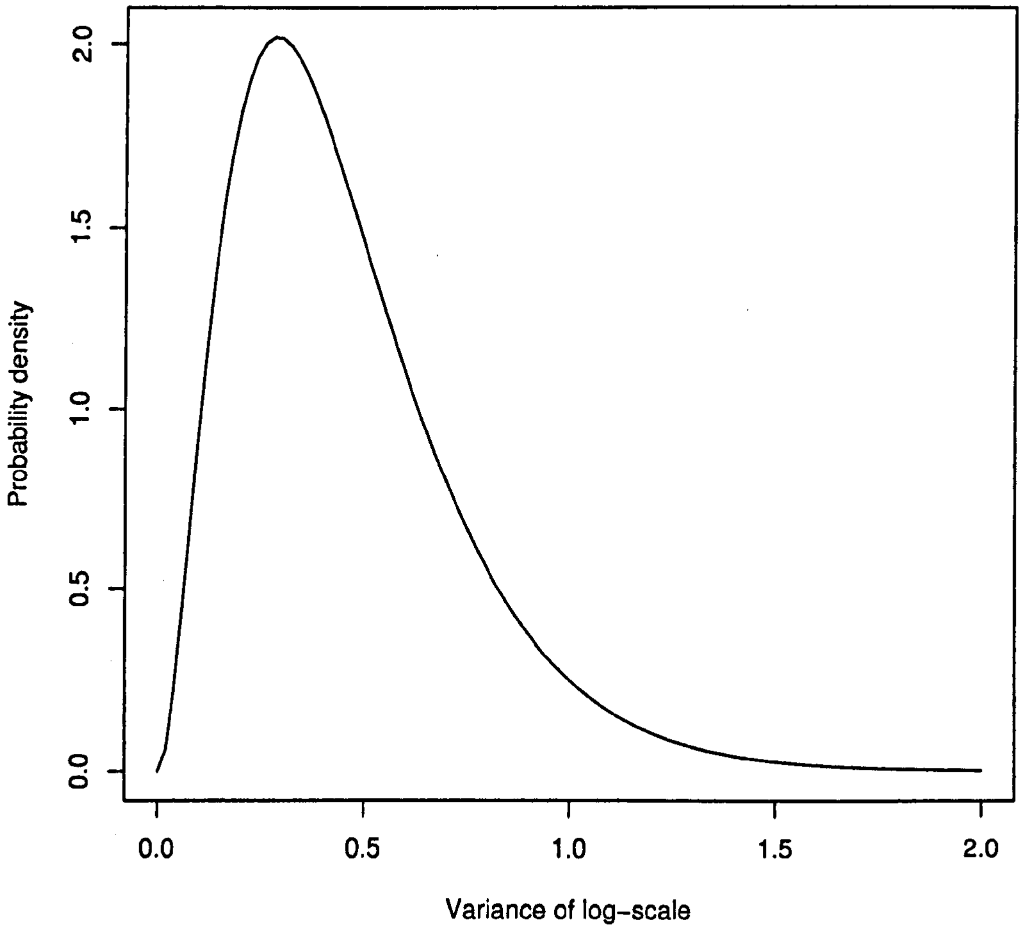


Figure 5.2: Plot to show the probability density function for σ^2, the variance of the log-scale

This prior for σ^2 does not have an upper limit on possible values of σ^2 but constrains the probable values to generally lie in the range 0 to 2 with values greater than 2 only occurring with a very small probability.

5.5 Application to the observed data sets

The three observed data sets analysed in Briggs *et al* (2005) are summarised in their Table 3. To help understand the nature of the data sets we produce the Five-number summaries as follows

data set	mean	sd	minimum	1st quartile	median	3rd quartile	maximum
CPOU	518	1145	12	102	172	399	10734
IV Fluids	2693	7083	123	318	421	621	73167
Paramedics	4233	7961	32	1151	2016	4069	130043

Table 5.7: Five-number summary for the three observed datat sets

All three observed data sets are skewed to the right with CPOU and Paramedics possessing distributional shapes that are broadly similar.

The IV Fluids data set is somewhat different with a longer thicker right tail. In Table 3 of Briggs *et al* (2005) we can observe that IV Fluids exhibits the smallest skewness and kurtosis but the largest CoV as the above indicates.

These properties may be displayed in the grouped data tables below

	population	percentage of values in the range						
data set	mean (pm)	> pm	> 2pm	> 3pm	> 4pm	> 5pm	> 10pm	> 20pm
CPOU	518	19.65	9.57	6.79	4.53	3.60	1.65	0.10
IV Fluids	2693	16.46	13.10	10.41	7.39	5.79	2.10	0.34
Paramedics	4233	23.97	11.07	6.26	3.40	2.70	0.86	0.16

Table 5.8: Grouped data table, values above the population mean

where we can observe that IV Fluids has a right tail that is much heavier than that for either CPOU or Paramedics, while considering this table with the next table

data set	population mean (pm)	percentage of values in the range						
		≤ pm/20	≤ pm/10	≤ pm/5	≤ pm/4	≤ pm/3	≤ pm/2	≤ pm
CPOU	518	2.77	5.33	25.41	36.63	50.10	61.52	80.35
IV Fluids	2693	0.08	22.61	68.60	77.33	80.52	81.61	83.54
Paramedics	4233	2.59	3.29	13.98	22.46	34.40	52.81	76.03

Table 5.9: Grouped data table, values below the population mean

we can observe that CPOU and Paramedics are mainly clustered between pm/10 and 2pm with IV Fluids concentrated between pm/20 and pm/4.

From Table 3 of Briggs *et al* (2005) we note that the CoV of each of the three observed data sets is around the value 2.

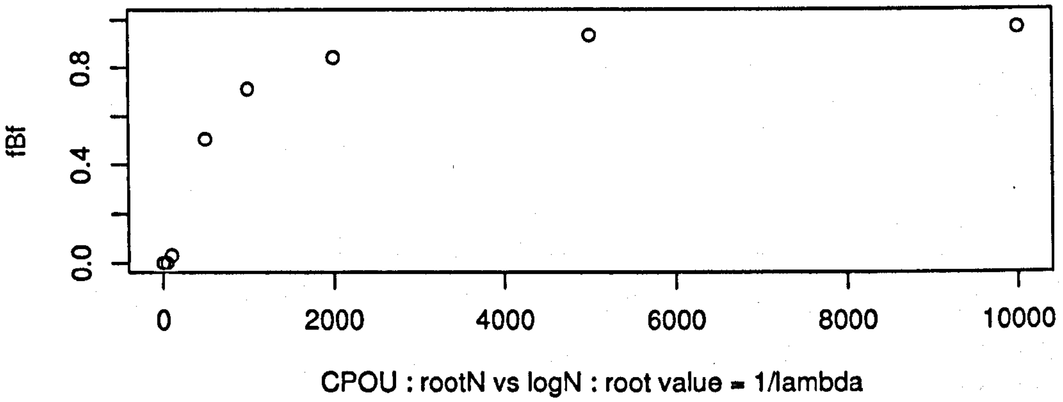
In Figure 2 of Briggs *et al* (2005) they show histograms of the log transformed observed cost data sets and the tail properties that we have described above become much more visually evident.

From Table 4 of Briggs *et al* (2005) we can observe that Paramedics is the only one of the three observed data sets whose log transformed histogram exhibits negative, albeit marginally, skewness with the highest coefficient of kurtosis and lowest coefficient of variation.

The work that we have undertaken in Chapter 2 developed, in Section 2.6, the theory to enable fBf's to be defined for the comparison of a range of rootNormal vs logNormal and logNormal vs rootGamma candidate models. The numerical results presented in Section 2.8 showed these comparisons for the pMDI+ data set.

We will now present numerical results for the comparisons of comparable ranges of candidate models for the CPOU, IV Fluids and Paramedics data sets.

The first plot in Figure 5.3 shows the fBf for comparing the rootNormal vs logNormal models for the CPOU data set where the value of the fBf → 1 as $\lambda \rightarrow 0$.



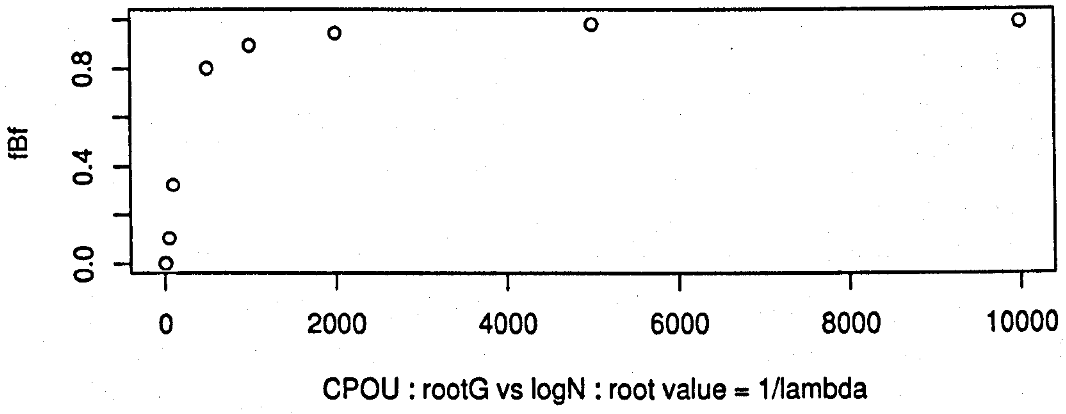


Figure 5.3: Plots comparing candidate models for the CPOU data set

The second plot in Figure 5.3 shows the fBf for comparing the rootGamma vs logNormal models for the CPOU data set where the value of the fBf → 1 as $\lambda \rightarrow 0$.

The first plot in Figure 5.4 shows the fBf for comparing the rootNormal vs logNormal models for the IV Fluids data set where the value of the fBf → 1 as $\lambda \rightarrow 0$.

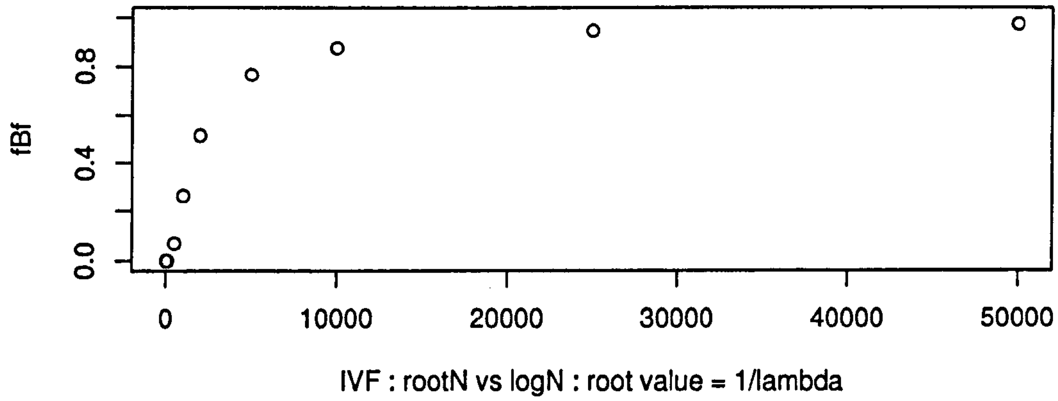


Figure 5.4: Plots comparing candidate models for the IV Fluids data set

The second plot in Figure 5.4 shows the fBf for comparing the rootGamma vs logNormal models for the IV Fluids data set where the value of the fBf → 1 as $\lambda \rightarrow 0$.

The first plot in Figure 5.5 shows the fBf for comparing the rootNormal vs logNormal models for the Paramedics data set where the value of the fBf → 1 as $\lambda \to 0$, although the value of the fBf is above 1 for values of $1/\lambda \geq 55$ with the peak fBf value of 1.26 occurring when $1/\lambda = 109$. Hence a logNormal model was favoured over a rootNormal model for values of $1/\lambda < 55$, whereas a rootNormal model was very weakly favoured over a logNormal model for values of $1/\lambda \geq 55$.

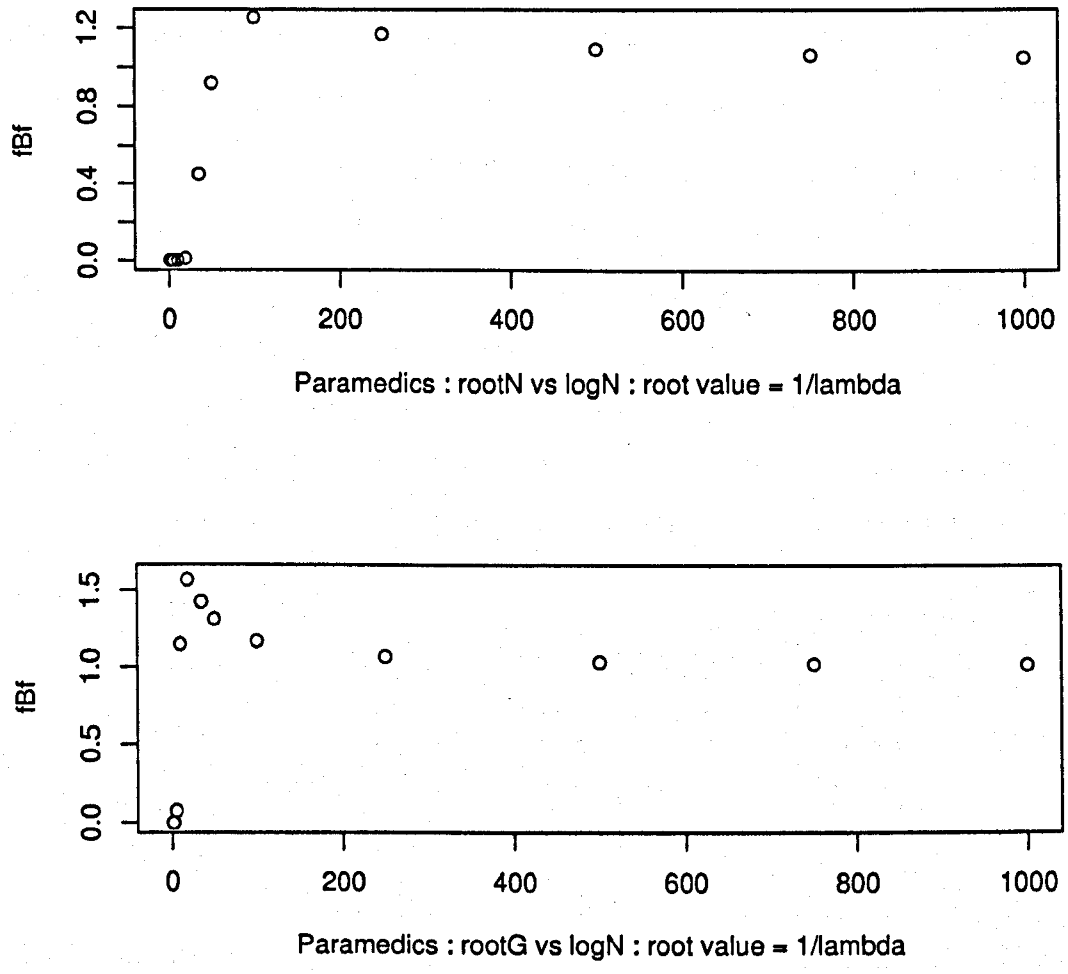


Figure 5.5: Plots comparing candidate models for the Paramedics data set

The second plot in Figure 5.5 shows the fBf for comparing the rootGamma vs logNormal models for the Paramedics data set where the value of the fBf → 1 as

$\lambda \longrightarrow 0$, although the value of the fBf is above 1 for values of $1/\lambda \geq 10$ with the peak fBf value of 1.57 occurring when $1/\lambda = 19$. Hence a logNormal model was favoured over a rootGamma model for values of $1/\lambda < 10$, whereas a rootGamma model was weakly favoured over a logNormal model for values of $1/\lambda \geq 10$.

As $\lambda \longrightarrow 0$ the fBf could barely distinguish between the logNormal and the other models.

We will use the logNormal as the data model for the CPOU, IV Fluids and Paramedics data sets because the only exception against it being the favoured model was for Paramedics when there was a region of very weak preference against.

We now wish to follow Briggs *et al* (2005) by applying the three estimators to the three observed data sets for sample size 20 with the Bpe to be evaluated using the default prior$R_Y \sim$ 1+G(11.5,6). The "sample" function in the R programming package, used here for replications "without replacement", does not allow control of the seed value and hence the samples drawn may show undue variation in their properties (representation of the population first and second order moments) for the different estimators due to sampling errors as indicated in the table following for the Paramedics data set

	population	10K		1000K	
	values	low	high	low	high
mean	4233	4204	4233	4234	4235
sd	7961	7832	8061	7961	7964
CoV	1.88	1.86	1.90	1.88	1.88

Table 5.10: Simulation results based on the total sample size for 10K and 100K replications of sample size 20, without replacement, from the Paramedics data set

where the values in the columns of the table show typical results to indicate the variation possible based on different replications ie different seed values.

When we follow Briggs *et al* (2005) by applying the three estimators to the three observed data sets with the Bpe evaluated when $R_Y \sim 1+G(11.5,6)$ for sample size 20, to reduce the sampling variation we will use 1000K replications and hence obtain

	population	sm		exp(lm + lv/2)		Bpe	
data set	mean	RMSE	average	RMSE	average	RMSE	average
CPOU	518	253	518	233	479	186	396
IV Fluids	2693	1570	2694	1648	2002	1459	1411
Paramedics	4233	1770	4232	1653	4270	1199	3754

Table 5.11: Simulation results based on drawing samples without replacement from the three large observed data sets

which is not inconsistent with their results for sample size 20 as shown in their Table 5, where they found that the RMSE was lowest for the Paramedics and CPOU data sets for the exp(lm + lv/2) estimator but lowest for the IV fluids data set for the sm estimator.

It is not unexpected that our Bpe achieves the lowest RMSE for all three observed data sets and is therefore, in the sense of the lowest RMSE, the best estimator as we know from Table 5.6 that our Bpe estimator performs best for logNormal data when CoV = 2. Inherent within the phrase "performs best" is the knowledge that our default prior used to determine the Bpe has been primarily "trained" by distributions with CoV = 2.

We remind ourselves that the Bpe estimator performs best for IV Fluids and we will explore the reasons for this in the next section.

5.6 Other data generating models

Of the three observed data sets that were examined in Briggs *et al* (2005) we noted in Section 5.5 that CPOU and Paramedics may be reasonably modelled as logNormal distributions while IV Fluids, although positively skewed, appears to possess a thicker tail than either a logNormal or Gamma model would suggest.

In Briggs *et al* (2005) they examined the performance of the exp(lm + lv/2) estimator when it was misapplied to Gamma distributed data. To determine how the same three estimators perform if different data generating models were used the following two models, Weibull and Pareto, were examined in the sense that data was simulated from the two models and then the same three estimators were used to generate RMSE's.

This an extension of the misapplication of the exp(lm + lv/2) estimator to Gamma distributed data introduced in Briggs *et al* (2005) because we now have four parametric data generating models (Gamma, Weibull, logNormal and Pareto) of increasingly heavier tails (the Weibull only conditionally as will be explained) and then the same three estimators were used to generate RMSE's.

5.6.1 Weibull model

If Y follows a Weibull distribution then this is defined by the two parameters $k > 0$ for shape and $\lambda > 0$ for scale. We also note that whenever $k < 1$ then this Weibull distribution possesses a heavy tail.

We also know that for a Weibull distribution its mean, μ_Y, and variance, σ_Y^2, are related to k and λ as

$$\mu_Y = \lambda\Gamma\left(1+\frac{1}{k}\right) = \frac{\lambda}{k}\Gamma\left(\frac{1}{k}\right) \qquad \text{and}$$

$$\sigma_Y^2 = \lambda^2\Gamma\left(1+\frac{2}{k}\right) - \mu_Y^2.$$

Remembering that $\sigma_Y^2 = \mathrm{CoV}^2\mu_Y^2$ and that $\mu_Y = 1000$, in this case, we can then show that

$$(1+\mathrm{CoV}^2)\left[\Gamma\left(\frac{1}{k}\right)\right]^2 = 2k\Gamma\left(\frac{2}{k}\right)$$

which we can solve for k by numerical methods for any value of CoV and hence obtain

$$\lambda = \frac{1000k}{\Gamma\left(\frac{1}{k}\right)}.$$

5.6.2 Pareto model

If Y follows a Pareto distribution then this is defined by the two parameters $k > 0$ for shape and $a > 0$ for location. We also note that the Pareto distribution always possesses a heavy tail.

We also know that for a Pareto distribution its mean, μ_Y, and variance, σ_Y^2, are related to k and a as

$$\mu_Y = \frac{ka}{k-1}, \text{ for } k > 1 \qquad \text{and}$$

$$\sigma_Y^2 = \frac{a^2k}{(k-1)^2(k-2)} = \frac{\mu_Y^2}{k(k-2)}, \text{ for } k > 2.$$

So again remembering that σ_Y^2= $\mathrm{CoV}^2\mu_Y^2$ and that $\mu_Y = 1000$, in this case, we can show that

$$k^2 - 2k - \frac{1}{\mathrm{CoV}^2} = 0$$

which we can solve for k by numerical methods for any value of CoV and hence obtain

$$a = \frac{k-1}{k}1000.$$

5.6.3 Simulation results

For both the Weibull and Pareto distributions we continue with a population mean of 1000.

For the Weibull distribution λ = 1095.21 and k = 4.54221 when CoV = 0.25, whereas λ = 575.250 and k = 0.54269 when CoV = 2 and we note that because $k < 1$ this distribution possesses a heavy tail.

For the Pareto distribution a = 804.806 and k = 5.12311 when CoV = 0.25, while a = 527.864 and k =2.11803 when CoV = 2.

To be able to draw samples from the Weibull and the Pareto distributions that reasonably reflected the population first and second moments required some consideration of the seed value and number of replications to be used. While the choice of seed value 0710 with 10K replications was satisfactory for the Weibull distribution it was necessary to search for seed value 3843495 combined with 1000K replications to obtain results that were acceptable for the Pareto distribution as the following table shows

		population		sample values				number of
Distribution	CoV	mean	sd	mean	sd	CoV	seed	replications
Weibull	0.25	1000	250	1001	250	0.249	0710	10K
Pareto	0.25	1000	250	1000	250	0.250	3843495	1000K
Weibull	2.00	1000	2000	1003	1994	1.988	0710	10K
Pareto	2.00	1000	2000	1001	2027	2.025	3843495	1000K

Table 5.12: Parametric distributions, mean = 1000, sample size = 20

The result of taking the seed value and number of replications indicated in Table 5.12 for sample size n = 20, evaluating the Bpe when $R_Y \sim 1+G(11.5,6)$, produces

Distribution	CoV	sm		exp(lm + lv/2)		Bpe	
		RMSE	average	RMSE	average	RMSE	average
Weibull	0.25	56	1001	56	1006	60	1025
Pareto	0.25	56	1000	54	999	58	1009
Weibull	2.00	447	1003	8519221	122444	408	782
Pareto	2.00	453	1001	177	960	170	984
TRMSE	-	1012	-	8519508	-	626	-

Table 5.13: Estimated root mean square error by underlying distribution and estimator for different coefficients of variation

The results obtained in the Table above are not unexpected from the results and discussion in Section 5.4.

However, the difficulty in drawing representative samples from right skewed distributions has been vividly illustrated by the potential problems in arriving at Table 5.12.

It is, of course, still entirely feasible and logical to make comparisons between the three estimators as the tail of the distribution becomes increasingly heavier for a constant value of the CoV.

We will note at this point that there is some discrepancy between the results obtained in Table 5.6 and those in Table 1 of Briggs *et al* (2005) when we apply the exp(lm + lv/2) estimator to Gamma, or lighter tailed, distributions. We will return to this topic in Section 5.7.

If we combine the results for sample size n = 20 from Table 5.6 and also Table 5.13, evaluating the Bpe when $R_Y \sim$ 1+G(11.5,6), we obtain

		sm		exp(lm + lv/2)		Bpe	
Distribution	CoV	RMSE	average	RMSE	average	RMSE	average
Gamma	0.25	56	1000	56	1002	61	1019
Weibull	0.25	56	1001	56	1006	60	1025
logNormal	0.25	56	1000	56	1002	62	1017
Pareto	0.25	56	1000	54	999	58	1009
Gamma	2.00	448	1001	4.89E+48	4.89E+45	553	668
Weibull	2.00	447	1003	8519221	122444	408	782
logNormal	2.00	447	1000	459	1079	309	862
Pareto	2.00	453	1001	177	960	170	984
TRMSE	-	2019	-	4.89E+48	-	1681	-

Table 5.14: Estimated root mean square error for sample size 20 by underlying distribution and estimator for different coefficients of variation

When CoV = 0.25 then all three estimators produce comparable results.

It is, however, when CoV = 2 that we can observe that as the tail of the parametric distribution from which values have been sampled becomes heavier then the Bpe performs better when compared with the sm in the sense of RMSE.

So the Bpe performs best when the data generating distribution has been correctly specified as the logNormal or indeed if it has been misapplied and, in particular, if the actual distribution possesses a heavier tail.

Although we have not explored this option there will exist distributions, for example mixtures, whose thickness of tail is intermediate between the Weibull and logNormal for which the Bpe performs better than the sm when CoV = 2.

5.7 Theoretical considerations

While Briggs *et al* (2005) obtained their results for estimates of the population mean from simulations, it is also possible to obtain theoretical results for their parametric simulations.

If the data $y_1, y_2, \ldots, y_n$ are the n observed values of the independent and identically distributed (iid) random variables $Y_1, Y_2, \ldots, Y_n$, then if we express Y as $\log Y = X$ we can define

$$\bar{Y} = \frac{1}{n}\Sigma Y_i \quad , \quad \bar{X} = \frac{1}{n}\Sigma X_i \quad \text{and} \quad S^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$$

with corresponding estimates from the log transformed data $x_i = \log y_i$

$$\text{lm} = \frac{1}{n}\Sigma x_i \quad \text{and} \quad \text{lv} = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$

Furthermore, we need to note that they considered only the case when the populations from which they drew their simulated values had a population mean of $\mu_Y = 1000$ with the population variance σ_Y^2 related to the mean through the relationship

$$\text{Coefficient of Variation} \equiv \text{CoV} = \frac{\sigma_Y}{\mu_Y} \quad \text{or} \quad \sigma_Y^2 = (\text{CoV})^2\mu_Y^2.$$

For the sample mean estimator $\bar{Y}$ as $\mathbb{E}\{\bar{Y}\} = \mu_Y = 1000$ we know that $\bar{Y}$ is an unbiased estimator and hence

$$\text{MSE } \{\bar{Y}\} = \text{ var } \{\bar{Y}\} = \frac{\sigma_Y^2}{n} = \frac{\mu_Y^2}{n} \times (\text{CoV})^2 \quad \text{and} \quad \text{RMSE } \{\bar{Y}\} = \mu_Y \times \frac{\text{CoV}}{\sqrt{n}}$$

where these results are applicable for any distribution with finite second moment, in particular the Gamma or the logNormal distributions.

We can present these results in the table below in a comparable format to Table 1 in Briggs *et al* (2005) and our Table 5.1, to observe that, with perhaps one exception, there is close agreement between the sampled values shown in Table 1 of Briggs *et al* (2005) and our theoretical values

	Sample sizes				
CoV	20	50	200	500	2000
0.25	56	35	18	11	6
0.50	112	71	35	22	11
1.00	224	141	71	45	22
1.50	335	212	106	67	34
2.00	447	283	141	89	45

Table 5.15: Theoretical RMSE for sm estimator

Briggs *et al* (2005) refer to the paper by O'Hagan and Stevens (2002) and their use of the logNormal estimator $\widehat{\mu}_Y = \exp(\bar{X} + S^2/2)$ applied to a logNormal distribution, where we have replaced the O'Hagan and Stevens use of Y by X.

This properties of this estimator were derived when $Y \sim \text{logN}(\mu, \sigma^2)$ with the mean value of Y defined as $\mu_Y = \exp(\mu + \sigma^2/2)$ with variance $\sigma_Y^2 = \mu_Y^2[\exp(\sigma^2) - 1]$. If we express Y as $\log Y = X \sim \text{N}(\mu, \sigma^2)$ then following O'Hagan and Stevens (2002) we can obtain

$$\mathbb{E}\{\widehat{\mu}_Y\} = \exp(\mu + \sigma^2/2n)\left(1 - \frac{\sigma^2}{n-1}\right)^{-\frac{n-1}{2}}$$

and so $\widehat{\mu}_Y$ is a (positively) biased estimator, whenever $\sigma^2 < n - 1$, contrary to the assertion made in Briggs *et al* (2005), with

$$\begin{aligned} \text{MSE}\{\widehat{\mu}_Y\} &= \text{var}\{\ \exp(\bar{X} + S^2/2)\ \} + (\text{bias})^2 \\ &= [\mathbb{E}\{\widehat{\mu}_Y\}]^2 \left[\exp(\sigma^2/n)\left(1 - \frac{2\sigma^2}{n-1}\right)^{-\frac{n-1}{2}}\left(1 - \frac{\sigma^2}{n-1}\right)^{n-1} - 1\right] \\ &+ \ (\mathbb{E}\{\widehat{\mu}_Y\} - \mu_Y)^2. \end{aligned}$$

In Briggs *et al* (2005) we know that $1000 = \mu_Y = \exp(\mu + \sigma^2/2)$ and also that $\text{CoV}^2\mu_Y^2 = \sigma_Y^2 = \mu_Y^2[\exp(\sigma^2) - 1]$ and so for their range of values of CoV and n we can compute the theoretical values of RMSE $\{\widehat{\mu}_Y\}$ using the relationships $\sigma^2 = \log(\text{CoV}^2 + 1)$ and $\mu = \log 1000 - \sigma^2/2$ and we can present these results in the table below in a comparable format to Table 1 in Briggs *et al* (2005) as

	Sample sizes				
CoV	20	50	200	500	2000
0.25	56	35	18	11	6
0.50	113	71	35	22	11
1.00	229	140	69	43	22
1.50	345	203	98	61	31
2.00	460	259	123	77	38

Table 5.16: Theoretical RMSE for $\exp(\bar{X} + S^2/2)$ estimator, evaluated for the logNormal distribution

and observe that, with perhaps one exception, there is close agreement between the sampled values shown in Table 1 of Briggs *et al* (2005) and our theoretical values when the sample size is large and/or CoV is small.

If we consider Table 5.15 as derived from logNormal data then we are able to make direct comparisons by individual cells with Table 5.16 for the theoretical values we have calculated and we can see that for CoV = 0.25 or 0.50 the values of the RMSE for both estimators are approximately equal for all sample sizes.

However, for larger values of CoV the $\exp(\bar{X} + S^2/2)$ RMSE is smaller than the sm RMSE for sample sizes of 50 and above but is greater for sample size 20.

If we take a fixed value of the CoV, say 1.50, and then examine the values of the RMSE that arise for the sm and the $\exp(\bar{X} + S^2/2)$ estimators as the sample size n varies then we can present the results in the table below. We can see that there is a value of n, in this case 26, which identifies the point where for values of n below 26 then the sm RMSE is less than the $\exp(\bar{X}+S^2/2)$ RMSE and comparable values of n exist for other values of the CoV, although the size of n decreases as the value of CoV increases.

Sample	Estimators	
size(n)	sm	$\exp(\bar{X}+S^2/2)$
50	212	203
40	237	229
30	274	270
26	294	294
25	300	301
20	335	345
15	387	416

Table 5.17: Theoretical RMSE for estimators, evaluated when CoV = 1.50 for the logNormal distribution

Whilst the sm RMSE, RMSE $\{\bar{Y}\} = \mu_Y \times \frac{\text{CoV}}{\sqrt{n}}$, is a simple function of n and CoV, the $\widehat{\mu}_Y = \exp(\bar{X}+S^2/2)$ RMSE is a more complex function of n and σ^2 because

$$\begin{aligned} \text{MSE}\{\widehat{\mu}_Y\} &= \exp(2\mu+\sigma^2/n)\left(1-\frac{\sigma^2}{n-1}\right)^{-(n-1)} \\ &\times \left[\exp(\sigma^2/n)\left(1-\frac{2\sigma^2}{n-1}\right)^{-\frac{n-1}{2}}\left(1-\frac{\sigma^2}{n-1}\right)^{n-1}-1\right] \\ &+ \left[\exp(\mu+\sigma^2/2n)\left(1-\frac{\sigma^2}{n-1}\right)^{-\frac{n-1}{2}}-\exp(\mu+\sigma^2/2)\right]^2 \qquad (5.1) \end{aligned}$$

However, both RMSE's possess the expected property that as n increases and/or CoV decreases then the value of the RMSE's decreases.

Briggs *et al* (2005) asserted that when samples have been truly drawn from a logNormal distribution then the lognormal estimator is more precise as observed in their Table 1. However, because their results are based on (a limited number of) simulations, their sampling errors have masked the more subtle true results that can be obtained theoretically.

5.8 Applying the $\exp(\bar{X} + S^2/2)$ estimator to Gamma distributions

For sample size n if we represent the data as $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ then they are observations on the corresponding iid random variables $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$.

If $Y \sim \mathrm{G}(a, b)$ then

$$f(y) = \frac{a^b y^{b-1} \exp(-ay)}{\Gamma(b)} \quad \text{where} \quad y > 0 \quad \text{and} \quad a, b > 0.$$

If we make the definition $\log Y = X$ then to apply the $\exp(\bar{X} + S^2/2)$ estimator we need to determine the distribution of the log transformation of Y which is

$$f(x) = \frac{a^b \exp(bx) \exp[-a \exp(x)]}{\Gamma(b)} \quad \text{where} \quad -\infty < x < \infty \quad \text{and} \quad a, b > 0$$

and then evaluate the mean and variance of this log transformation.

If we now examine the moment generating function of X then we are able to obtain the closed form

$$\begin{aligned} M(s) &= \mathbb{E}\{e^{sX}\} \\ &= \int \frac{\exp(sx) a^b \exp(bx) \exp[-a \exp(x)]}{\Gamma(b)} dx \\ &= \frac{\Gamma(s+b)}{\Gamma(b) a^s}. \end{aligned}$$

We can remind ourselves that $a, b > 0$ and we wish to evaluate the case $s = 0$, hence we can restrict our attention to the range $s + b > 0$.

To obtain moments for X we will need to differentiate the Gamma function and to undertake this we can appeal to the Weierstrass identity

$$\Gamma(v) = e^{-\gamma v}\frac{1}{v}\prod_{n=1}^{\infty}\frac{e^{\frac{v}{n}}}{1+\frac{v}{n}} \quad \text{where} \quad \gamma = 0.577215665 \quad \text{is known as Euler' s constant}$$

which is valid for all complex numbers, except negative integers.

As we will confine our attention to $v > 0$ then $\Gamma(v) > 0$ and we will find it analytically more convenient to work with $\log\Gamma(v)$ as

$$\log\Gamma(v) = -\gamma v - \log v + \sum_{n=1}^{\infty}\left[\frac{v}{n} - \log\left(1+\frac{v}{n}\right)\right].$$

As we can differentiate the Gamma function infinitely often then we can obtain

$$\frac{d}{dv}\log\Gamma(v) = -\gamma - \frac{1}{v} + \sum_{n=1}^{\infty}\frac{v}{n(v+n)} \quad \text{and}$$

$$\frac{d^2}{dv^2}\log\Gamma(v) = \frac{1}{v^2} + \sum_{n=1}^{\infty}\frac{1}{(v+n)^2}.$$

We will work with the cumulant generating function of X which we define as $K(s) = \log M(s)$ and so can obtain for X its mean as $K'(0)$ with variance $K''(0)$.

Hence we have

$$K(s) = \log\Gamma(s+b) - \log\Gamma(b) - s\log a \quad \text{with}$$

$$K'(s) = -\gamma - \frac{1}{s+b} + \sum_{n=1}^{\infty}\frac{s+b}{n(s+b+n)} - \log a \quad \text{where}$$

$$K'(0) = -\gamma - \frac{1}{b} + \sum_{n=1}^{\infty}\frac{b}{n(b+n)} - \log a$$

and similarly

$$K''(s) = \frac{1}{(s+b)^2} + \sum_{n=1}^{\infty}\frac{1}{(s+b+n)^2} \quad \text{where}$$

$$K''(0) = \frac{1}{b^2} + \sum_{n=1}^{\infty}\frac{1}{(b+n)^2}.$$

It is then possible to use these relationships to evaluate the mean and variance for the log of Gamma distributions as shown below

	Distributions			
	Gamma		log of Gamma	
CoV	a=scale	b=shape	mean	variance
0.25	0.01600	16.00	6.8754	0.0644
0.50	0.00400	4.00	6.7774	0.2838
1.00	0.00100	1.00	6.3305	1.6449
1.50	0.00044	0.44	5.4519	6.0456
2.00	0.00025	0.25	4.0666	17.1973
2.50	0.00016	0.16	2.1494	40.3917

Table 5.18: Mean and variance for log of Gamma distributions

where the series for the cumulants converge very slowly and 20,000 terms have been used.

We want to be able to determine theoretical RMSE's when the $\exp(\bar{X} + S^2/2)$ estimator is applied to Gamma distributions and to proceed we have now made the assumption that the log of Gamma distributions will follow approximately Normal distributions with $\log Y = X \overset{\text{approx}}{\sim} \mathrm{N}(\mu, \sigma^2)$.

The Gamma distribution is defined on the positive part of the real line and is right skewed. The log of Gamma distribution is defined on the real line and is left skewed. The assumption of approximate Normality does become less realistic as the CoV increases.

So, continuing with the assumption of approximate Normality for the log of Gamma distributions it is then possible to obtain approximate theoretical results for the RMSE of the estimator $\widehat{\mu}_Y = \exp(\bar{X} + S^2/2)$, in a comparable way to those obtained in earlier in Section 5.7 with a population mean of 1000, as

CoV	RMSE	$\mathbb{E}\{\widehat{\mu}_Y\}$
0.25	58	1002
0.50	133	1020
1.00	705	1383
1.50	38286	10317
2.00	∞	4.67E+11
2.50	∞	∞

Table 5.19: Approximate RMSE and $\mathbb{E}\{\widehat{\mu}_Y\}$ for $\exp(\bar{X} + S^2/2)$ and n = 20

These results have been obtained because when we deduced the moments of $\exp(\bar{X} + S^2/2)$, following O'Hagan and Stevens (2002), we needed to evaluate $\mathbb{E}\{e^{\frac{S^2}{2}}\}$ for the mean and the resulting Gamma integral only converged when $\sigma^2 < n - 1$. Similarly to deduce the variance we needed to evaluate $\mathbb{E}\{e^{S^2}\}$ and the resulting Gamma integral only converged when $\sigma^2 < (n-1)/2$.

We can then present results in the table below in a format comparable to that in Table 1 in Briggs *et al* (2005)

	Sample sizes				
CoV	20	50	200	500	2000
0.25	58	37	18	12	6
0.50	133	83	43	28	17
1.00	705	461	329	299	283
1.50	38285	7761	4511	4065	3858
2.00	∞	2.27E+08	799762	446177	343139
2.50	∞	∞	1.88E+12	3.44E+10	7.86E+09

Table 5.20: Approximate theoretical RMSE for the $\exp(\bar{X} + S^2/2)$ estimator, evaluated for the Gamma distribution

and observe that although there is reasonable agreement between these results and those presented in Table 1 of Briggs *et al* (2005) when CoV is 0.25 and 0.5 the two sets of results quickly diverge thereafter and we can see that the approximate theoretical results are never less than those derived by simulation and presented in Briggs *et al* (2005). When results are derived by simulation it will not, of course, be possible to replicate theoretical values of infinity and very large values should be seen instead.

The General Power Transformation (GPT) was introduced by Box and Cox (1964) as

$$X = \begin{cases} \lambda^{-1}(Y^{\lambda} - 1) & : \quad \lambda \neq 0 \\ \log Y & : \quad \lambda = 0 \end{cases}$$

for any strictly positive Y where we have already used these ideas in Chapter 2.

For CoV $= 2$ if we examine the results of a GPT on $Y \sim \mathrm{G}(0.00025, 0.25)$ for a sequence of $\lambda = 1/2, 1/4, 1/8, 1/16, \ldots$ then we can observe that $\lambda = 1/8$ produces an approximate Normal transformation, although with range $(-8, \infty)$.

If we evaluate the first two moments for this 1/8 GPT of the random variable $Y \sim \mathrm{G}(0.00025, 0.25)$ we find that, considering only the positive part of the real line, we obtain mean $= 6.75$ and variance $= 31.26$. If we had considered the whole of the real line then it is easy to show that, although the mean would have been smaller, the variance would have been larger than 31.26.

If we look at the calculations in Table 5.18 for the CoV $= 2$ row then we can see that for the log of Gamma distribution we obtained a variance of 17.2 and so the variance obtained when a better approximation to Normality has been obtained is even higher.

We believe that this fully justifies the results in Tables 5.19 and 5.20 and that they give conservative estimates of the RMSE for Gamma distributions with the $\exp(\bar{X} + S^2/2)$ estimator.

5.9 Shrinkage estimators

For sample size n if we represent the data as $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ then they are observations on the corresponding iid random variables $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$.

If $\widehat{\theta}(\mathbf{Y})$ is an estimator of an unknown parameter θ then in classical statistics the usual properties of interest for $\widehat{\theta}(\mathbf{Y})$ are bias and MSE. It may be possible to choose a shrinkage estimator $c\,\widehat{\theta}(\mathbf{Y})$, where we do not restrict to c = 1, that possesses better properties - particularly lower MSE.

We wish to examine how our results may alter when using the sample mean shrinkage estimator $c\bar{Y}(= c$ sm), where we do not restrict to $c = 1$.

Firstly we note that $\mathbb{E}\{c\bar{Y}\} = c\,\mu_Y = 1000\,c$ which does not necessarily equal $\theta = 1000 = \mu_Y$ and so we recognise that in general this is a biased estimator. Hence

$$\begin{aligned}\mathrm{MSE}\{c\bar{Y}\} &= \mathrm{var}\{c\bar{Y}\} + (\mathrm{bias})^2 \\ &= c^2\frac{\mu_Y^2}{n}\,\mathrm{CoV}^2 + (c-1)^2\mu_Y^2 \\ &= \frac{\mu_Y^2}{n}\,[c^2\mathrm{CoV}^2 + n(c-1)^2]\end{aligned}$$

which can be easily shown to attain its minimum when c^* is defined as

$$c^* = \frac{n}{n + \mathrm{CoV}^2}$$

and at this value

$$\mathrm{MSE}\{c^*\bar{Y}\} = \frac{\mu_Y^2}{n}\,\mathrm{CoV}^2 \times c^* = c^* \times \mathrm{MSE}\{\bar{Y}\}$$

and hence RMSE$\{c^*\bar{Y}\}$ <RMSE$\{\bar{Y}\}$ because $0 < c^* < 1$.

When $n = 20$ and CoV $= 2$ then the minimum value of RMSE$\{c\bar{Y}\}$ is attained when $c^* = 0.8333$ and the smaller RMSE is achieved at the expense of bias in the shrinkage estimator which produces the value $\mathbb{E}\{c^*\bar{Y}\} = 833$.

As either n becomes large and/or CoV becomes small then $c^* \rightarrow 1$ and hence $\text{RMSE}\{c^* \bar{Y}\} \rightarrow \text{RMSE}\{\bar{Y}\}$, as is shown in the table below when compared with Table 5.15

	Sample size(n)				
CoV	20	50	200	500	2000
0.25	56	35	18	11	6
0.50	111	71	35	22	11
1.00	218	140	71	45	22
1.50	318	208	105	67	34
2.00	408	272	140	89	45

Table 5.21: RMSE for shrinkage sample mean estimator evaluated at c^*

Briggs *et al* (2005) refer to the paper by O'Hagan and Stevens (2002) and their use of the logNormal estimator $\exp(\bar{X} + S^2/2)$, evaluated for a logNormal distribution.

If we now examine the shrinkage estimator $\widehat{\mu}_c = c\exp(\bar{X} + S^2/2) = c\,\widehat{\mu}_Y$ then it has the following properties

$$\mathbb{E}\{\widehat{\mu}_c\} = c\,\mathbb{E}\{\widehat{\mu}_Y\}$$

$$\begin{aligned}
\text{MSE}\{\widehat{\mu}_c\} &= \text{var}\{c\,\exp(\bar{X} + S^2/2)\ \} + (\text{bias})^2 \\
&= c^2\,\text{var}\{\ \exp(\bar{X} + S^2/2)\ \} + (c\,\mathbb{E}\{\widehat{\mu}_Y\} - \theta)^2 \\
&= c^2\,(\mathbb{E}\{\widehat{\mu}_Y\})^2 \left[\exp(\sigma^2/n)\left(1 - \frac{2\sigma^2}{n-1}\right)^{-\frac{n-1}{2}} \left(1 - \frac{\sigma^2}{n-1}\right)^{n-1} - 1\right] \\
&+ \ (c\,\mathbb{E}\{\widehat{\mu}_Y\} - \mu_Y)^2
\end{aligned}$$

when $\mathbb{E}\{\widehat{\mu}_Y\} = \exp(\mu + \sigma^2/2n)\left(1 - \frac{\sigma^2}{n-1}\right)^{-\frac{n-1}{2}}$ and $\theta = \mu_Y = 1000$ (in this case).

and the MSE$\{\widehat{\mu}_c\}$ attains its minimum when

$$c^{**} = \frac{\mu_Y}{\mathbb{E}\{\widehat{\mu}_Y\}\exp(\sigma^2/n)\left(1-\frac{2\sigma^2}{n-1}\right)^{-\frac{n-1}{2}}\left(1-\frac{\sigma^2}{n-1}\right)^{n-1}}.$$

We can now present the results of calculating the optimum values of c for the two estimators $c\bar{Y}$ and $c\ \exp(\bar{X}+S^2/2)$ as

	$c\bar{Y}$				$c\ \exp(\bar{X}+S^2/2)$			
	Sample size				Sample size			
CoV	10	15	20	30	10	15	20	30
0.25	0.994	0.996	0.997	0.998	0.991	0.994	0.995	0.997
0.50	0.976	0.984	0.988	0.992	0.963	0.975	0.981	0.988
1.00	0.909	0.938	0.952	0.968	0.861	0.907	0.930	0.954
1.50	0.816	0.870	0.899	0.930	0.724	0.816	0.862	0.907
2.00	0.714	0.790	0.833	0.882	0.582	0.719	0.788	0.857

Table 5.22: Optimum values of c for the $c\bar{Y}$ estimator and also the $c\ \exp(\bar{X}+S^2/2)$ estimator applied to the logNormal distribution

where we can observe that as either n becomes large and/or CoV becomes small then c^* and $c^{**} \to 1$ when the RMSE for the estimator and its corresponding minimum value shrinkage estimator tend to the same value.

It is only when $n < 30$ and CoV > 1 that the value of c that minimises the RMSE becomes significantly less than 1.

However we still need to evaluate the effect on the resulting value of the RMSE which we will show next, although we must not lose sight of the fact that we do not know the value of c that minimises the RMSE as we do not know the population value for the CoV.

We are now able to present RMSE's in two tables which compare the sm = $\bar{Y}$, $\widehat{\mu}_Y = \exp(\bar{X} + S^2/2)$ with the shrinkage versions of these estimators computed at their optimum values of $c^* \bar{Y}$ and $c^{**} \exp(\bar{X} + S^2/2)$ and also the Bpe (using the default prior for R_Y and seed value 0710 with 10K replications) where each cell in the tables below contains the RMSE for the estimators in the positions shown

$c^* \bar{Y}$	Bpe	$c^{**} \exp(\bar{X} + S^2/2)$
$\bar{Y}$		$\exp(\bar{X} + S^2/2)$

We can deduce from Tables 5.23 and 5.24 below that the effect of using the shrinkage estimators is not generally large and is not even detectible unless n is sufficiently small and CoV sufficiently large

Although values for c other than 1 and its optimum (in the sense of minimum RMSE) have been examined across the range of combination of values for the CoV and the sample size n they do not appear to offer any significant advantages and $c = 1$ has been retained as its simplicity is very appealing.

The recommendation for choice of estimator, when the population value for CoV and the underlying parametric distribution is unknown, is to use the Bpe because for values of CoV ≥ 1 the Bpe has the lowest RMSE. For CoV $= 0.5$ the Bpe has comparable RMSE to the other estimators when $n \geq 50$ and its only when CoV $= 0.5$ with $n \leq 30$ and also when CoV $= 0.25$ that the the Bpe does not have the lowest RMSE.

These recommendations have been established when comparing the sample mean shrinkage estimator (for Gamma and logNormal distributions) against the Bpe and $\{\exp(\bar{X} + S^2/2)\}$ shrinkage estimators for logNormal distributions.

	Sample size			
CoV	10	15	20	30
0.25	79 79 96 79 80	64 65 74 65 65	56 56 62 56 56	46 46 50 46 46
0.50	156 158 183 158 162	128 129 145 129 131	111 111 123 112 113	91 91 98 91 92
1.00	302 310 294 316 345	250 251 242 258 270	218 217 211 224 229	180 177 173 183 183
1.50	429 446 359 474 561	361 359 301 387 416	318 309 266 335 345	264 251 223 274 270
2.00	535 562 409 632 818	459 450 348 516 570	408 387 313 447 460	343 314 265 365 351

Table 5.23: RMSE : values of n from 10 to 30

	Sample size			
CoV	50	200	300	400
0.25	35 35 40 35 35	18 18 23 18 18	14 14 20 14 14	13 13 18 13 13
0.50	71 70 73 71 71	35 35 36 35 35	29 29 30 29 29	25 25 27 25 25
1.00	140 137 134 140 140	71 68 68 71 69	58 56 55 58 56	50 48 48 50 48
1.50	208 194 177 212 203	105 97 94 106 98	86 79 77 87 80	75 68 67 75 69
2.00	272 242 213 283 259	140 121 116 141 123	115 98 95 115 100	100 85 83 100 86

Table 5.24: RMSE : values of n from 50 to 400

5.10 Comparisons with other estimators for the mean of logNormal distributions

In a series of papers published from 1997 onwards Xiao-Hua Zhou, either alone or with others, looked at a number of aspects of estimation for the logNormal distribution in a health cost context.

We will look here at Zhou (1998) where he compares his four point estimators, $\widehat{\theta}_i$ for $i = 1, 2, 3, 4$, for the mean of a logNormal distribution and evaluates them in a classical framework using the expected mean square error. The comparisons are made for the relative mean square error

$$\mathbb{E}\left\{\frac{(\widehat{\theta}_i - \theta)^2}{\theta^2}\right\}$$

which has the technical advantage that the population logNormal mean θ does not need to be specified (for his four point estimators).

We will continue with our notation of Section 5.7, where we have interchanged the Zhou (1998) use of X and Y. So if $Y \sim \text{logN}(\mu, \sigma^2)$ then the mean value of Y is defined as $\theta = \mu_Y = \exp(\mu + \sigma^2/2)$ with variance $\sigma_Y^2 = \mu_Y^2[\exp(\sigma^2) - 1]$. If we express Y as $\log Y = X \sim \text{N}(\mu, \sigma^2)$ then the four estimators that he compared are

the sample mean $\widehat{\theta}_1 = \bar{Y}$,
a uniformly minimum variance unbiased (UMVU) estimator,
a maximum likelihood estimator (ML) $\widehat{\theta}_3 = \exp\left(\bar{X} + \frac{n-1}{2n} S^2\right)$
a conditionally minimal (MSE) estimator (CMMSE) $\widehat{\theta}_4 = \exp(\bar{X}) g_{n-1}\left(\frac{n-4}{2(n-1)} S^2\right)$
where $g_m(t) = \sum_{r=0}^{\infty} \frac{1}{r!} \frac{m+2r}{m} \left(\frac{m}{m+1} t\right)^r \prod_{i=1}^{r} \frac{m}{m+2i}$

and because of the complex nature of some of the expressions for the MSE for these estimators he undertakes numerical comparisons and takes σ^2 to be 0.19 to 4.94 incremented by 0.19. He also takes the sample size, n, to be 7, 11, 20, 30, 40, 50, 155 and 200.

He is able to show that, for any fixed value of n, the mean square error of the conditionally minimum MSE estimator is uniformly smaller than that of the UMVU estimator, the ML estimator or the sample mean.

We will concentrate here on the case $n = 20$ and consider the Relative MSE (RelMSE) for the following estimators

the sample mean $\widehat{\theta}_1 = \bar{Y}$,
the logNormal estimator $\widehat{\mu}_Y = \exp(\bar{X} + S^2/2)$
from O'Hagan and Stevens (2002), which is clearly closely related to
the maximum likelihood estimator $\widehat{\theta}_3 = \exp\left(\bar{X} + \frac{n-1}{n} S^2/2\right)$
of Zhou (1998),
his conditionally minimal MSE estimator $\widehat{\theta}_4$
and, evaluated when $R_Y \sim 1+\mathrm{G}(11.5,6)$, our Bpe.

We know from Zhou (1998) that the RelMSE $\{\widehat{\theta}_4\}$ does not involve θ and is a function of σ^2 and n only.

Also from Zhou (1998) and Section 5.7 we know that the RelMSE $\{\widehat{\theta}_1\}$ does not involve θ and is a function of σ^2 and n only as

$$\text{RelMSE }\{\widehat{\theta}_1\} = \frac{\text{MSE }\{\widehat{\theta}_1\}}{(\theta)^2} = \frac{\text{var }\{\bar{Y}\}}{(\mu_Y)^2} = \frac{(\text{ CoV })^2}{n} = \frac{\exp(\sigma^2) - 1}{n}$$

remembering that in this Zhou (1998) context we are only dealing with logNormal distributions.

To evaluate the RelMSE $\{\exp(\bar{X} + S^2/2)\}$ we use Equation 5.1 from Section 5.7 to show RelMSE $\{\exp(\bar{X} + S^2/2)\}$ = RelMSE $\{\widehat{\mu}_Y\}$ where

$$
\begin{aligned}
\text{RelMSE } \{\widehat{\mu}_Y\} &= \frac{\text{MSE } \{\widehat{\mu}_Y\}}{(\mu_Y)^2} \\
&= \left\{ \exp(2\mu + \sigma^2/n) \left(1 - \frac{\sigma^2}{n-1}\right)^{-(n-1)} \right. \\
&\times \left[\exp(\sigma^2/n) \left(1 - \frac{2\sigma^2}{n-1}\right)^{-\frac{n-1}{2}} \left(1 - \frac{\sigma^2}{n-1}\right)^{n-1} - 1 \right] \\
&+ \left. \left[\exp(\mu + \sigma^2/2n) \left(1 - \frac{\sigma^2}{n-1}\right)^{-\frac{n-1}{2}} - \exp(\mu + \sigma^2/2) \right]^2 \right\} \\
&\div \ [\exp(\mu + \sigma^2/2)]^2 \\
&= \exp(\sigma^2/n - \sigma^2) \left(1 - \frac{\sigma^2}{n-1}\right)^{-(n-1)} \\
&\times \left[\exp(\sigma^2/n) \left(1 - \frac{2\sigma^2}{n-1}\right)^{-\frac{n-1}{2}} \left(1 - \frac{\sigma^2}{n-1}\right)^{n-1} - 1 \right] \\
&+ \left[\exp(\sigma^2/2n - \sigma^2/2) \left(1 - \frac{\sigma^2}{n-1}\right)^{-\frac{n-1}{2}} - 1 \right]^2
\end{aligned}
$$

which is a function of σ^2 and n only.

To determine the RelMSE {Bpe} we are unable to obtain the result in a closed analytical format and have needed to obtain numerical results by simulation, for the specific value of $\mu_Y = 1000$ (so far). However it may be that the value of the RelMSE {Bpe} will not be too sensitive to changes in the value of μ_Y. This can be examined by evaluating the RelMSE for the Bpe for 10K replications for the 0710 seed value, sample size 20, $R_Y \sim$ 1+G(11.5,6) for logNormal distributions with the CoV = 0.25, 1 and 2, and with 1000K replications for CoV = 5 and 10, over a range of values for μ_Y where the results are presented in Table 5.25 overleaf. While we can observe small changes in the value of the RelMSE as μ_Y increases this is not so significant as to not be able to say that the value of the RelMSE for the Bpe may be determined when $\mu_Y = 1000$ and then compared with the RelMSE for the other estimators (where it was unnecessary to specify the value of μ_Y).

	CoV				
μ_Y	0.25	1.00	2.00	5.00	10.0
250	0.003888	0.044526	0.097655	0.263609	0.446111
500	0.003886	0.044490	0.097670	0.263827	0.446407
1000	0.003885	0.044455	0.097686	0.264046	0.446703
2000	0.003883	0.044419	0.097702	0.264264	0.446998
5000	0.003880	0.044373	0.097724	0.264553	0.447389

Table 5.25: RelMSE for values of μ_Y and CoV

Whilst we can obtain numerical evaluations of the theoretical results for the first three estimators we cannot obtain theoretical results for the Bpe. We are able to obtain results by simulation for our first two estimators and, by comparing these results with those obtained theoretically, can justify the direct comparison between the Zhou (1998) theoretical CMMSE and our Bpe obtained by simulation.

The Zhou (1998) paper takes σ^2 to be between 0.19 and 4.94 which, if we note the relationship $\text{CoV}^2 = \exp(\sigma^2) - 1$ from our earlier work on the logNormal distribution in Section 5.7, tells us that CoV varies between 0.46 and 11.78. So although the Zhou (1998) paper covers most of the same range of CoV, 0.5 to 2, as the Briggs *et al* (2005) paper it also recognises that much more skew logNormal distributions may be possible.

This extreme skewness will significantly increase the care that will be necessary to obtain samples that adequately represent the population from which they have been drawn in the sense of first and second order moments, as the following table indicates.

Following the work in Section 5.2, for an increasing sequence of p values, the p-percentiles for the random variable Y, following the two distributions shown whose mean = 1000 are

	logNormal distribution	
p values	CoV = 2	CoV = 11.78
90	2273	1460
99	8555	14888
99.9	22548	81317
99.99	50067	328945
99.999	100070	1106715
99.9999	185981	3277924
99.99999	327454	8831256
99.999999	552729	22097919
99.9999999	901730	52090121
99.99999999	1430110	11685871

Table 5.26: p-percentiles

where we can see that as the the p value increases, then the p-percentile values for CoV = 11.78 are significantly greater than those for CoV = 2.

To be able to draw samples from the logNormal distributions that reasonably reflected the population first and second order moments from which they have been drawn required very careful consideration of the seed value and number of replications to be used. While the choice of seed value 0710 with 100K replications was satisfactory for values of CoV less than 1.89, the number of replications was increased to 1000K for values of CoV up to 6 and for values of CoV greater than 6 then the seed value 6561684 combined with 1000K replications was necessary to obtain acceptable results as shown in the following table

population				sample values		
σ^2	CoV	mean	sd	mean	sd	CoV
0.19	0.457	1000	457	1000	457	0.457
0.38	0.680	1000	680	1000	680	0.680
0.57	0.877	1000	877	1000	877	0.877
0.76	1.067	1000	1067	999	1067	1.068
0.95	1.259	1000	1259	999	1260	1.261
1.14	1.458	1000	1458	999	1461	1.462
1.33	1.668	1000	1668	999	1672	1.673
1.52	1.890	1000	1890	1000	1891	1.890
1.71	2.128	1000	2128	1000	2129	2.128
1.90	2.385	1000	2385	1000	2386	2.385
2.09	2.662	1000	2662	1000	2663	2.662
2.28	2.963	1000	2963	1000	2963	2.962
2.47	3.290	1000	3290	1000	3290	3.288
2.66	3.646	1000	3646	1000	3645	3.643
2.85	4.036	1000	4036	1000	4031	4.029
3.04	4.462	1000	4462	1000	4451	4.449
3.23	4.927	1000	4927	1000	4909	4.907
3.42	5.438	1000	5438	1001	5407	5.404
3.61	5.997	1000	5997	1001	5948	5.945
3.80	6.611	1000	6611	1000	6687	6.688
3.99	7.284	1000	7284	1000	7361	7.362
4.18	8.023	1000	8023	1000	8096	8.096
4.37	8.834	1000	8834	1000	8896	8.894
4.56	9.725	1000	9725	1000	9765	9.761
4.75	10.704	1000	10704	1001	10707	10.701
4.94	11.780	1000	11780	1001	11728	11.719

Table 5.27: logNormal distributions, sample size = 20

We will use the above range of the variance of the log-scale σ^2 in the following plots.

For the following four plots we will use a solid line for the numerical evaluation of theoretical results and a dotted line for simulated values.

The plot below shows that the theoretical (smtheo) and simulated (smsim) values for the sample mean estimator are very close.

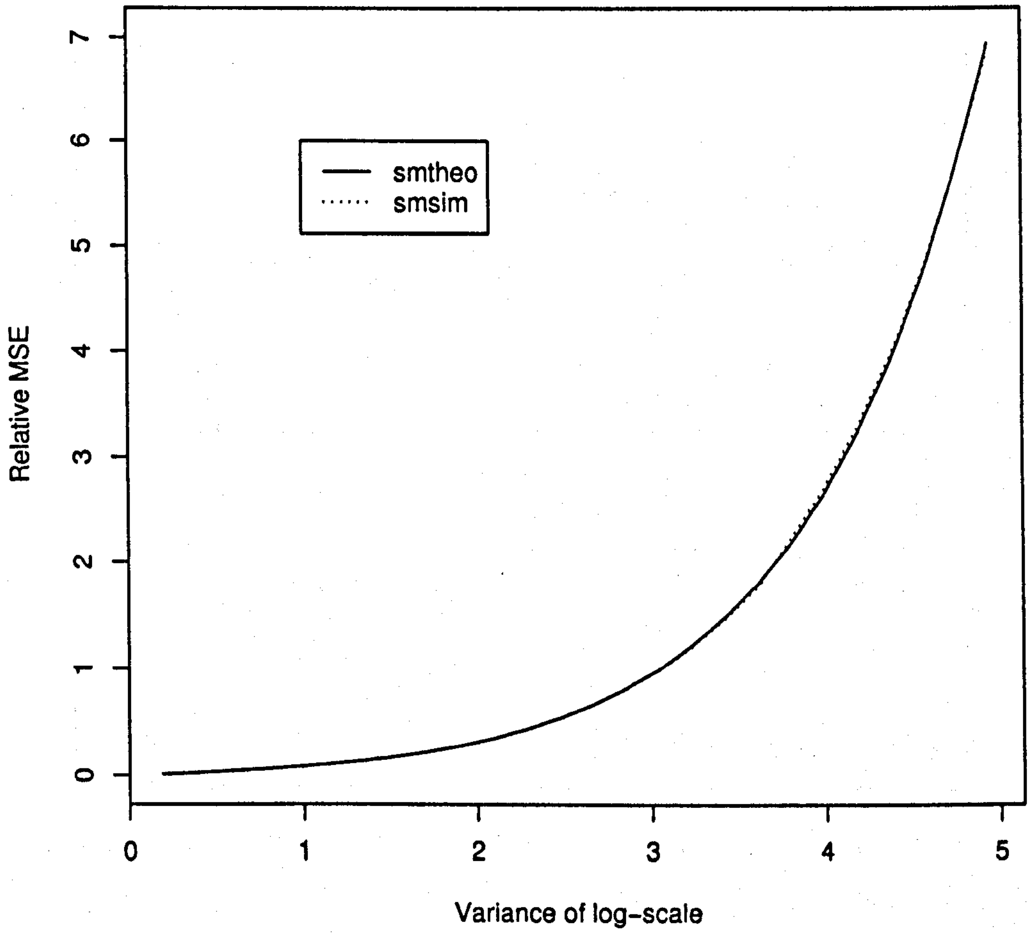


Figure 5.6: Plot showing the theoretical and simulated sample mean estimators

This plot shows that the theoretical and simulated values for the $\exp(\bar{X}+S^2/2)$ estimator show close agreement when σ^2 or the CoV has small values. However, as σ^2 increases, there is some evidence that the simulated values are lower than the theoretical values. This is not entirely unexpected because the results shown in Table 5.27 are for the whole of the sample, ie 1000K replications for sample size 20 or 20,000K values in total. There will be more variation in individual samples of size 20 and this will become more manifest as the CoV increases - particularly for second order moments.

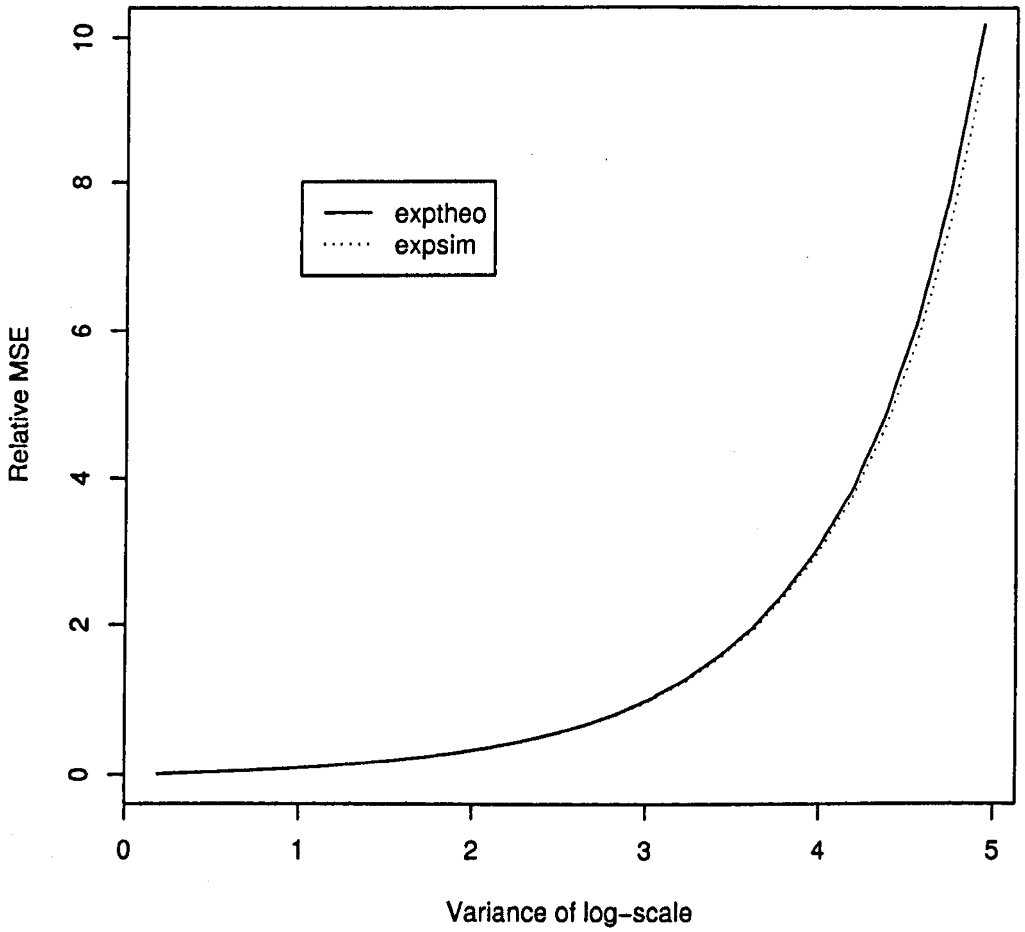


Figure 5.7: Plot showing the theoretical and simulated exponential estimators

This plot compares the theoretical CMMSE with the simulated Bpe and from the two previous plots we can be confident that this is a valid comparison, when σ^2 is below four in particular.

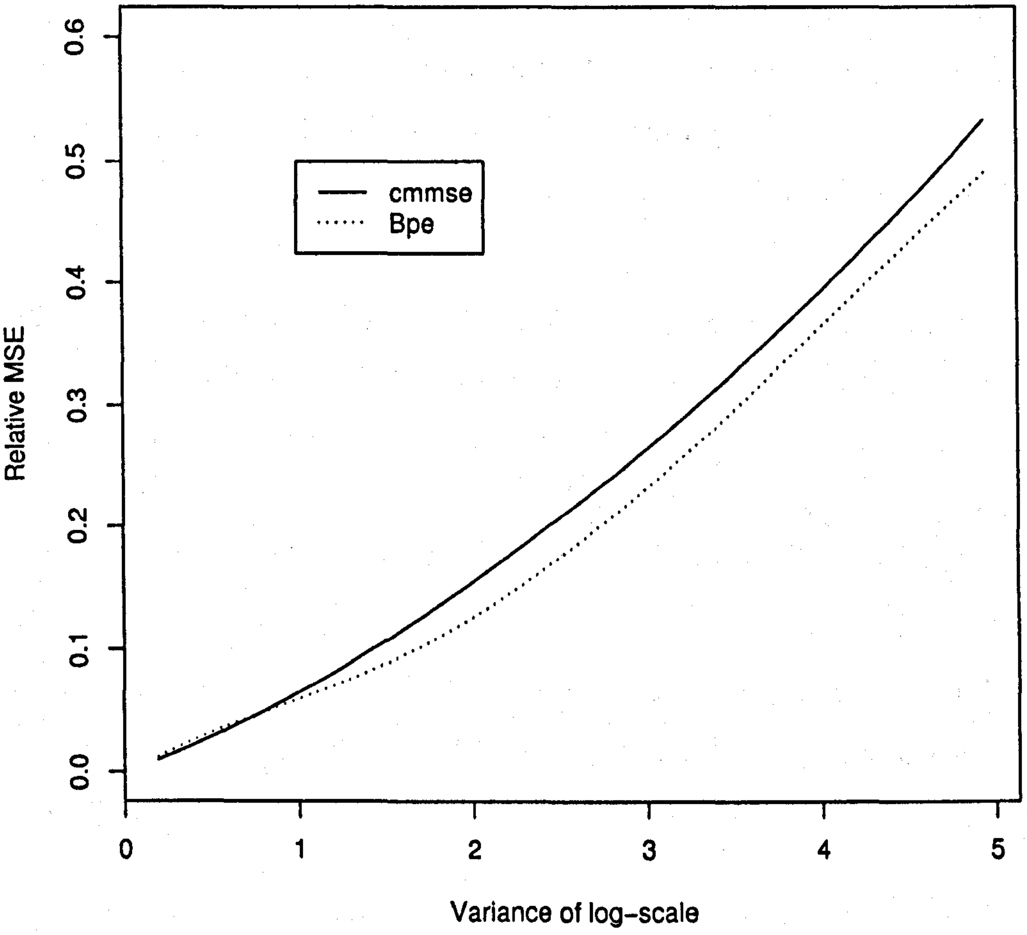


Figure 5.8: Plot comparing the CMMSE and Bpe estimators

We may conclude that our Bpe has a lower Relative MSE than the Zhou (1998) CMMSE for $\sigma^2 > 1$. It is worth noting in passing that our Bpe uses the default prior $R_Y \sim 1+G(11.5,6)$ which has been trained for both logNormal and Gamma distributions, using the values CoV = 0.25 and 2.00, $\mu_Y = 1000$ and sample size 20 when establishing its value for the TRMSE.

If we search for a default prior for R_Y that minimises the Total RelMSE (TRelMSE) defined as

$$\text{TRelMSE } = \Sigma \text{ RelMSE}$$

which gives equal weights to the two parametric models considered, namely the logNormal distribution for CoV = 0.457438 and CoV = 11.78008 (corresponding to the range limits chosen by Zhou of $\sigma^2 = 0.19$ and 4.94 respectively), then we arrive at $R_Y \sim$ 1+G(5,7.5) when $\mu_Y = 1000$ and sample size 20.

The default prior used earlier was $R_Y \sim$ 1+G(11.5,6) and the changes have arisen because in the Zhou type calculations (TRelMSE) only the logNormal was considered as a parametric model, the range of values for CoV was changed and because the RelMSE gives a much higher weighting (the square of the value for RMSE) to the larger RelMSE, which occurs when CoV = 11.78, for constant μ_Y.

Plot 5.9 overleaf compares the theoretical CMMSE with the simulated Bpe using $R_Y \sim 1 +$ G(5,7.5). Comparing Plots 5.8 and 5.9 it is possible to see that the Bpe plots are of the same type of shape although Plot 5.8 is not so pronounced as Plot 5.9.

Again we can identify a region where the Bpe has lower RelMSE than the CMMSE estimator and this occurs for values of $\sigma^2 > 3$.

We have derived two default priors from particular loss functions and logics to minimise totals of these loss functions. In both of the cases analysed we have been able to establish regions of the range of the Variance of the log-scale σ^2 where the Bpe estimator performs better, in the sense of lower RelMSE, than the CMMSE estimator.

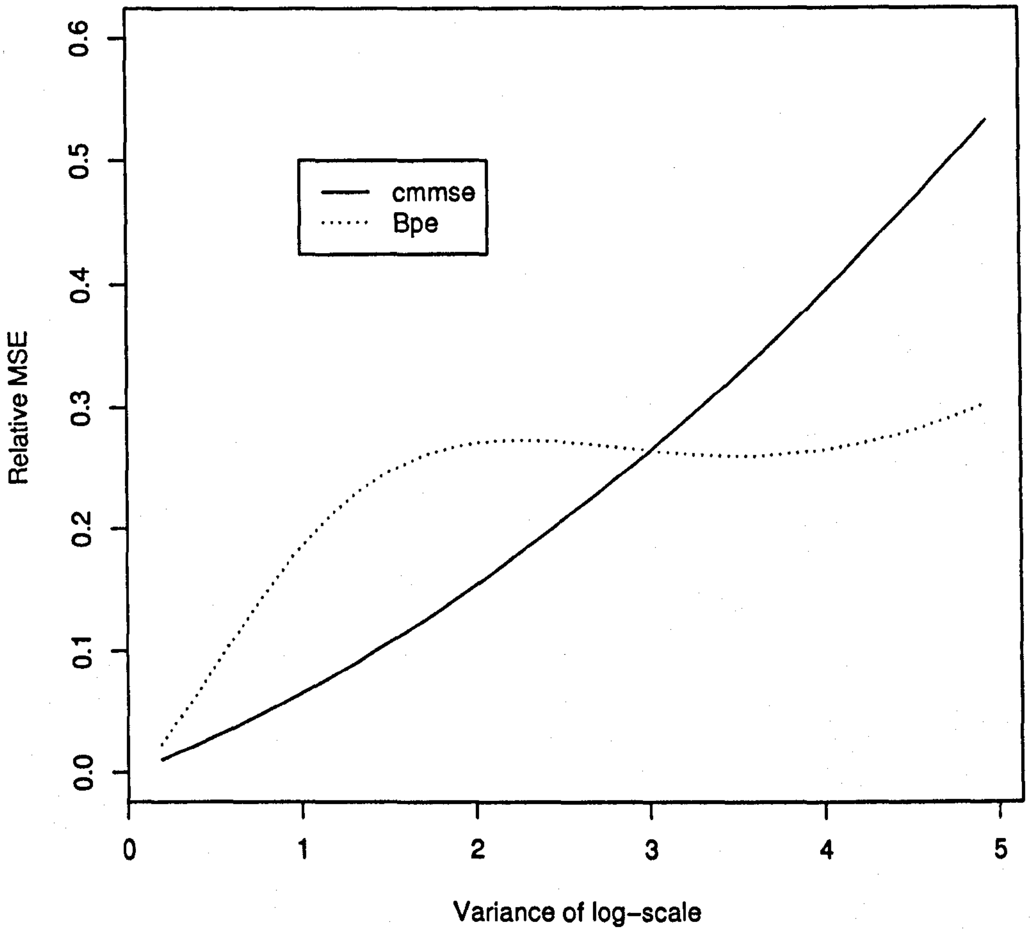


Figure 5.9: Plot comparing the CMMSE and Bpe estimators

Plot 5.9 above does illustrate the care that is needed when choosing default priors because if the prior has been trained by a very specific set of parametric models, as opposed to a set of parametric models that reflects the possibility of the broader underlying conditions that may arise in practise, then it will respond to those very specific set of models. The results that will be produced, in the sense of MSE, while they may be very good for some circumstances may be much worse for others. The default prior $R_Y \sim 1+\mathrm{G}(11.5,6)$ has been trained over such a range of broad underlying conditions and does not in general perform particularly badly.

Chapter 6

Elicitation

6.1 Introduction

The contents of Section 6.2 are based in the main on O'Hagan *et al* (2006) with its associated references where our main innovation is a method of eliciting beliefs about a shape parameter.

Elicitation may be described as the process of a facilitator capturing an expert's prior beliefs about an unknown quantity. We will follow the convention of ascribing the gender of the facilitator as male and the expert as female. The facilitator seeks to obtain her beliefs in the form of a probability distribution which both expresses her uncertainty and also enables him to combine the two sources of information, her prior beliefs and the data, into a single source of information, the posterior distribution.

The data model for costs selected in Chapter 2 is the logNormal distribution which has two parameters, μ for scale and σ^2 for shape, which are not of course directly observable. Moreover, the role of these parameters is not immediately obvious. We do however need our expert's prior beliefs about μ and σ^2 expressed as their joint distribution and so it is the hyperparameters in the joint distribution that are of interest.

If we can find a way to express this joint distribution in terms of functions of the parameters that are independent then the process of elicitation becomes easier as we now need to elicit two independent marginal distributions. If, furthermore, each of the two independent distributions only involve (a function of) one of the parameters then these marginal distributions are univariate with respect to the parameters and may be considerably easier to elicit than their joint distribution.

It is generally recognised as good practise to elicit beliefs about quantities that are directly observable. We will follow this practise wherever possible but will need a different approach when eliciting beliefs about the shape parameter.

Although we have talked here about using a logNormal data model we will only tell our expert, see The elicitation in the Appendix, that our data model is to be a distribution that takes non-negative values, unbounded above and right skewed.

The logNormal distribution has a number of properties that may be considered to be useful for this problem.

Its Median is $\exp(\mu)$ and so is a simple function of μ alone and this is the way that we will choose to model the scale parameter.

Its shape parameter is σ^2 and we now look to determine a function of σ^2 alone that is (statistically) independent of the median, where independence means knowing the value of one of the parameters does not give the expert any information about the value of the other parameter.

Kurtosis, the degree of peakedness of a distribution, is defined as the fourth standardized moment μ_4/σ^4 where μ_4 is the fourth moment about the mean and σ^2 is the variance. This is a ratio of two different quantities. Also commonly used is the measure of (excess) kurtosis, $(\exp(4\sigma^2) + 2\exp(3\sigma^2) + 3\exp(2\sigma^2) - 6)$, and either quantity contains σ^2 alone but is not directly observable.

Skewness, the degree of asymmetry of a distribution, is defined as the third standardized moment μ_3/σ^3 where μ_3 is the third moment about the mean and σ^2 is the variance. This is a ratio of two different quantities. Skewness may be defined as $([\exp(\sigma^2)+2]\sqrt{\exp(\sigma^2)-1})$ which contains σ^2 alone but is not directly observable.

The coefficient of variation, defined as the ratio of standard deviation to mean, is useful as a relative measure of dispersion, or variation, but is the ratio of two different quantities. The coefficient of variation, or CoV, is $[\exp(\sigma^2)-1]^{\frac{1}{2}}$ and so σ^2 alone is present but this ratio is not directly observable.

The logNormal distribution is unimodal where the value of its mode is expressed as $\exp(\mu)/\exp(\sigma^2)$. The ratio of its median to mode is $\exp(\sigma^2)$. As the mode of a logNormal distribution is a visible feature we do have concerns that the mode will become an anchor point if it is used in an elicitation. Anchoring is the description given when elicited values remain close to some initial value and the elicitation yields values that are too conservative. While the ratio of median to mode is a function of σ^2 alone it is once again the ratio of two different quantities and the ratio is not directly observable.

The $\Phi(q)$-quantile for the logNormal distribution is $\exp(\mu + q\sigma)$ and so the quantile ratio, defined as the ratio of its $\Phi(q)$-quantile to its median, is $\exp(q\sigma)$. This ratio is a function of σ^2 alone. The logNormal distribution is uni-modal and positively skewed. In any plot of this logNormal distribution the mode is a dominant feature and it's possible that this may influence an expert's judgement. To reduce this influence we have chosen to work with quantiles whose theoretical values always lie on the same side of the mode, in this case to the right, where the range of numerical values is also larger than to the left. We will work with the $\frac{1}{2}$-quantile, or median, and the $\frac{3}{4}$-quantile. So the quantile ratio is a ratio of similar quantities but the quantile ratio is not directly observable.

We have outlined above five functions of σ^2 alone that we may consider as candidates to elicit prior beliefs for the shape parameter, namely

kurtosis
skewness
coefficient of variation
ratio of median to mode
quantile ratio

None of these candidate functions are directly observable and we believe that to make the best choice we have to eliminate the worse options.

Kurtosis and skewness are mainly used in descriptive statistics to summarise the properties of samples. We believe it would be difficult for an expert to form prior beliefs about them for a population.

The coefficient of variation and the ratio of median to mode are both ratios of different quantities and we believe it would be more difficult for an expert to form prior beliefs about them than a ratio of similar quantities.

The Quantile Ratio is a ratio of similar quantities and so arises in a natural way and we believe represents the best opportunity for an expert to express prior beliefs and it is the Quantile Ratio that we have chosen to work with here.

The Median (a value expressed in monetary units) and Quantile Ratio (a value expressed as a (dimensionless) number) are statistically independent if knowing the value of one of them does not change the expert's beliefs about the other one.

Whilst there is no obvious reason to believe that the Median, $\exp(\mu)$, and the Quantile Ratio, $\exp(q\sigma)$, are dependent, the assumption that we now make of independence between them will of course need to be examined when conducting the elicitation.

Hence we can restructure the problem so that we will need to elicit prior beliefs for functions of each parameter alone and propose now to work with the Median, $\exp(\mu)$, for scale and the Quantile Ratio, $\exp(q\sigma)$, for shape. As μ and σ^2 are fixed but unknown for this population then so are the Median and Quantile Ratio.

We will work with a logNormal data model whose parameters μ and σ^2 will be restructured as the Median $M_Y = \exp(\mu)$ and the Quantile Ratio $R_Y = \exp(q\sigma)$.

We will also work with the Third Quartile $T_Y(\boldsymbol{\theta})$ for the logNormal random variable Y which to simplify nomenclature we will represent as T_Y and (although the theory developed so far in this chapter and elsewhere in this thesis relates to a Quantile Ratio) we have chosen to work with a ratio of quartiles, or a Quartile Ratio directly, in this elicitation. To help distinguish the Median, Third Quartile and Quartile Ratio from their elicited quantile values we will refer for the rest of this chapter, and most particularly when communicating with the expert, to the Quartile Ratio, denoted as R_Y, rather than the Quantile Ratio.

In The elicitation, located in the Appendix, we will describe the four stage procedure that we will follow for eliciting μ and σ^2, the expert's prior beliefs for the parameters of the logNormal data model, by eliciting her prior beliefs for the population Median and Quartile Ratio. Each of the stages will comprise a number of steps.

We propose in the first stage to elicit the expert's beliefs for her $\frac{1}{2}$-quantile values for M_Y (Median) and T_Y (Third Quartile) for the population as $m(M_Y)$ and $m(T_Y)$ respectively. We are then able to obtain her plausible value for the Quartile Ratio as $m(T_Y)/m(M_Y)$ and we will use this plausible value in stage two where we elicit the uncertainty in the expert's prior beliefs for the Quartile Ratio.

The Quartile Ratio is not a quantity that can be directly observed and it may therefore be difficult to elicit an expert's prior beliefs. The purpose of obtaining a plausible value is to provide guidance to the expert about the numerical value that the quartile ratio may take to ensure that the facilitator is able to conduct an elicitation that does represent our expert's prior beliefs.

In Section 6.3 we will describe the Case studies or elicitations that have been conducted.

The reasons why we were unable to use the 3CPO cost data set, which was generated between July 2003 and April 2007 and thus is relatively recent (when compared to the pMDI+ data set in particular which dates from 1991 and 1992), will be explained in Section 6.3.1.

For each of the elicitations that will be described in Sections 6.3.2 and 6.3.4 Peter Gregory is the facilitator. We will follow elicitation convention by assigning a male gender to the facilitator but will assign the true gender to the expert. For the two elicitations that were conducted, in both cases we were only able to follow The elicitation procedure in the Appendix up to and including the practise elicitation on distribution of salary where, in both cases, the random variable Y represents salary in the appropriate population. Comments on the two elicitations will be made in Sections 6.3.3 and 6.3.5 respectively.

The elicitations were conducted following the Sheffield Elicitation Framework using the Distribution Fitting Tool for one expert using the tertile elicitation method to fit positively skewed distributions.

Although we were unable to use the 3CPO data set, in Section 6.3.6 we will postulate what the outcome might have been if this had been possible for the data sets collected in Sheffield.

6.2 Theory

We will assume that the joint prior distribution for the Median and the Quartile Ratio factorises into independent priors for the Median and the Quartile Ratio. Hence we know from Section 4.2.4 that our joint prior distribution for μ and σ^2 may be factorised as a product of functions of μ alone and σ^2 alone. However, these prior distributions for μ and σ^2, except in a few special cases, will not follow any recognisable known parametric form.

We will work with the Median, M_Y, Third Quartile, T_Y, and also the particular Quartile Ratio, $R_Y = T_Y/M_Y$, requiring $q = 0.6745$.

It is worth noting the results from an experiment that used samples conducted by Peterson & Miller (1964). It concluded that its subjects were more proficient at estimating medians than means or variances when the sample distributions were highly skewed. Whilst our ultimate interest is eliciting our experts beliefs about the parameters we take account of these sample results as we start our process by eliciting beliefs about population quartiles.

The first stage concerns the Median M_Y and Third Quartile T_Y of the random variable Y that represents the cost (of treatment) in our population where the expert has (only) been told that an appropriate model for Y is a distribution that takes non-negative values and is skewed to the right.

Initially we concentrate on the Median. At the 1st step we will ask for the largest value for M_Y that the expert believes is possible. It is well known that experts are prone to over-confidence, see Kerens (1991), by not recognising all the uncertainty that is present. Whilst we would always elicit this value for a quantity with a finite range, this question asks the expert to recognise that very large cost values are possible. In the 2nd step we can, after noting $M_Y > 0$, elicit the $\frac{1}{2}$-quantile value for M_Y, which we denote as $m(M_Y)$, by asking the expert to determine $m(M_Y)$ such that M_Y is equally likely to be above or below this value.

We now concentrate on the population Third Quartile T_Y. At the 3rd step we ask the expert for the largest value for T_Y that the expert believes is possible. At the 4th step we elicit the $\frac{1}{2}$-quantile value for T_Y, which we denote as $m(T_Y)$, by asking the expert to determine $m(T_Y)$ such that T_Y is equally likely to be above or below this value.

While the theoretical values for M_Y and T_Y show that $T_Y > M_Y$ here we are dealing with the elicited values $m(M_Y)$ and $m(T_Y)$ and in the unlikely event that $m(M_Y) > m(T_Y)$ then further discussion and clarification will be necessary before repeating the two steps above.

In the second stage, after recognising that the Quartile Ratio is bounded below by 1, we initially calculate a plausible value for R_Y as $m(T_Y)/m(M_Y)$ and we tell the expert these values. Our 1st step, for reasons comparable to those in stage one, is to ask for the largest value for R_Y that the expert believes is possible. For our 2nd step we will then elicit values for the $\frac{1}{3}$-quantile and $\frac{2}{3}$-quantile for R_Y by asking the expert to determine $l(R_Y)$ and $u(R_Y)$ such that R_Y is equally likely to lie below $l(R_Y)$ as above $u(R_Y)$ as between these two values. In the 3rd step we will elicit the $\frac{1}{2}$-quantile value for R_Y, which we denote as $m(R_Y)$, by asking the expert to determine $m(R_Y)$ such that R_Y is equally likely to be above or below this value. In the 4th step we will fit a "1+Gamma" distribution to the elicited values and ask for visual confirmation before giving feedback.

The third stage is to elicit the prior beliefs for the population Median M_Y. After recognising that M_Y is bounded below by 0, we will remind the expert that, in the first stage, she has determined the largest value she thinks possible for M_Y. The 1st step is to elicit values for the $\frac{1}{3}$-quantile and $\frac{2}{3}$-quantile for M_Y by asking the expert to determine $l(M_Y)$ and $u(M_Y)$ so that M_Y is equally likely to lie below $l(M_Y)$ as above $u(M_Y)$ as between these two values. The 2nd step is where we remind the expert that we have obtained in the first stage the $\frac{1}{2}$-quantile value for M_Y denoted $m(M_Y)$. In the 3rd step we are then able to use the elicited values to fit a logNormal distribution and ask for visual confirmation before giving feedback.

In the fourth stage we will examine the assumption of independence between the median and the quartile ratio.

We will tell the expert that we want to determine whether knowing the value of the Median gives her any information about the value of the Quartile Ratio, rather than telling her that it is independence that is of interest. If the expert was told that independence was the focus of our attention then we believe that this may convey information to our expert and may influence the beliefs that she expresses.

From the second stage we have been able to determine a confirmed (marginal) distribution for our expert's prior beliefs for the Quartile Ratio and similarly for the Median from the third stage.

We will show the expert her confirmed (marginal) prior distribution for the Quartile Ratio and ask her if there is any particular value that could be chosen for the Median from her agreed (marginal) prior distribution would that cause her to want to change her beliefs about what is now her conditional prior distribution for the Quartile Ratio.

If the expert does believe that the Median (a value expressed in monetary units) and Quartile Ratio (a value expressed as a (dimensionless) number) are dependent then we will ask
why she believes that there is a relationship between them
what is that relationship ?

If the expert continues to hold the view of dependency between the Median and Quartile Ratio then this dependency will need to be captured in our model.

6.2.1 Comments on the elicitation procedure

To fit distributions to the elicited values we will use the SHELF distribution fitting tool, see www.tonyohagan.co.uk/shelf. Although we do need to be able to elicit median values to use SHELF, the choice of tertiles values to be elicited (rather than quartiles) is made to help the expert particularly when eliciting values for M_Y and R_Y.

It is unclear, see Garthwaite *et al* (2005), which percentiles should be elicited for the variable-interval method and so choosing quartiles does not appear to be an inappropriate choice.

We will use the results from stage one to establish a plausible value for the Quartile Ratio in stage two. We have elicited the expert's beliefs for the Quartile Ratio in stage two so that the expert is asked about a quantity other than the Median before returning to the Median in stage three. We would expect to be able to obtain a better elicitation for the Median in stage three because the expert will have retained less of her thought processes from stage one after this "interruption".

6.2.2 Fitting distributions to the elicited values

We propose to use the SHELF distribution fitting tool.

Median

We will fit a logNormal distribution for M_Y.

Quantile ratio

We will fit a "1+Gamma" distribution for R_Y, where $S_Y = R_Y - 1$ and hence $S_Y \sim$ Gamma.

6.3 Case studies

6.3.1 The 3CPO Study : introduction

The motivation for this thesis was the medical cost data set for the pMDI+ patient group. A further medical cost data set was obtained courtesy of Prof S Dixon for the 3CPO study, see http://www.sheffield.ac.uk/trial3cpo/ for more details.

The 3CPO trial generates patient level costs for the three arms of the trial where the aim was to recruit 1200 patients over 26 sites in England and Scotland. It was possible to extract the data for the major recruitment centres. Our aim was to identify finance experts who had an understanding of acute cardiogenic pulmonary oedema (the 3CPO medical study area) with whom we could conduct elicitations of the costs for the Standard and the two alternative treatments.

The NHS traditionally sets budgets and then manages costs on a "top down" basis. There is however increasing interest in "bottom up", or patient-level, costs. The Department of Health (DoH) is encouraging, whilst not making mandatory, the implementation of patient-level information and costing systems (PLICS). We have been able to obtain from the DoH a list of the organisations that are in the process of implementing PLICS and were able to identify those organisations that the DoH said were implementing PLICS and had recruited patients to the 3CPO trial.

Having contacted a number of the organisations that had been major recruiters to the 3CPO trial, including those listed as in the process of implementing PLICS, a positive response was obtained from a couple, namely York and Bristol. However, they were both in the early stages of implementation and it has not been possible, in 2010, to identify personnel who possess the skill set required for them to be considered an expert capable of elicitation.

It was possible to identify clinicians who understood the medical condition but not costs or finance personnel who were beginning to understand patient-level costs but did not understand the medical condition. The skill set required by an expert should be developed over the coming months and years but is not available in 2010.

6.3.2 The first elicitation

For this elicitation the expert was Miss Irena Peel, BSc ACA MBA, Financial Controller, Royal Institute of British Architects (RIBA), on 24th January 2010.

Results obtained for the distribution of salary at RIBA can be summarised as

values elicited	M_Y	T_Y	R_Y
upper bound	35	40	1.5
median	30	38	1.2
lower tertile	20	-	1.1
upper tertile	35	-	1.3

Table 6.1: Initial values elicited for M_Y, T_Y and R_Y

where the plausible value for R_Y was approximately 1.3 and the entries for M_Y and T_Y represent thousands of pounds. There are logical lower bounds for M_Y and R_Y of 0 and 1 respectively.

As can be seen from Table 6.1 the value elicited for the upper tertile for M_Y was equal to the value that had been elicited earlier for the upper bound. When this was discussed with the expert she felt that she had been too restrictive with her initial belief for the value of the upper bound and this was increased to 40.

The elicited values that were used to fit the distributions were

values elicited	M_Y	T_Y	R_Y
upper bound	40	40	1.5
median	30	38	1.2
lower tertile	20	-	1.1
upper tertile	35	-	1.3

Table 6.2: Elicited values for M_Y, T_Y and R_Y used to fit a distribution

with SHELF fitting a logNormal distribution with mean $\mu = 10.181$ and variance $\sigma^2 = 0.441^2$ for M_Y.

It can be seen from Figure 6.1 below that although we have chosen to fit a logNormal distribution the alternative offered (for positively skewed distributions) in SHELF, a Gamma distribution, is a comparable distribution.

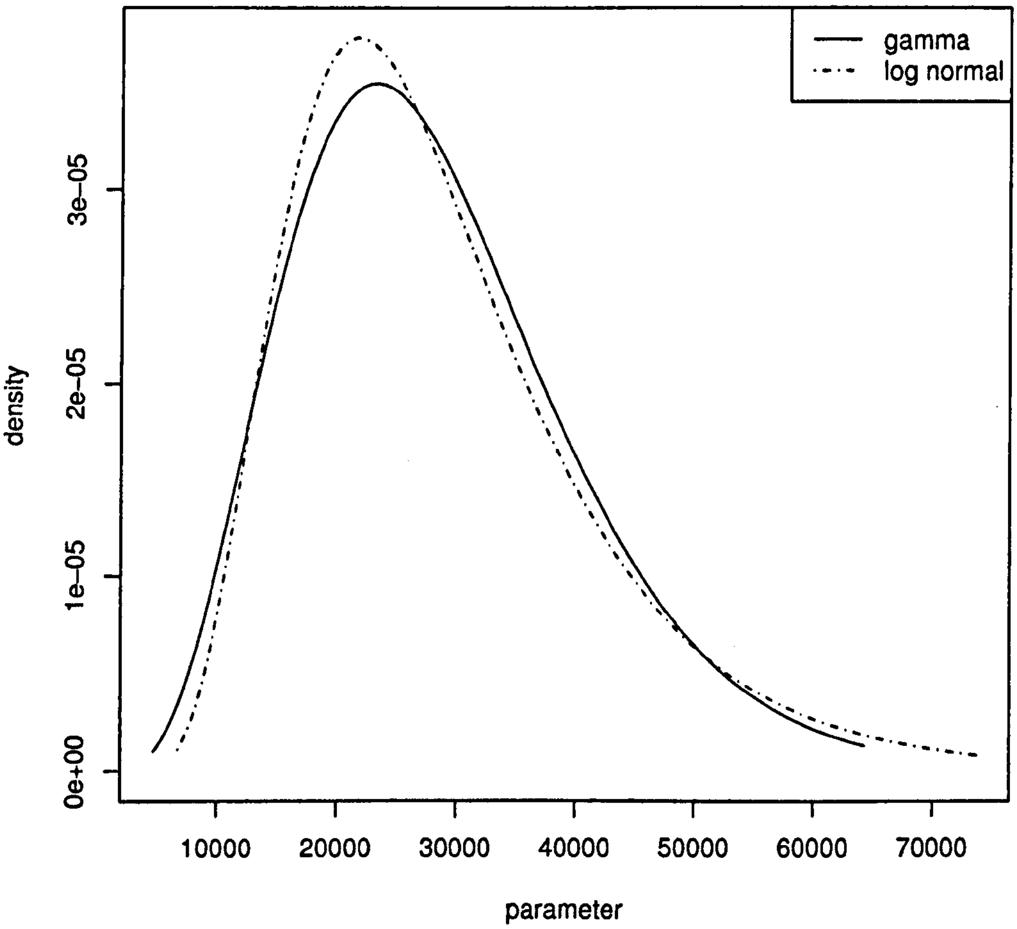


Figure 6.1: Plot comparing the fitted logNormal and Gamma distributions for M_Y

The expert felt that the plot of the logNormal distribution fitted by SHELF for M_Y did represent her prior beliefs although the elicited values for M_Y suggested negative skewness and her upper bound of £40,000 was the 0.827 quantile value of the fitted logNormal distribution which might not have adequately captured her upper bound as "a value that it is extremely unlikely" for M_Y to exceed.

For R_Y we want to fit a 1+Gamma distribution and when 1 was subtracted from the elicited values for R_Y SHELF fitted the Gamma distribution with scale parameter 5.924 and shape parameter 1.388.

It can be seen from the chart below which compares the two alternatives offered (for positively skewed distributions) in SHELF that these distributions do show some divergence.

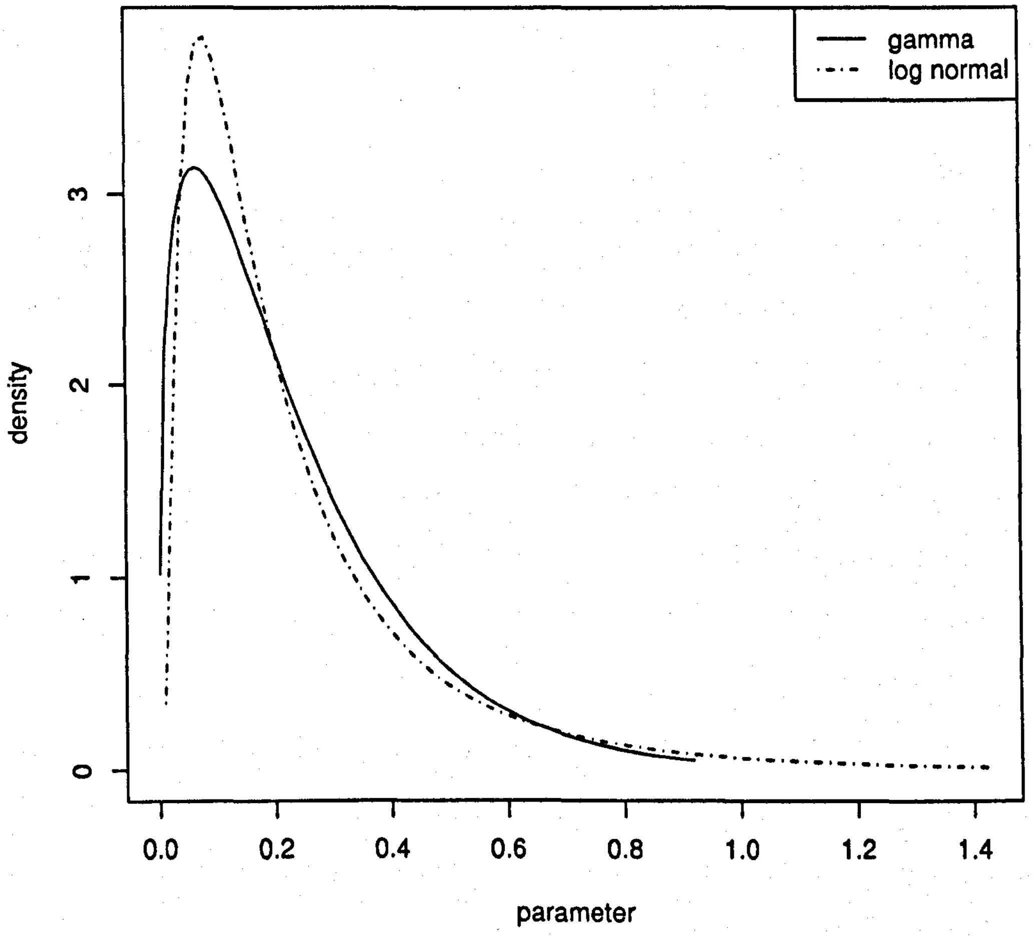


Figure 6.2: Plot comparing the fitted logNormal and Gamma distributions for R_Y

When the Excel 1+Gamma chart was shown to the expert she felt that the chart did represent her prior beliefs although the elicited values for R_Y suggested a symmetric distribution and her upper bound of 1.5 was the 0.901 quantile value of the fitted 1+Gamma distribution which once again might not have adequately captured her upper bound as "a value that it is extremely unlikely" for R_Y to exceed.

6.3.3 Comments on the first elicitation

Some practical points

1. It would be helpful to have a hand held calculator available (might be needed to calculate the plausible value of R_Y).

2. Additional note pads and pens to be provided.

3. Ensure that the elicitation is conducted with plenty of desk space. The PC to be kept to one side as it is only used in the fitting and feedback part and not in the main (time consuming) part of capturing the expert's prior beliefs.

4. The documentation that had to be completed from SHELF did seem potentially repetitive. Some changes to the documentation had been made to reflect the procedure that all the briefing to the expert took place at the meeting on a "face to face " basis. Further modifications were made to record all the elicited values on one piece of paper.

5. The facilitator had developed more user friendly visual feedback of the fitted distributions using three colour charts in Excel to give individual feedback for M_Y and R_Y plus a combined chart showing smaller versions simultaneously of the two individual charts for the question about independence.

When visual feedback for M_Y was given the chart did not produce the results desired but a small modification has corrected that situation.

6. After reflection on the elicitation and re-reading the SHELF documentation the facilitator decided to adopt a more proactive role for subsequent elicitations.

Changes to The elicitation in the Appendix

1. A definition of elicitation is now made on page 1 of The elicitation in the Appendix to ensure that the ideas that the process of elicitation will require are introduced at the start of the elicitation session.

2. At the first elicitation the expert was offered a choice of two page 6 charts to illustrate Normal distributions. The chart that is now included was chosen by the expert at the first elicitation.

3. Positively skewed distributions are introduced on page 12 with the fourth paragraph containing the key points. However, the sentence "The larger the value of the Quartile Ratio, the greater the positive skew of values of the distribution" may not convey the intended meaning. The words were therefore illustrated by reference to the BBC article http://news.bbc.co.uk/1/hi/magazine/7581120.stm that shows a chart for The UK Income Distribution in 2006/7.

However it was felt during this first elicitation that this needed strengthening even further and this was achieved by adding in what is now page 16 to illustrate the effect of changing the value of R_Y with the comments about the Quartile Ratio on page 13 restated.

As the value of σ^2, or $R_Y = \exp(0.6745\sigma)$, increases (with $M_Y = \exp(\mu)$ held fixed) the value at which the mode is attained reduces. However, the value of the probability density function for the modal value is a minimum when $\sigma^2 = 1$, or $R_Y = 1.963$, but this subtlety has not been made explicit in the chart on page 16 and the values shown of $R_Y = 1.5$ and 3 indicate the general relationship between R_Y and the shape of the probability density function.

Note that this strengthening was conducted immediately before the elicitation of what have now been described as specific values. Although they do represent the $\frac{1}{2}$-quantile value, $\frac{1}{3}$-quantile value and $\frac{2}{3}$-quantile values for R_Y there is no reason why they need to be referred to as such, which potentially confuses the expert.

When this first elicitation was conducted the lack of questions from the expert to the facilitator about the Quantile Ratio did concern the facilitator and added to the resolve to strengthen this part of the elicitation.

4. When elicitation for the Median was undertaken, now page 19, there was evidence from the expert's queries to the facilitator that she had forgotten the population property that the Median controlled. Hence what is now page 18 has been added in to strengthen the expert's understanding immediately before the elicitation of what again have been described as specific values.

6.3.4 The second elicitation

For this elicitation the expert was Mr Richard Gregory, BA ACA MBA, Head of Risk Control, Northern Rock, on 3rd April 2010.

The results that were obtained can be summarised as

values elicited	M_Y	T_Y	R_Y
upper bound	25	30	1.75
median	20	25	1.50
lower tertile	16	-	1.20
upper tertile	22	-	1.70

Table 6.3: Initial values elicited for M_Y, T_Y and R_Y

where the plausible value for R_Y was 1.25 and the entries for M_Y and T_Y represent thousands of pounds. There are logical lower bounds for M_Y and R_Y of 0 and 1 respectively.

As can be seen from Table 6.3 the value elicited for the upper tertile for R_Y was close to the value that had been elicited earlier for the upper bound. When this was discussed with the expert he felt that he had been too restrictive with his initial belief for the value of the upper bound and this was increased to 1.8.

The elicited values that were then used to fit the initial distribution were

values elicited	M_Y	T_Y	R_Y
upper bound	25	30	1.8
median	20	25	1.5
lower tertile	16	-	1.2
upper tertile	22	-	1.7

Table 6.4: Elicited values for M_Y, T_Y and R_Y used to fit an initial distribution

with SHELF fitting a logNormal distribution with mean $\mu = 9.845$ and variance $\sigma^2 = 0.260^2$ for M_Y.

It can be seen from the chart below that although we have chosen to fit a logNormal distribution the alternative offered (for positively skewed distributions) in SHELF, a Gamma distribution, is a comparable distribution.

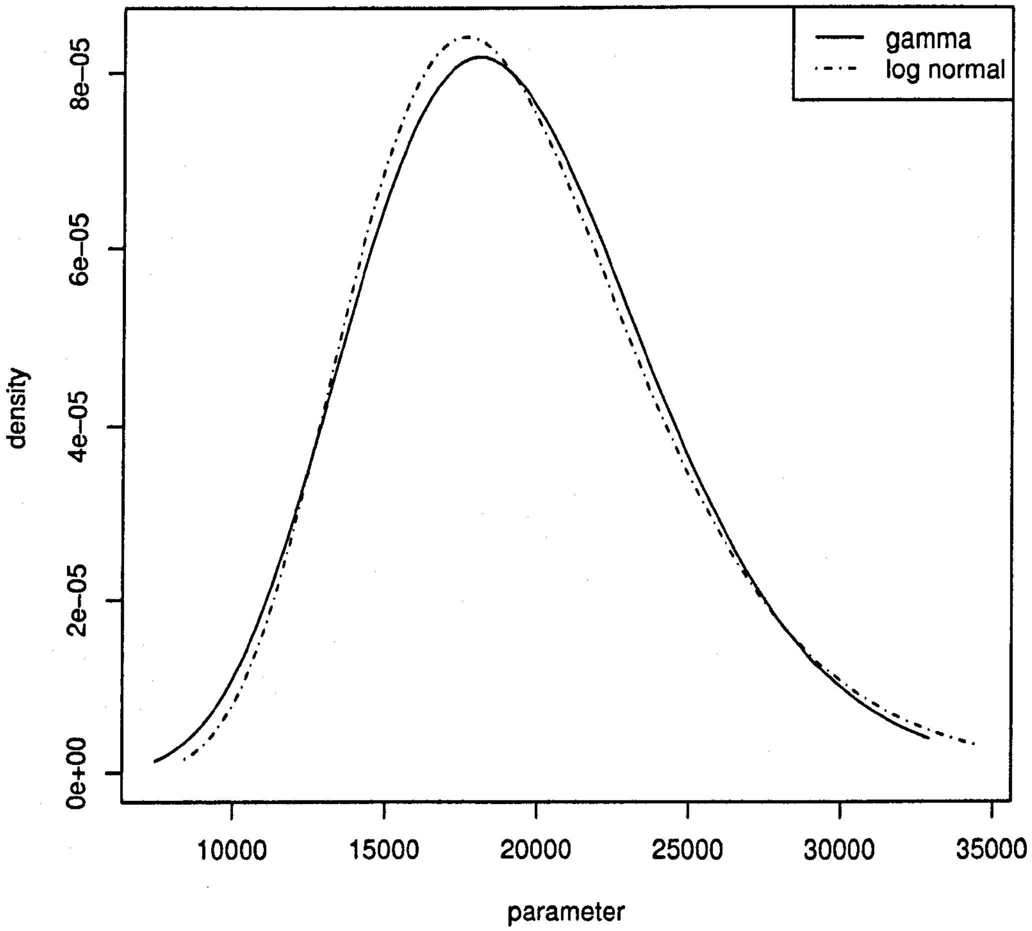


Figure 6.3: Plot comparing the fitted logNormal and Gamma distributions for M_Y

The expert felt that the plot of the logNormal distribution fitted by SHELF for M_Y did represent his prior beliefs although the elicited values for M_Y suggested negative skewness and his upper bound of £25,000 was the 0.860 quantile value of the fitted logNormal distribution which might not have adequately captured his upper bound as "a value that it is extremely unlikely" for M_Y to exceed.

For R_Y we want to fit a 1+Gamma distribution and when 1 was subtracted from the elicited values for R_Y SHELF fitted the Gamma distribution with scale parameter 2.690 and shape parameter 1.339.

It can be seen from the chart below which compares the two alternatives offered (for positively skewed distributions) in SHELF that these distributions do show some divergence

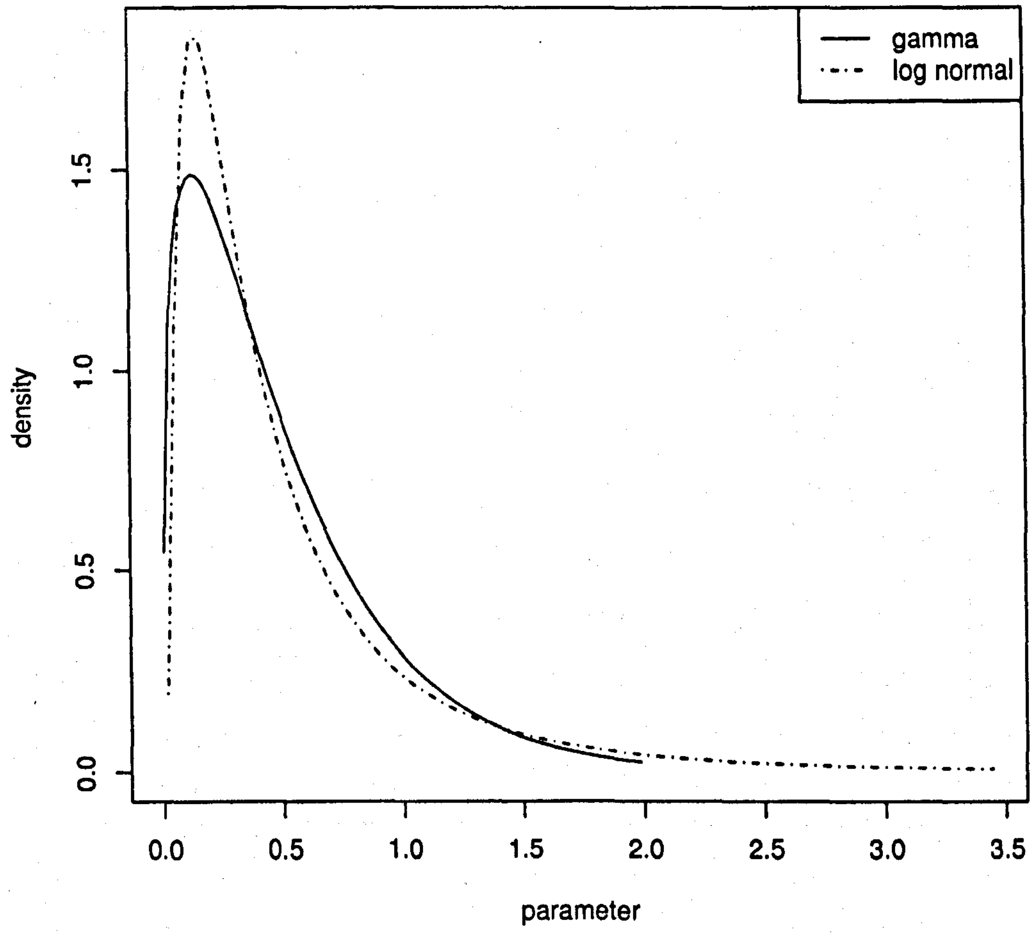


Figure 6.4: Plot comparing the fitted logNormal and Gamma distributions for R_Y

When the Excel 1+Gamma chart was shown to the expert this stimulated some discussion, particularly about the upper tail. He did not believe that the chart had captured the prior beliefs that he possessed and this was confirmed by examining the quantile fitted values that SHELF produces. It became apparent during this discussion that the expert felt that he possessed a more detailed understanding of salary distribution in the upper tail of Y.

The facilitator suggested that a way to obtain the increase in upper tail weight that the expert wanted was to increase the value of the upper bound to 1.9. SHELF then fitted the Gamma distribution with scale parameter 2.532 and shape parameter 1.310 as is shown in the chart below

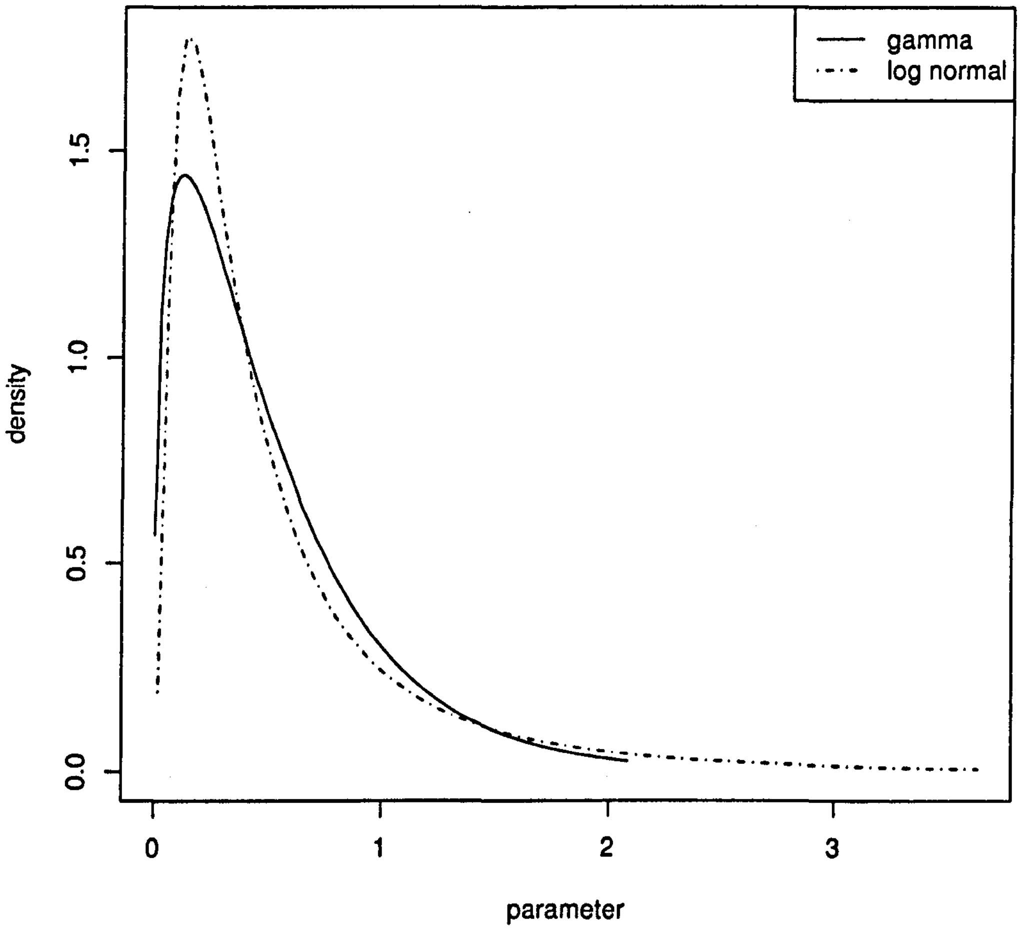


Figure 6.5: Plot comparing the fitted logNormal and Gamma distributions for R_Y

and again this required a visual examination of the revised Excel 1+Gamma chart as well as the quantile fitted values that SHELF produces before he was satisfied with this fitted distribution.

When the Excel 1+Gamma chart was shown to the expert he felt that the chart did represent his prior beliefs although the elicited values for R_Y suggested negative skewness and his upper bound of 1.9 was the 0.836 quantile value of the fitted 1+Gamma distribution which again might not have adequately captured his upper bound as "a value that it is extremely unlikely" for R_Y to exceed.

When the final question was posed for the expert, see page 20 of The elicitation in the Appendix, which is about the independence of the Median and Quartile Ratio the expert at first felt that for a small value of the Median then the value of the Quartile Ratio would be large. This reflected the way that he tended to think of the elicited values of the Quartile Ratio as the Third Quartile value divided by the Second Quartile value.

For a workforce of c.4000, as employed during the early part of 2010, he was thinking about the ratio of the salary of the 3000th ranked employee (determined when counting down) to the 2000th ranked employee (when counting up) where all Northern Rock employees were considered in this elicitation (including part-time, evening and weekend only employees) which gave the workforce a large, relatively lowly paid, clerical slew.

The facilitator refreshed the expert's knowledge of the Median as a value that was measured in pounds whereas the Quartile Ratio was a dimensionless number. After a break to allow the expert to reflect he decided that he did not feel that particular values of the Median would want him to change his beliefs about the Quartile Ratio and so implicitly accepted the independence of the Median and the Quartile Ratio.

6.3.5 Comments after both elicitations

The following observations were noted

1. All four upper bounds were too low, or restrictive, and the facilitator should have discussed this with the experts.

2. The facilitator was conducting his first and second elicitations here and felt, as was indeed the case, that he was progressing along a very steep learning curve. Although, when conducting the second elicitation, he became more proactive in his interaction with the expert it is evident that he should, within his level of expertise, increase his proactive involvement for any subsequent elicitations.

3. None of the four elicitations produced values that represented positively skewed distributions. The SHELF package followed the choice made by the facilitator to fit positively skewed distributions and produced the best fit that it could make. As SHELF fits both a logNormal and Gamma distribution if the "positively skewed" option has been chosen then "the sum of squared differences between the elicited probabilities and fitted probabilities" are shown for both distributions. While this may be used to decide which of these two distributions is more relevant the values do also indicate how good is the "absolute" accuracy of fit.

4. The experts agreed four unimodal distributions as representing their prior beliefs and probably these unimodal distributions were themselves within the experts' "comfort zone".

5. None of the four elicitations produced values that represented positively skewed distributions although in each case the expert did agree that positively skewed distributions represented their prior beliefs. This apparent "conflict" could be attributed to the experts' training, their inexperience in thinking about elicited values or the facilitator's lack of expertise in conducting elicitations or perhaps some combination of these factors.

The following changes to The elicitation are proposed for the future

1. In The elicitation in the Appendix there are three references to "determine the largest value that you believe possible", for the Median and Third Quartile on page 15 and also the Quantile Ratio on page 17. These pages of The elicitation would be amended by inserting the additional part "determine the smallest value that you think possible" for

M_Y on page 15, before (a)
T_Y on page 15, between (b) and (c),
R_Y on page 17, before (a).

2. In The elicitation in the Appendix there are two places where we will use SHELF to fit a positively skewed distribution to the elicited values for the Quantile Ratio (on page 17) and the Median (on page 19). These pages of The elicitation would be amended by inserting the additional part shown below, for R_Y on page 17 between (c) and (d) and for M_Y on page 19 between (b) and (c)

"when the five elicited values have been determined by the expert, the facilitator will give an informal feedback by producing a sketch to indicate the shape of the distribution that has been suggested by the elicited values".

In particular the sketch will show whether the shape is indicated as being negatively, symmetric or positively skewed and, from the smallest and largest values, how quickly the shape decreases to zero. It is not unreasonable to expect the distribution of prior beliefs for quantities that do not take negative values (M_Y and R_Y - 1) to be positively skewed and if an expert's elicited values suggested a different shape then this would be discussed with the expert before formally using SHELF to fit a positively skewed distribution.

6.3.6 The 3CPO Study : the Sheffield data sets

The study was able to recruit 1069 patients, from July 2003 until the end of April 2007, over its 26 sites with complete data available for 1052 patients. The Northern General Hospital, Sheffield was ranked the second largest site with 134 patients recruited with complete data. As the facilitator had hoped to conduct an elicitation in Sheffield, preliminary analysis was undertaken on the total costs presented for the 134 patients recruited in Sheffield as well as the 1052 for the study as a whole.

Costs, which commenced once the patient had presented at an Emergency Department, were collected according to a simple additive model for each patient with a "fixed" component representing the cost of the treatment arm to which the patient had been allocated and a "variable" component representing the cost of the time spent in the different type(s) of possible treatment. The same unit costs were used across the study.

Summary statistics for per patient total cost are presented below for the three arms of the 3CPO study as a whole (namely Standard, CPAP and NIPPV) in the style of Briggs *et al* (2005), where for a Normal distribution Skewness = 0 and Kurtosis = 3, as

	Standard	CPAP	NIPPV
n	364	337	351
Mean	3715	4085	4325
Sd	3497	3790	3917
Skewness	3.23	2.80	2.49
Kurtosis	20.14	14.39	10.65
CoV	0.94	0.93	0.91

Table 6.5: 3CPO study : cost summary statistics

and, for the patients recruited in Sheffield

	Standard	CPAP	NIPPV
n	45	42	47
Mean	4006	4381	4646
Sd	2790	3348	4634
Skewness	1.66	1.12	1.98
Kurtosis	6.57	3.18	6.23
CoV	0.70	0.76	1.00
Second quartile	3398	3375	3075
Third quartile	5488	5370	5155
Plausible r	1.62	1.59	1.68

Table 6.6: Sheffield study : cost summary statistics

where the sample second quartile value is the $2\text{x}(\frac{n+1}{4})^{th}$ ranked sample value and the sample third quartile value is the $3\text{x}(\frac{n+1}{4})^{th}$ ranked sample value of the ordered observations with the plausible sample quartile value the ratio of these quartile values. If the rank number is non-integer then if the decimal part is 0.5 take the average of the two ranked sample values immediately above and below, otherwise round to the nearest rank number.

We can observe that the CoV for the Standard and CPAP data sets are 0.70 and 0.76 respectively.

While the sample numbers are much smaller for the Sheffield study compared to the 3CPO study as a whole there do appear to be some differences between these two sets of data. In particular, the mean values are higher for Sheffield but Prof Dixon was unable to offer an explanation for this observation.

The 3CPO costs were measured in £ but to avoid unnecessary repetition the £ has been suppressed throughout this Section.

If we now look at the summary statistics for the log of cost for the patients recruited in Sheffield we have

	Standard	CPAP	NIPPV
n	45	42	47
Mean	8.07	8.12	8.07
Sd	0.72	0.76	0.86
Skewness	-0.52	-0.05	0.19
Kurtosis	3.70	2.79	2.94
CoV	0.09	0.09	0.11

Table 6.7: Sheffield study : log cost summary statistics

where for a Normal distribution Skewness = 0 and Kurtosis = 3.

We can observe that the log of cost for the Standard data set exhibits negative skewness with the CPAP data set showing this marginally.

We will now present numerical results for the comparisons of comparable ranges of candidate models (to those presented in Section 2.8 for the pMDI+ data set) for the Standard, CPAP and NIPPV data sets collected for those patients recruited in Sheffield.

The first plot in Figure 6.6 shows the fBf for comparing the rootNormal vs logNormal models for the Standard data set where the value of the fBf $\rightarrow 1$ as $\lambda \rightarrow 0$, although the value of the fBf is above 1 for values of $1/\lambda \geq 3$ with the peak fBf value of 2.29 occurring when $1/\lambda = 5$. Hence a logNormal model was favoured over a rootNormal model for values of $1/\lambda < 3$, whereas a rootNormal model was very weakly favoured over a logNormal model for values of $1/\lambda \geq 3$.

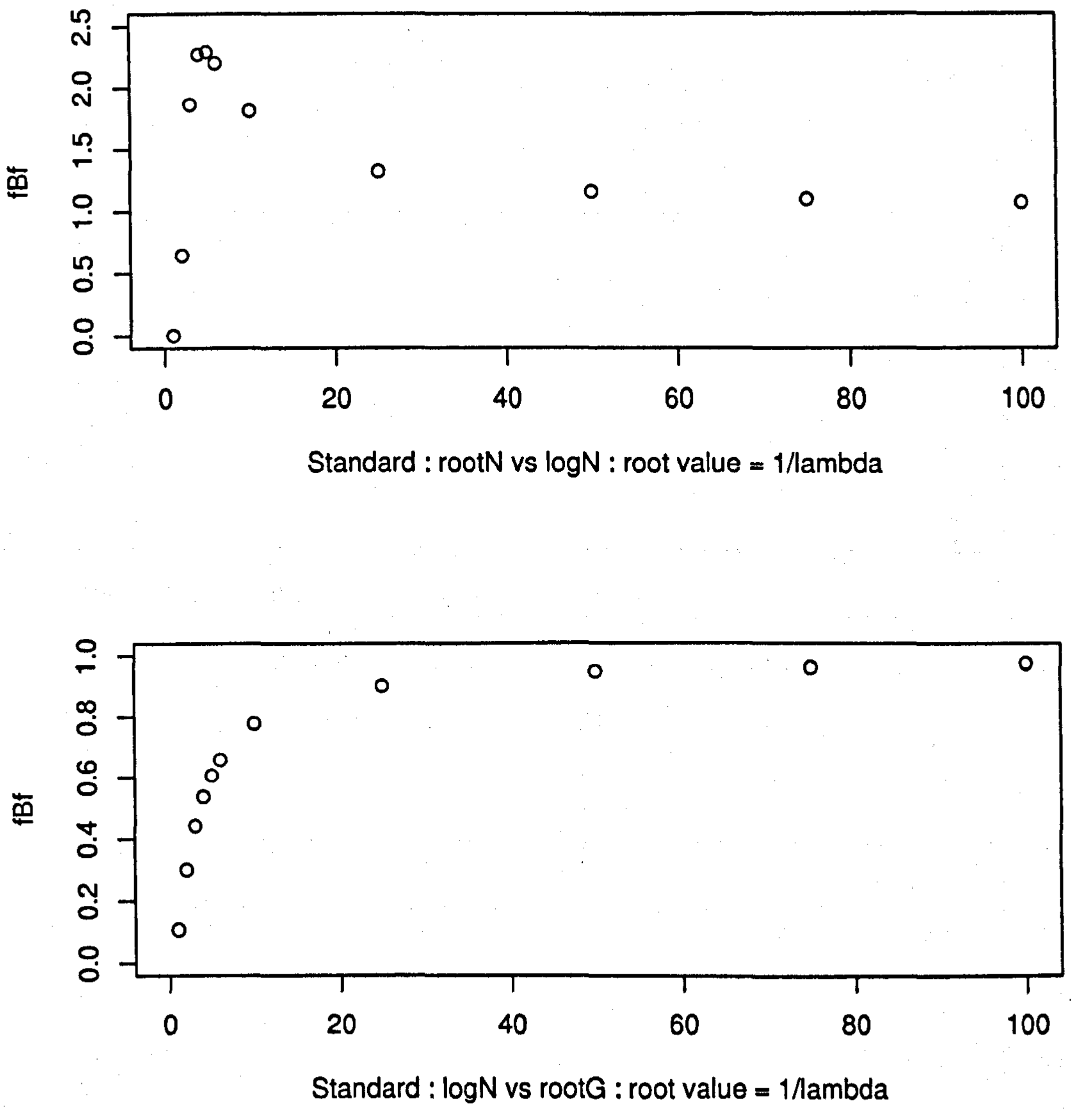
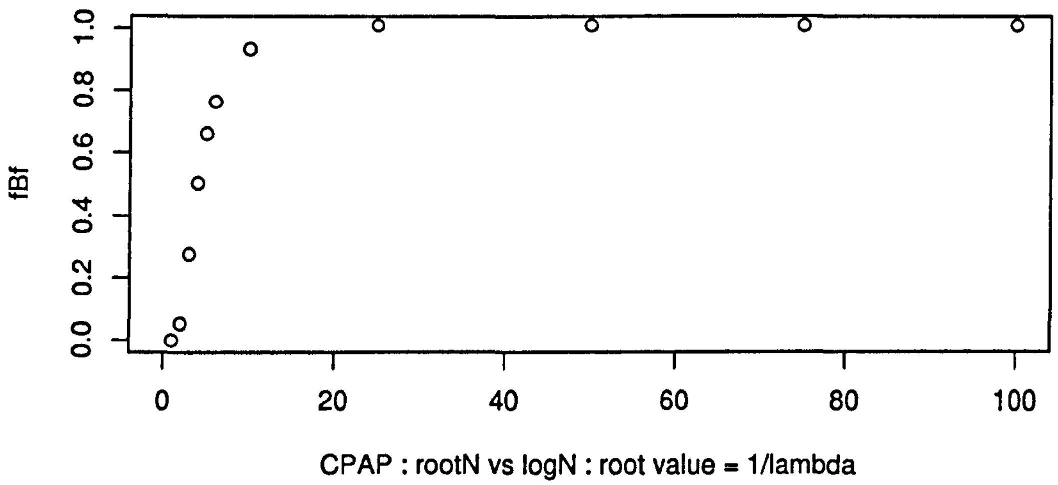


Figure 6.6: Plots comparing candidate models for the Standard data set

The second plot in Figure 6.6 shows the fBf for comparing the logNormal vs rootGamma models for the Standard data set where the value of the fBf → 1 as $\lambda \rightarrow 0$ and the preference for rootGamma models is strong for values of $1/\lambda$ close to 1 but quickly decreases. As $\lambda \rightarrow 0$ then the fBf could barely distinguish between the logNormal and the other models.

The first plot in Figure 6.7 shows the fBf for comparing the rootNormal vs logNormal models for the CPAP data set where the value of the fBf → 1 as $\lambda \to 0$.



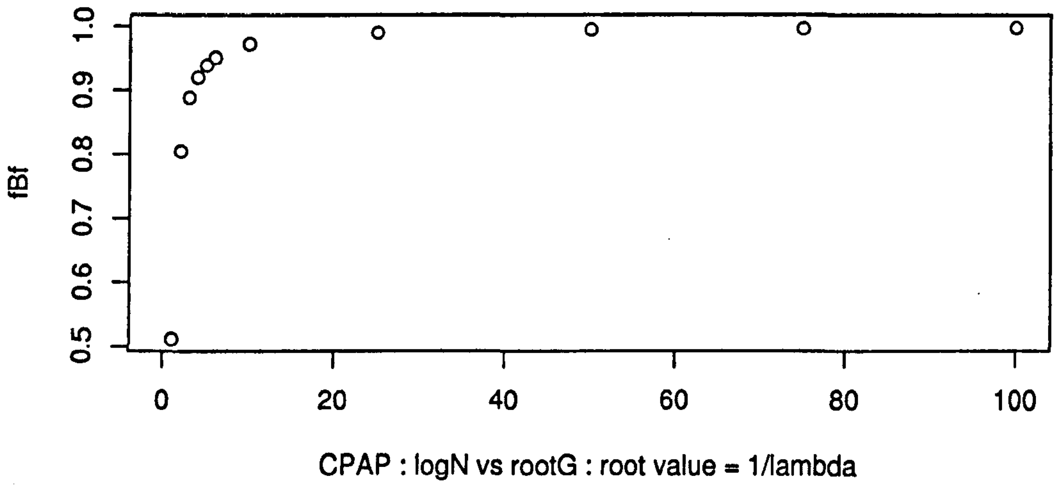


Figure 6.7: Plots comparing candidate models for the CPAP data set

The second plot in Figure 6.7 shows the fBf for comparing the logNormal vs rootGamma models for the CPAP data set where the value of the fBf → 1 as $\lambda \to 0$ and there is a weak preference for rootGamma models for values of $1/\lambda$ close to 1. However, as $\lambda \to 0$, then the fBf could barely distinguish between the logNormal and the other models.

The first plot in Figure 6.8 shows the fBf for comparing the rootNormal vs logNormal models for the NIPPV data set where the value of the fBf → 1 as $\lambda \rightarrow 0$.

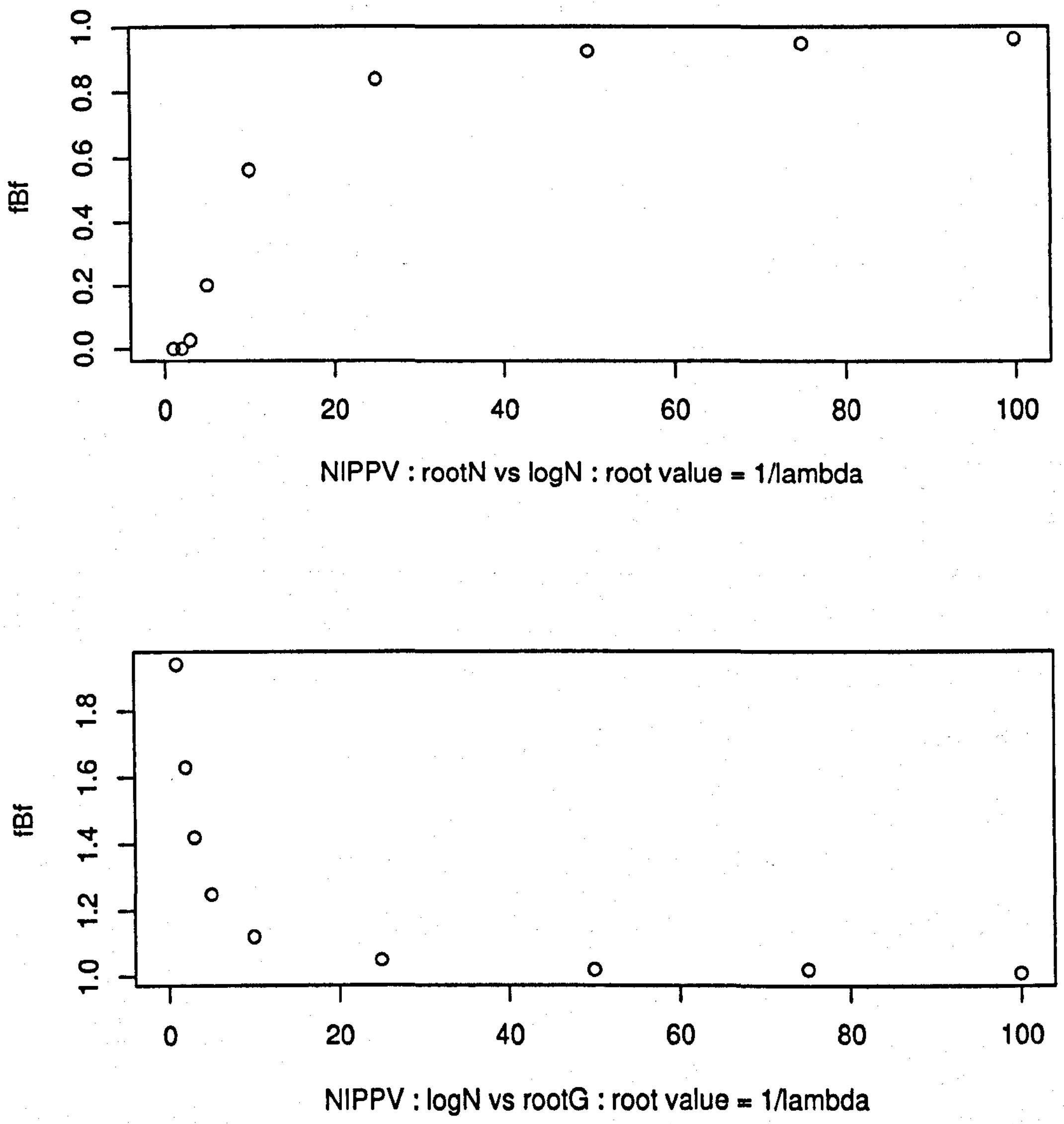


Figure 6.8: Plots comparing candidate models for the NIPPV data set

The second plot in Figure 6.8 shows the fBf for comparing the logNormal vs rootGamma models for the NIPPV data set where the value of the fBf → 1 as $\lambda \rightarrow 0$.

The logNormal was the favoured model over rootNormal and also rootGamma models for the NIPPV data set. However, as $\lambda \rightarrow 0$, then the fBf could barely distinguish between the logNormal and the other models.

However, for the Standard and CPAP data sets Gamma was the favoured model.

If we produce an amended version of Table 5.6 as Table 6.8 below, using a value of CoV = 0.75 as representative of the Standard and CPAP data set values

		sm		exp(lm + lv/2)		Bpe	
Distribution	CoV	RMSE	average	RMSE	average	RMSE	average
Gamma	0.25	56	1000	56	1002	61	1019
Gamma	0.75	168	1000	225	1088	194	1064
Gamma	2.00	448	1001	4.89E+45	4.89E+45	553	668

Table 6.8: Estimated root mean square error for sample size 20 for Gamma with different estimators using 1000K replications

then we can see that when using the RMSE to compare our Bpe, which uses a logNormal data model, with the other estimators our Bayesian model is reasonably robust against the misapplication to a Gamma distribution with CoV = 0.75.

The histograms for log of cost shown on the next page as Figure 6.9 do indicate that logNormal distributions appears to be not unreasonable data models.

We will therefore continue with our logNormal data model for all three trial arms.

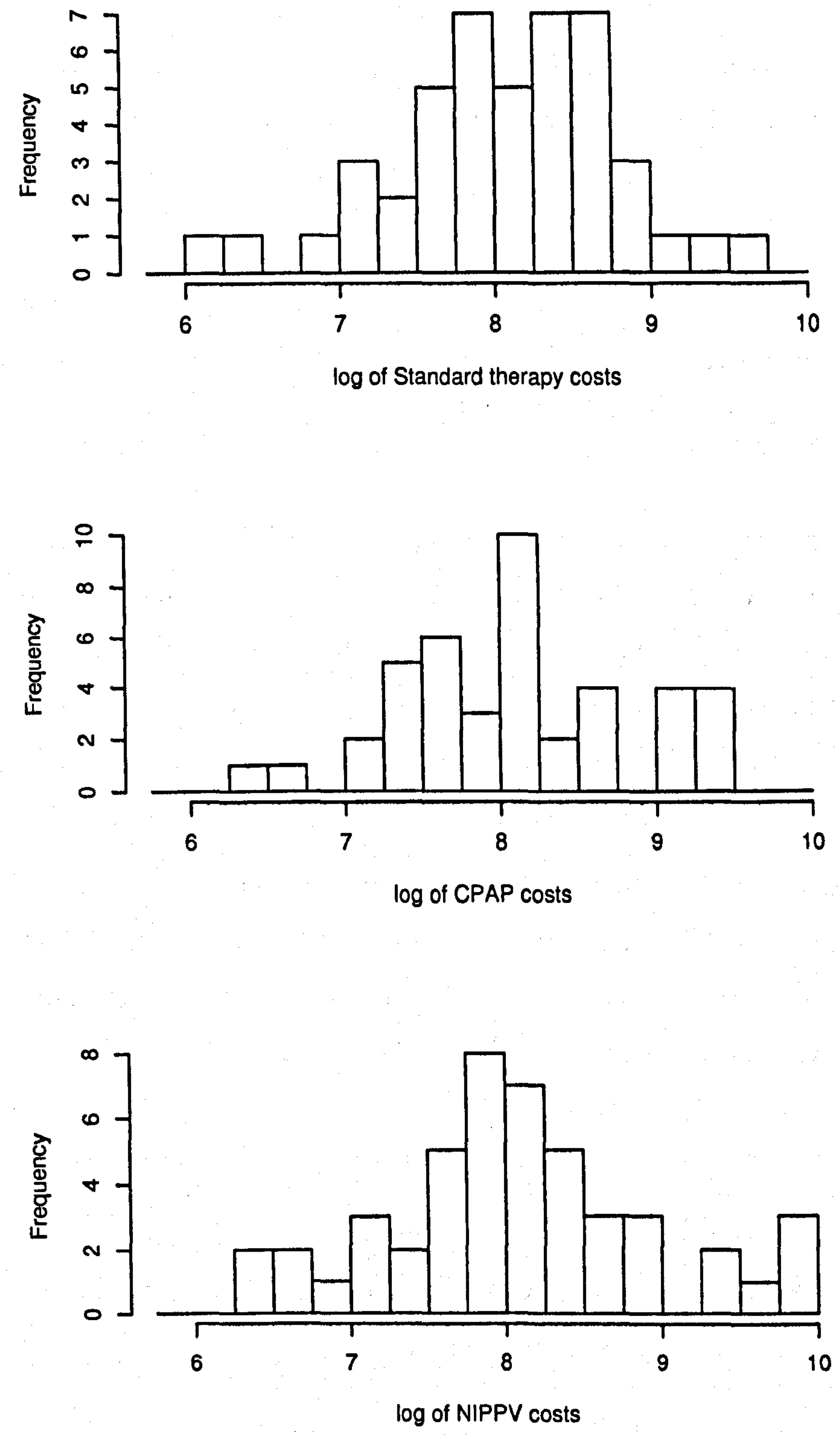


Figure 6.9: Histograms showing the distribution of log cost for the Sheffield study

Although we have been unable to elicit prior beliefs for any recruitment centre in the 3CPO study data set it is of interest to postulate what the outcome might have been if this had been possible for Sheffield and the rest of this chapter is devoted to that analysis and we will suppress the use of "Sheffield" when referring to any of the legs of 3CPO.

We will commence by establishing the possible default priors may have been elicited from an expert for the Standard therapy and will then consider alternative prior beliefs for the other two legs of 3CPO.

In Chapter 5 we have examined the application of the Bayesian model that has been developed to available data sets. In Section 5.2.3 we have examined the choice of prior beliefs for the median. Table 5.3 compares a number of options when M_Y follows a logNormal distribution and concludes that the estimated RMSE is reasonably robust to the choice of parameter values when M_Y follows a $\text{logN}(\mu, \sigma^2)$ distribution over a range of values that we might expect to encounter.

For the two elicitations that have been reported earlier in this chapter SHELF was used to fit logNormal distributions to M_Y which gave mean $\mu = 10.181$ and variance $\sigma^2 = 0.441^2$ and mean $\mu = 9.845$ and variance $\sigma^2 = 0.260^2$ respectively. We can observe that while the two values of μ are in the range that have been considered in Chapter 5 the variance is much smaller.

To examine a range of values for μ and σ^2 we use the Standard therapy data set with prior beliefs of $R_Y \sim 1+\text{G}(11.5,6)$ to predict the mean value for one unobserved member of this cost population as

μ	σ^2		
	1	10^2	100^2
15	4544	4176	4173.00
8	4169	4173	4172.97
0	3809	4169	4172.93
-8	3482	4165	4172.89
-15	3195	4162	4172.85

Table 6.9: Sheffield Standard therapy : predicted mean cost

where we can observe that, for a fixed value of σ^2, as the value of μ increases then the predicted mean value also increases. When the value of σ^2 is small then the size of μ strongly influences the predicted mean value.

We can observe from Table 6.9 that for $\mu = 8$, a value almost exactly equal to the sample mean of the log of cost, the predicted mean cost is, as expected, robust against the value of σ^2 used, over the range shown.

To examine the influence of smaller values of σ^2 in more detail we use the Standard therapy data set with prior beliefs of $R_Y \sim 1+\text{G}(11.5,6)$ to predict the mean value for one unobserved member of this cost population as

μ	σ^2									
	0.1^2	0.2^2	0.3^2	0.4^2	0.5^2	0.6^2	0.7^2	0.8^2	0.9^2	1^2
10	46743	9109	5468	4824	4573	4447	4371	4324	4292	4269
8	4006	4101	4135	4151	4158	4163	4165	4167	4168	4169
0	292	436	769	1425	2827	3245	3475	3625	3731	3809

Table 6.10: Sheffield Standard therapy : predicted mean cost for small σ^2

where we have taken values of μ as 8 and also a value for μ above and below 8.

We can see that when $\mu = 8$ then the predicted mean cost is extremely robust against σ^2 down to a value of $\sigma^2 = 0.7^2$, very robust down to $\sigma^2 = 0.3^2$ and even reasonably robust when $\sigma^2 = 0.1^2$.

To choose default prior distributions for M_Y and R_Y that are located around reasonable values we will use their mean values to measure location because it always exists for our choice of prior distributions and to measure the spread of values that we wish to choose we will use their standard deviation to measure spread or uncertainty of our prior beliefs.

As we have been unable to elicit prior beliefs from an expert to conduct this analysis we have chosen prior beliefs for M_Y that are reasonably robust against the choice of the uncertainty parameter value and if we work with $M_Y \sim$ G(8,10^2) as the default prior for our analysis of the 3CPO data sets it expresses reasonable uncertainty while being centred around plausible values for M_Y. We want the different data sets to determine different posterior mean values for one unobserved member of this cost population, where appropriate.

In Section 5.4 we have examined the choice of prior beliefs for the quantile ratio to develop the default prior $R_Y \sim$ 1+G(11.5,6) which minimised the TRMSE and yields a mean of 1.522 and sd of 0.213 for R_Y.

It is of interest to compare what has been proposed as the default prior with the two distributions that were elicited earlier. For R_Y we want to fit a 1+Gamma distribution and when 1 was subtracted from the elicited values for R_Y for the first elicitation SHELF fitted the Gamma distribution with scale parameter 5.924 and shape parameter 1.388 which yields a mean of 1.234 and sd of 0.199 for R_Y. For the second elicitation SHELF fitted the Gamma distribution with scale parameter 2.532 and shape parameter 1.310 which yields a mean of 1.517 and sd of 0.452 for R_Y.

Hence while the mean values are different, and there is certainly no reason to expect them to be (even) approximately equal, the sd from the second elicitation is larger which reflects the greater uncertainty of the expert's beliefs about R_Y.

If we now apply the three prior beliefs for R_Y to the Standard data set for Sheffield using $M_Y \sim \text{logN}(8,10^2)$ as the default prior for M_Y then the expected value is 4173 for the default prior of $R_Y \sim 1+\text{G}(11.5,6\ )$ with 4143 from the first elicitation and 4211 from the second elicitation. So the values are approximately equal and we will proceed with $R_Y \sim 1+\text{G}(11.5,6)$ as the default prior.

To simulate the possible results from an elicitation we will start by formulating the expert's prior views about the Standard therapy costs as the default priors $M_Y \sim \text{logN}(8,10^2)$ and $R_Y \sim 1+\text{G}(11.5,6)$, with skewness of $2(\text{shape})^{-\frac{1}{2}} = 0.82$ and uncertainty or sd $= 0.213$ for R_Y, to find the posterior mean value for one unobserved member of this cost population when the same prior beliefs are applied to all three arms of the 3CPO study as

therapy type	posterior mean value	sample mean
Standard	**4173**	4006
CPAP	4487	4381
NIPPV	4615	4646

Table 6.11: Sheffield study : mean values

where we will make comparisons with the posterior mean value for an unobserved Standard therapy member.

To examine alternative prior beliefs for $R_Y \sim 1+\text{G}$ we will allow the skewness to increase to 1.41 or to reduce to 0.47 while the uncertainty may increase to 0.43 or to reduce to 0.11, which we represent as $R_Y \sim 1+\text{G}(\text{scale},\text{shape})$ with the nine pairs of parameter values shown in Table 6.12 below

	uncertainty		
skewness	reduce	default	increase
increase	(13,2)	(6.7,2)	(3.3,2)
default	(23,6)	(11.5,6)	(5.7,6)
reduce	(40,18)	(20,18)	(9.8,18)

Table 6.12: G(scale,shape) parameter values

When $M_Y \sim$ logN(8,10^2) and the nine possible prior beliefs for $R_Y \sim$ 1+G are applied to the CPAP data set then Table 6.13 below shows the posterior mean value for one unobserved member

	uncertainty		
skewness	reduce	default	increase
increase	4334	4460	4554
default	4282	4487	4663
reduce	4315	4724	5291

Table 6.13: Sheffield study : CPAP , sample mean 4381

where the posterior mean value for an unobserved Standard therapy member was 4173. Hence for all of the range of prior beliefs examined here the posterior mean value for CPAP was higher than that for the Standard therapy.

When $M_Y \sim$ logN(8,10^2) and the nine possible prior beliefs for $R_Y \sim$ 1+G are applied to the NIPPV data set then Table 6.14 below shows the posterior mean value for one unobserved member

	uncertainty		
skewness	reduce	default	increase
increase	4435	4606	4737
default	4350	4615	4848
reduce	4352	4834	5490

Table 6.14: Sheffield study : NIPPV

and for all of the range of prior beliefs examined here the posterior mean value for NIPPV was higher than that for the Standard therapy.

It is possible to find a prior belief for R_Y that will produce posterior mean values that are less than the posterior mean value for the Standard therapy. For $R_Y \sim$ 1+G(187.5,56.25), yielding skewness of $2(\text{shape})^{-\frac{1}{2}} = 0.27$ and uncertainty of sd = 0.04 for R_Y, we find that

Therapy type	Expected value
CPAP	3916
NIPPV	3861

Table 6.15: Sheffield study : R_Y mean=1.300 & sd=0.040

but for this case there are very strong prior beliefs expressed for R_Y, which with very little positive skew is approximately Normal while R_Y is centred around 1.3.

So, apart from exceptional prior beliefs, such as those above, our expert will still conclude that the Standard therapy will have the lowest posterior mean value of the three therapies compared.

Chapter 7

Discussion

The problem that we have considered in this thesis is how to forecast the cost of treating an unobserved member of a population for some medical intervention. As this cost would be used to produce a budget we will only work to establish a point estimate - although a credible interval, either equal-tailed (which is available as part of the WinBUGS summary statistics of the posterior distribution) or highest posterior density, may be of interest in some non-financial scenarios.

For our Bayesian analysis we want to incorporate an expert's prior beliefs with an appropriate data model to produce the posterior expected mean value which would then be used as the forecast cost value for one unobserved member of the population.

In Chapter 2 we established using Bayes factors that, for the pMDI+ cost data set, a logNormal distribution was the best choice. However, in Chapter 5 when considering the fBf for the Paramedics data set there was a region of very weak preference against the logNormal distribution whilst in Chapter 6 a Gamma distribution was favoured for the Standard and CPAP data sets. Whilst all the observed data sets were "noisy" the only preferences against the logNormal data model arose from those data sets whose log distribution possessed a negative skewness.

We have approached authors of published papers for access to other cost data sets but, amongst other reasons, lack of permission to release the data, has not made this possible.

Parametric distributions that possess a heavier tail than the logNormal were considered in Chapter 5 while there is a reference to a practical problem where a Weibull distribution is considered to be an appropriate model in Oakley & Clough (2010). It would be worthwhile developing the Bayes factor methodology for other data models to enable wider comparisons to be undertaken.

The concepts behind forecasting future members of a finite population were introduced in Chapter 3 which led to the problem of interpreting the outcome from WinBUGS when using customary noninformative priors. Discussions with Prof Roberts and Prof Forster indicate that it is not necessarily straight forward to determine when the results from WinBUGS, which are necessarily finite, do indicate values that are infinite. This is particularly true for logNormal data that has a small shape parameter although WinBUGS does come with a warning that an understanding of the theory behind Bayesian statistics is required before using WinBUGS.

Having decided that the logNormal distribution was the best data model for the pMDI+ cost data set the rest of this thesis was devoted to establishing models for the joint prior belief for the logNormal parameters that satisfied three criteria

1. we can show analytically the existence of the posterior predictive moments
2. we can determine the value of the posterior predictive mean
3. we can elicit an expert's prior beliefs.

In Chapter 4 we developed a way to model prior beliefs for the logNormal data model that ensured posterior predictive moments were finite. In particular a novel way to model the prior belief for the shape parameter was proposed when numerical integration was available to speed the computation for particular choices for the prior distributions. To enable comparisons with classical estimators to be made in Chapter 5 default priors were determined. The parameter values for the shape default prior were trained by the parametric distributions and the values of the CoV that were used and also the Root Mean Square Error as the loss function. The choice of parametric distributions and the values of the CoV used do represent the typical range that might be encountered and weighting each result equally, aka the Laplace Criteria of Uncertainty, does produce results that have been shown to be robust, in the sense of the value of the RMSE, against misapplication of the data model.

The choice of RMSE was continued from the Briggs paper but may not be the most appropriate choice. Further research with end users may establish that some asymmetric form of loss function, not necessarily involving Squared Errors could be more relevant to the practical situation under investigation.

Data sets that arise in practical situations do not necessarily follow parametric forms and the performance of the logNormal data model with the default priors was encouragingly robust, in the sense of the value of the RMSE, against any misapplication of the data model - particularly when data arose from distributions with heavier tails than the logNormal.

The methodology of conducting an Elicitation is outlined in Chapter 6 where SHELF has been used to obtain the best fit for the elicited values. Before SHELF was fortuitously made available we were examining the classical problem of fitting parameters analytically from elicited values, for a logNormal distribution for M_Y and a 1+Gamma distribution for R_Y. Garthwaite & O'Hagan (2000) develop a simple method for the logNormal distribution that only requires three quantile values to be elicited. We had been seeking to establish a comparable method for the Gamma distribution for R_Y - 1 and it would be of interest to establish whether this does exist.

Although the default prior was developed to enable comparisons to be made when it was not possible to elicit an expert's prior beliefs we have also produced, and revised after use, a method that can be followed to elicit an expert's prior beliefs for our Bayesian model.

The case studies reported in Chapter 6 only represent part of what we would like to have undertaken and we would want to try to conduct elicitations with financial experts from within the NHS for the 3CPO data sets.

Finally, to summarise this thesis, the problem that we have considered is how to forecast the cost of treating unobserved members of a population for some medical intervention. As a Bayesian analysis we have incorporated an expert's prior beliefs with an appropriate data model to produce the posterior expected mean value.

The Bayesian model that we have developed, a logNormal data model and the quantile ratio to model the shape prior beliefs, does perform better (see Table 5.6) when using the default prior, in the sense of the Root Mean Square Error, than either the sample mean or the exp(lm + lv/2) estimators when data is simulated from Gamma and logNormal distributions. This better performance has also been demonstrated to be true for the three observed data sets that were introduced in Briggs *et al* (2005), see Table 5.11.

A Gamma data model is the preferred choice of many Health Economists, which was derived from a private conversation with Dr Richard Grieve of the LSHTM and see also Willans & Kowgier (2008). Our Bayesian model uses a data model that has a heavier tail than the Gamma distribution and, as shown in Table 5.14, performs better than the sample mean or the exp(lm + lv/2) estimators when data is derived from a range of tail weights and skewness and is relatively insensitive to misapplication to data models other than logNormal.

Glossary

Gamma function

The Gamma function is defined as

$$\Gamma(b) = \int_0^\infty v^{b-1} \exp(-v) dv$$

where we will only work with $b > 0$.

The Gamma function along the positive real axis is a convex function that takes strictly positive values. It has a minimum at $\Gamma(1.4616) = 0.8856$ and its value approaches ∞ as $b \longrightarrow 0$ or ∞.

Important properties that we will use are

$$\Gamma(b+1) = b\Gamma(b) \quad , \qquad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \quad , \qquad \Gamma(1) = 1$$

and the Gauss multiplication theorem, which is defined when m is a positive integer as the finite product

$$\Gamma(b)\Gamma\left(b+\frac{1}{m}\right)\Gamma\left(b+\frac{2}{m}\right)\cdots\Gamma\left(b+\frac{m-1}{m}\right) = (2\pi)^{\frac{m-1}{2}} m^{\frac{1}{2}-mb}\Gamma(mb).$$

Gamma distribution

The random variable V is said to follow a Gamma distribution, which we denote by $V \sim \mathrm{G}(a,b)$ for $v > 0$ and a, $b > 0$, if

$$f_V(v) = \frac{a^b v^{b-1} \exp(-av)}{\Gamma(b)}.$$

The scale parameter for V is a with shape parameter b.

Inverse Gamma distribution

The random variable W is said to follow an Inverse Gamma distribution, which we denote by $W \sim \mathrm{IG}(a,b)$ for $w > 0$ and a, $b > 0$, if

$$f_W(w) = \frac{a^b w^{-(b+1)} \exp(-\frac{a}{w})}{\Gamma(b)}$$

derived from $1/W = V \sim \mathrm{G}(a,b)$.

Bayesian paradigm

Bayes' theorem allows us to combine our two sources of information about the parameters $\boldsymbol{\theta}$, namely prior beliefs and data, into a single source of information.

It is expressed in symbols as

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta}) \times f(\mathbf{y}|\boldsymbol{\theta})$$

or in words as

the posterior is proportional to the prior times the likelihood

where $p(\boldsymbol{\theta}|\mathbf{y})$ is called the posterior distribution for $\boldsymbol{\theta}$ after observing the data $\mathbf{y}$, $\pi(\boldsymbol{\theta})$ represents our prior beliefs about $\boldsymbol{\theta}$ before observing the data $\mathbf{y}$ and $f(\mathbf{y}|\boldsymbol{\theta})$ represents the likelihood, $L(\boldsymbol{\theta};\mathbf{y})$, because $L(\boldsymbol{\theta};\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})$.

Likelihood

We will work with the continuous random variable Y whose realised value will be the observation y with probability density function

$$f_Y(y|\boldsymbol{\theta}) = \frac{d}{dy}p\{Y \leq y|\Theta = \boldsymbol{\theta}\} \qquad \forall y$$

and, unless it is necessary to avoid confusion, we will use the usual abbreviation of $f_Y(y|\boldsymbol{\theta})$ as $f(y|\boldsymbol{\theta})$.

Whenever we have n observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ then we will assume that this data set is the realised values from the independent identically distributed random variables $(Y_1, Y_2, \ldots, Y_n)$.

We are then able to formulate the likelihood of $\mathbf{y}|\boldsymbol{\theta}$ as

$$f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}) = \prod f(y_i|\boldsymbol{\theta}).$$

Normal distribution

The random variable X is said to follow a Normal distribution, which we denote by $X \sim \mathrm{N}(\mu, \sigma^2)$ for $-\infty < x < \infty$ and $\sigma^2 > 0$, if

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

The scale parameter for X is σ^2 with location parameter μ.

logNormal distribution

The random variable Y is said to follow a logNormal distribution which we denote by $Y \sim \mathrm{logN}(\mu, \sigma^2)$, or alternatively $Y \sim \mathrm{logN}(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\mu, \sigma^2)$, for $y > 0$ and $\sigma^2 > 0$, if

$$f_Y(y) = (2\pi\sigma^2)^{-\frac{1}{2}} y^{-1} \exp\left[-\frac{(\log y - \mu)^2}{2\sigma^2}\right]$$

derived from $\log Y = X \sim \mathrm{N}(\mu, \sigma^2)$ where the scale parameter for Y is μ with shape parameter σ and the median of $Y|\boldsymbol{\theta}$ is $\exp(\mu)$.

Quantiles

The q-quantile of a random variable V is the value v_q such that for $q \in [0,1]$ $p\{V \leq v_q\} = q$.
If $Z \sim \mathrm{N}(0,1)$ then $\Phi(z) = p\{Z \leq z\}$.
If $\log Y = X \sim \mathrm{N}(\mu, \sigma^2)$ then $p\{Y \leq \exp(\mu + q\sigma)\} = \Phi(q)$ and hence the $\Phi(q)$-quantile of $Y|\boldsymbol{\theta}$ is $\exp(\mu + q\sigma)$.

Percentiles

The p-percentile of a random variable V is the value v_p such that for $p \in [0,100]$ $p\{V \leq v_p\} = p$.

Student's *t* distribution

The random variable X is said to follow a t distribution, which we denote by $X \sim t_\nu(\mu, \sigma^2)$ for $-\infty < x < \infty$, $\sigma^2 > 0$ and ν a positive integer, if

$$f_X(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi\sigma^2}} \left[1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right]^{-\frac{(\nu+1)}{2}}.$$

$\log t$ distribution

The random variable Y is said to follow a $\log t$ distribution which we denote by $Y \sim \log t_\nu(\mu, \sigma^2)$ for $y > 0$, $\sigma^2 > 0$ and ν a positive integer, if

$$f_Y(y) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi\sigma^2}} y^{-1} \left[1 + \frac{(\log y - \mu)^2}{\nu\sigma^2}\right]^{-\frac{(\nu+1)}{2}}$$

derived from $\log Y = X \sim t_\nu(\mu, \sigma^2)$.

References

Al-Eideh, BM, Al-Refau, ASA & Sbeith, WM (2004). Modelling the CPI using a lognormal diffusion process and implications on forecasting inflation. *IMA Journal of Management Mathematics* **15** (**1**), 39-51.

Ashby, D. Bayesian Methods. *Biostatistics in clinical trials* **2**. Wiley, Chichester.

Berger, JO & Bernado, JM (1989).
Estimating a Product of Means : Bayesian Analysis With Reference Priors. *Journal of the American Statistical Association* **84** (**405**), 200-207.

Berger, JO & Pericchi, LR (1996).
The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association* **91** (**433**), 109-122.

Bernado, JM (1979). Reference Posterior Distributions for Bayesian Inference (with Discussion). *Journal of the Royal Statistical Society Series* B **41**, 113-147.

Black, WC (1990). The CE Plane : A Graphic Representation of Cost-Effectiveness. *Medical Decision Making* **10**, 212-214.

Box, GEP & Cox, DR (1964). An analysis of transformations (with Discussion). *Journal of the Royal Statistical Society Series* B **26**, 211-252.

Box, GEP & Tiao, GC (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Massachusetts.

Briggs, A (1999). A Bayesian approach to stochastic cost-effectiveness analysis. *Health Economics* **8** (**3**), 257-261.

Briggs, A, Nixon, R, Dixon, S & Thompson, S (2005). Parametric modelling of cost data : some simulation evidence. *Health Economics* **14** (**4**), 421-428.

Cassel, C-M, Särndal, C-E & Wretman, JH (1993). *Foundations of Inference in Survey Sampling.* Krieger Publishing Company, Florida.

Chen, DG (2002). A Bayesian model with a bivariate normal-lognormal prior distribution and a nolinear mixed-effect model for a regional fish stock-recruitment meta-analysis. *Proceedings of the American Statistical Association.*

Cochran, WG (1939). The Use of the Analysis of Variance in Enumeration by Sampling. *Journal of the American Statistical Association* **34** (**207**), 492-510.

Congdon, P (2001). *Bayesian statistical modelling.* Wiley, Chichester.

Cook, ET (1913). *The Life of Florence Nightingale.* Archive CD Books Ltd.

De Groot, MH (1970). *Optimal Statistical Decisions.* McGraw-Hill, New York.

Drummond, MF & O'Brien, BJ (1993). Clinical importance, statistical significance and the assessment of the economic and quality-of-life outcomes. *Health Economics* **2** (**3**), 205-212.

FDA (2010). Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. Docket No 2006D-0191.

Fieller, EC (1954). Some problems in Interval Estimation. *Journal of the Royal Statistical Society Series* B **16**, 175-185.

Friedman, LM, Furberg, CD & DeMets, DL (1998). *Fundamentals of Clinical Trials.* Springer-Verlag, New York.

Garthwaite, PH, Kadane, JB & O'Hagan, A (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association* **100** (**470**), 680-700.

Garthwaite, PH & O'Hagan, A (2000). Quantifying expert opinion in the UK water industry : an experimental study. *The Statistician* **47** (**4**), 455-477.

Gelman, A, Carlin, JB, Stern, HS & Rubin, DB (1995). *Bayesian Data Analysis.* Chapman & Hall, London.

Graham, LD, Smith, SD & Dunlop, P (2005). Lognormal Distribution Provides an Optimum Representation of the Concrete Delivery and Placement Process. *Journal of Construction Engineering and Management* **131** (**2**), 230-238.

Grieve, R, Nixon, R, Thompson, SG & Normand, C (2005). Using multilevel models for assessing the variability of multinational resource use and cost data. *Health Economics* **14** (**2**), 185-196.

Jeffreys, H (1961). *Theory of Probability.* Oxford University Press, Oxford.

Kass, RE & Raftery, AE (1995). Bayes Factors.
Journal of the American Statistical Association **90** (**430**), 773-795.

Kass, RE & Wasserman, L (1996). The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association* **91** (**435**), 1343-1370.

Kawahara, H, Reese, ED, Kitayama, T, Sasaki, S & Suto, Y (2008).
Extracting Galaxy Cluster Gas Inhomogeneity from X-Ray Surface Brightness : A Statistical Approach and Application to Abell 3667.
The Astrophsyical Journal **687** (**2**), 936-950.

Keren, G (1991). Calibration and probability judgements : conceptual and methodological issues. *Acta Psychologica* **77**, 217-273.

Khan, HMR, Haq, MS & Provost, SB (2005). Bayesian prediction for the log-normal model under Type II censoring. *CAMS Report 0405-17.* Center for Applied Mathematics and Statistics, New Jersey Institute of Technology.

Körner, TW (1988). *Fourier Analysis.* University Press, Cambridge.

Korpalski, S, Bruce, E, Holden, L & Klonne, D (2005). Dislodgeable foliar residues are lognormally distributed for agricultural re-entry studies.
Journal of Exposure Science & Environmental Epidemiology **15** (**2**), 160-163.

Lempers, FB (1971). *Posterior Probabilities of Alternative Linear Models.* University Press, Rotterdam.

Liljas, B, Ståhl, E & Pauwels, RA (1997). Cost-effectiveness Analysis of a Dry Powder Inhaler (Turbuhaler®) versus a Pressurised Metered Dose Inhaler in Patients with Asthma. *Pharmacoeconomics* **12** (**2 Pt 2**), 267-277.

Longfellow, HW (1857). Santa Filomena. *The Atlantic Monthly.*
1 (**1**), 22-23. Theatlantic.com.

Malliavin, P (1995). *Integration and Probability* . Springer-Verlag, New York.

Manca, A, Rice, N, Sculpher, MJ & Briggs, AH (2005). Assessing generalisability by location in trial-based cost-effectiveness analysis : the use of multilevel models. *Health Economics* **14** (**5**), 471-485.

May, JH, Strum, DP & Vargas, LG (2000). Fitting the Lognormal Distribution to Surgical Procedure Times. *Decision Sciences* **31** (**1**), 129-148.

Mould, RF, Lederman, M, Tai, P & Wong, JKM (2002).
Methodology to predict long-term cancer survival from short-term data
using Tobacco Cancer Risk and Absolute Cancer Cure models.
Physics in Medicine and Biology **47** (**22**), 3893-3924.

Neti, PSV & Howell, RW (2006). Log Normal Distribution of Cellular Uptake of Radioactivity : Implications for Biological Responses to Radiopharmaceuticals. *The Journal of Nuclear Medicine* **47** (**6**), 1049-1058.

Nightingale, F (1859). *A contribution to the sanitary history of the British army during the late war with Russia.* Harrison and Sons.
Harvard University Library Open Collections Program.

Novotny, V & Drozd, P (2000).
The size distribution of conspecific populations : the peoples of New Guinea.
Proceedings of The Royal Society of London - Series B : Biological Sciences
267 (**1446**), 947-952.

Oakley, JE & Clough, HE (2010). Sensitivity analysis in microbial risk assessment : Vero-cytotoxigenic *E.coli* 0157 in farm-pasteurized milk. *The Oxford Handbook of Applied Bayesian Analysis*, 69-89. Oxford University Press, Oxford.

O'Brien, BJ, Drummond, MF, Labelle, RJ & Willan, A (1994). In Search of Power and Significance : Issues in the Design and Analysis of Stochastic Cost-Effectiveness Studies in Health Care *Health Care* **32** (**2**), 150-163.

O'Hagan, A (1991). Discussion on Posterior Bayes factors (by M. Aitkin). *Journal of the Royal Statistical Society Series* B **53** (**1**), 136.

O'Hagan, A (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society Series* B **57** (**1**), 99-138.

O'Hagan, A (1997). Properties of Intrinsic and Fractional Bayes Factors. *Test* **6** (**1**), 101-118.

O'Hagan, A, Buck, CE, Daneshkhah, A, Eiser, JR, Garthwaite, PH, Jenkinson, DJ, Oakley, JE & Rakow T (2006). *Uncertain Judgements : Eliciting Expert Probabilities.* Wiley, Chichester.

O'Hagan, A & Forster, J (2004). *Kendall's Advanced Theory of Statistics* Volume 2B *Bayesian Inference.* Arnold, London.

O'Hagan, A & Stevens, JW (2001). A framework for cost-effectiveness analysis from clinical trial data. *Health Economics* **10** (**4**), 302-315.

O'Hagan, A & Stevens, JW (2002). The probability of cost-effectiveness. *BMC Medical Research Methodology* **2** (**5**).

O'Hagan, A & Stevens, JW (2003). Assessing and comparing costs : how robust are the bootstrap and methods based on asymptotic normality ? *Health Economics* **12** (**1**), 33-49.

O'Hagan, A, Stevens, JW & Montmartin, J (2000). Inference for the Cost-Effectiveness Acceptability Curve and Cost-Effectiveness Ratio. *Pharmacoeconomics* **17** (4), 339-349.

O'Hagan, A, Stevens, JW & Montmartin, J (2001). Bayesian Cost-Effectiveness Analysis from Clinical Trial Data. *Statistics in Medicine* **20**, 733-753.

Padgett, WJ & Johnson, MP (1983).
Some Bayesian lower bounds on reliability in the lognormal distribution.
The Canadian Journal of Statistics **11** (2), 137-147.

Pauwels, RA, Hargreave, FE, Camus, P, Burkowski, M & Ståhl, E (1996).
A one-year comparison of Turbuhaler® versus pressurized metered-dose inhaler (pMDI) in asthmatic patients. *Chest* **110,** 53-577.

Peterson, CR & Miller, A (1964). Mode, median and mean as optimal strategies. *Journal of Experimental Psychology* **68,** 363-367.

Poskitt, DS (1987). Precision, Complexity and Bayesian Model Determination. *Journal of the Royal Statistical Society Series* B **49** (**2**), 199-208.

Schwarz, G (1978). Estimating the Dimension of a Model.
The Annals of Statistics **6** (**2**), 461-464.

Smith, AFM & Spiegelhalter, DJ (1980).
Bayes Factors and Choice Criterea for Linear Models.
Journal of the Royal Statistical Society Series B **42** (**2**), 213-220.

Spiegelhalter, DJ, Abrams, KR & Myles, JP (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* Wiley, Chichester.

Spiegelhalter, DJ, Best, NG, Carlin, BP & van der Linde, A (2002).
Bayesian measures of model complexity and fit (with discussion).
Journal of the Royal Statistical Society Series B **64** (4), 583-639.

Spiegelhalter, DJ & Smith, AFM (1982).
Bayes Factors for Linear and Log-linear Models with Vague Prior Information. *Journal of the Royal Statistical Society Series* B **44** (**3**), 377-387.

Steele, C (2008). Use of the lognormal distribution for the coefficients of friction and wear. *Reliability engineering & systems safety* **93** (**10**), 1574-1576.

Tessella (2009/2011). The Opportunities and Advantages of Using Bayesian Statistics in Clinical Trials. *www.tessella.com* Capability statement.

Van Hout, BA, Al, MJ, Gordon GS & Rutten, FFH (1994). Costs, effects and C/E ratios alongside a clinical trial. *Health Economics* **3** (**5**), 309-319.

Willan, AR (2001). On the probability of cost-effectiveness using data from randomized clinical trials. *BMC Medical Research Methodology* **1** (**8**).

Willan, AR & Kowgier, ME (2008). Cost-effectiveness analysis of a multinational RCT with a binary measure of effectiveness and an interacting covariate. *Health Economics* **17** (**7**), 777-791.

Willan, AR & O'Brien, BJ (1996). Confidence intervals for cost-effectiveness ratios : an application of Fieller's theorem. *Health Economics* **5** (**3**), 203-211.

Zhou, X-H (1998). Estimation of the log-normal mean.
Statistics in Medicine **17,** 2251-2264.

Zellner, A (1971a). *An Introduction to Bayesian Inference in Econometrics.* Wiley, New York.

Zellner, A (1971b). Bayesian and Non-Bayesian Analysis of the Log-Normal Distribution and Log-Normal Regression.
Journal of the American Statistical Association **66** (**334**), 327-330.

Appendix

There are two parts to the Appendix.

The first part contains the WinBUGS code used in Section 3.4.3.

The second part contains The elicitation procedure.

WinBUGS code

```
model
      {
              mu ~ dnorm( 0 , 0.001 )
              tau ~ dgamma( 0.001 , 0.001 )
              for ( k in 1 : N ) {
                                      X[ k ] ~ dlnorm( mu , tau )
                                }
              ans ~ dlnorm( mu , tau )
      }
list( N = 26 )

 X[ ]
660.3283
194.5458
350.0813
377.6088
48.17163
242.0741
363.7085
79.278
340.5153
276.0091
321.8638
182.7198
240.5935
1138.358
325.4537
79.278
372.4307
72.74691
100.74
19871.29
26201
160.4786
174.6438
1740.847
329.4807
450.8718

list( ans = 1500 , mu = 6 , tau = 0.5 )
```

The elicitation

This section contains the procedure that the facilitator and the expert will follow when they meet and will be produced as a document separate from the rest of this thesis and it will be made available to the expert at the meeting. To control the flow of information a single sheet at a time will be provided.

It is an implicit assumption that the facilitator and expert will be meeting to discuss (budget) costs for a treatment that, in terms of costs, the facilitator is talking to an expert and that there are some relevant costs available which, preferably, the expert has not seen.

The elicitation

Definition

I would like to tell you what is meant by the term elicitation.

Elicitation is the name given to the process that will capture your knowledge about an unknown quantity and represent your prior beliefs in the form of a probability distribution.

I will explain this in more detail now.

Procedure

The procedure that we will follow comprises four parts

Overview

Introduction to the concepts involved

The elicitation that we will follow, with practise

The elicitation

The Sheffield Elicitation Framework

The Sheffield Elicitation Framework, SHELF, will be used to record our meeting and to produce the probability distributions from your prior beliefs.

Overview

We need to be certain that we both know what it is that we want to achieve.

We will be talking about costs measured in £ sterling and what I would like to determine are the beliefs that you hold about these costs before, or in other words prior to, any data is collected.

I will need to ask you about these prior beliefs in a particular way and will explain shortly the technical terms that we will use. Your prior beliefs are a valuable source of information about these costs and it is important that we use your beliefs.

However, the data is also a source of information about the costs and I will then combine these two sources of information to produce posterior knowledge, or in other words our knowledge after observing the data, about these costs.

We will then be able to use this posterior knowledge for making inferences about the (underlying population of) costs from which we have obtained our data sample, in particular the posterior mean value.

In this Bayesian approach, as it is known, we are looking to use all of the available information rather than just relying on the data as used in the traditional approach.

Introduction to the concepts involved

It is important that if either of us uses a particular word then we are both agreed on exactly what it means. I will now introduce the concepts that we will need to conduct this elicitation - which is the name given to the process that will capture your knowledge about an unknown quantity and represent your beliefs in the form of a probability distribution.

I will start this introduction by describing how we identify a statistical model for the data-generating process in terms of a probability distribution.

Probability distributions

We will work with unknown quantities that are continuous because any value within their range is possible. Their probability distributions will define the probability that an unknown quantity lies within some part of its range.

We will find it easiest to work with well known families of distributions, also known as parametric distributions. Parameters control particular properties of distributions and to specify which member of the family of distributions it is necessary to specify the value of the (typically two) parameters.

Perhaps the most widely known distribution is the Normal distribution. This distribution is usually represented as $\mathrm{N}(\mu, \sigma^2)$, where we use the symbol N to represent the Normal family of distributions and two parameters, μ and σ^2, are required to specify which Normal distribution. The Normal distribution has the same shape, symmetric about its central value, for all parameter values.

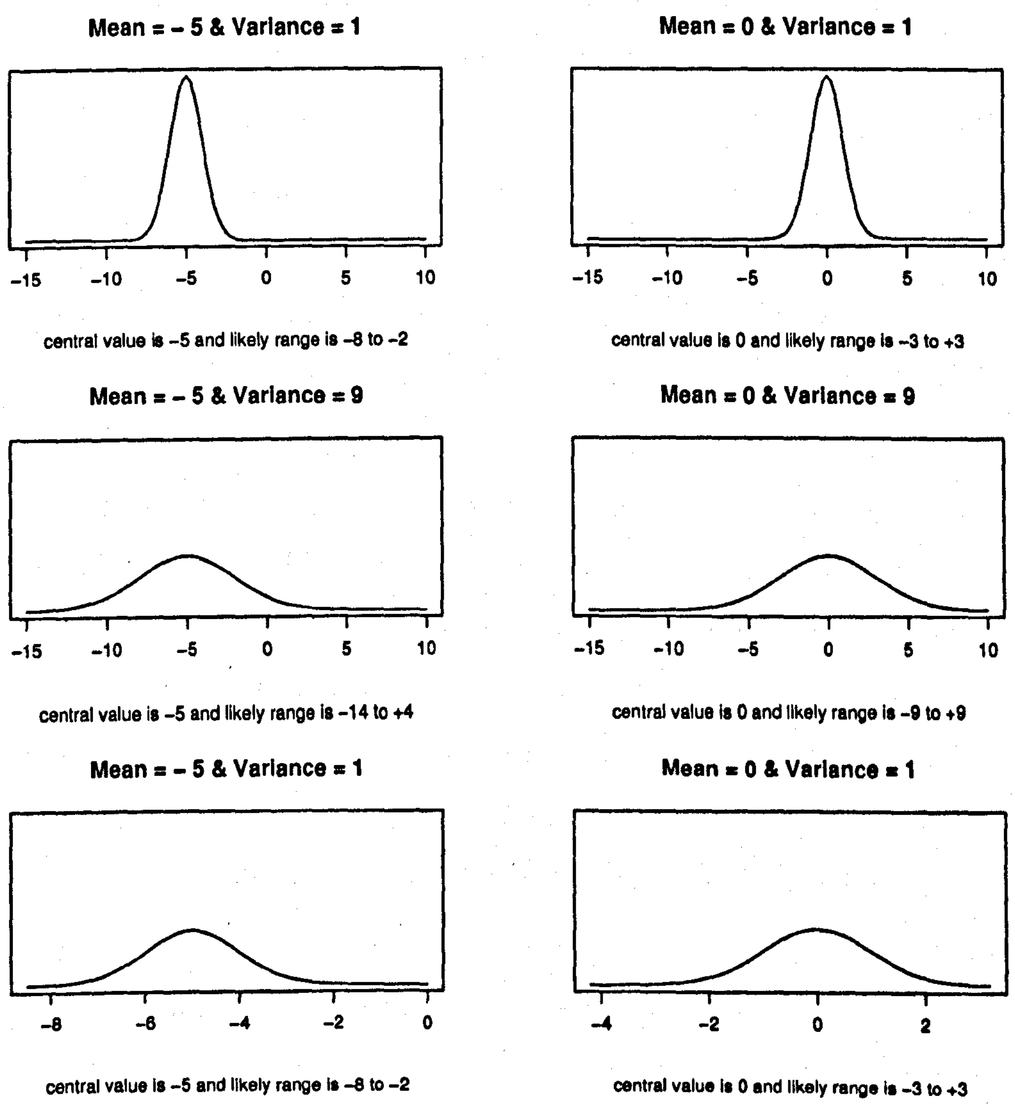
The parameter μ controls the location of the distribution (the central value about which the distribution is clustered) and $\sigma^2 > 0$ controls its scale (the range of values that are likely).

The examples on the following page show the effect of increasing μ and σ^2 on the location and range, but not the shape, of the Normal distribution.

Examples of the Normal distribution where

μ is the mean or location or central value and

σ^2 is the variance or scale or range of likely values



Mean = - 5 & Variance = 1

central value is -5 and likely range is -8 to -2

Mean = 0 & Variance = 1

central value is 0 and likely range is -3 to +3

Mean = - 5 & Variance = 9

central value is -5 and likely range is -14 to +4

Mean = 0 & Variance = 9

central value is 0 and likely range is -9 to +9

Mean = - 5 & Variance = 1

central value is -5 and likely range is -8 to -2

Mean = 0 & Variance = 1

central value is 0 and likely range is -3 to +3

Figure 1: Normal distributions

Uncertainty

When we work with a data-generating model then we need to recognize the two kinds of uncertainty that we have to distinguish between and then take account of

1. we are uncertain about the values of the parameters that precisely define our probability distributions because we lack complete knowledge about their values

2. we are uncertain about which outcomes randomness will produce.

An example where uncertainty arises because of lack of knowledge alone would be your ability to estimate my age.

If we had arranged this meeting by email without having met then you could have formed a prior belief about my age. Once we had met then you may have revised your beliefs and whilst you still could not be certain you would have been able to reduce your uncertainty.

Your beliefs about my age are yours and yours alone. You have formed personal beliefs and there is no reason to think that they may be the same as those held by any one else. This is exactly what we would expect when uncertainty arises because of lack of knowledge alone.

An example where randomness alone determines the outcome would be when tossing an unbiased die.

The probability of any of the faces arising at a single toss is $\frac{1}{6}$.

An example where both kinds of uncertainty are present would be the model used in simple linear regression of $Y_i = \beta_0 + \beta_1 X_i + \epsilon$.

The term ϵ represents the random part of the model. The parameters β_0 and β_1 are unknown and we wish to improve our knowledge by estimating their values.

Statistical summary

While a parameter controls a particular property of a distribution there are other ways to describe important features of the distribution, including the quartiles and in particular the median. They are collectively known as a statistical summary.

Quartiles

Quartiles divide the range of possible values that an unknown quantity Y may take into four sections in a particular way. We are interested here in the Second Quartile, or Median, and the Third Quartile.

The Median for Y may be denoted as the value m where $p\{Y \leq m\} = 0.5$ and clearly $p\{Y > m\} = 0.5$. We can see that an informal interpretation is that the median represents the "middle value" as the probability that the true value lies above the Median is equal to the probability that the true value lies below the Median.

Similarly for Y the Third Quartile may be denoted as the value t where $p\{Y \leq t\} = 0.75$. Here, the probability that the true value lies below the Third Quartile is three times the probability that the true value lies above the Third Quartile and the Third Quartile represents the "three quarters value".

We also need to introduce here the Quartile Ratio as

$$\frac{\text{Third Quartile}}{\text{Median}}$$

which will be explained in more detail shortly.

Positively skewed distributions

It is often the case that distributions are not symmetric and other possibilities include "J shaped" and skew. A distribution that can take any non-negative value (bounded below by zero but without an upper limit) and is positively skewed, also known as right skewed, is considered to be a good model for financial distributions eg the distribution of salaries within an organization.

As well as parameters it is possible to use functions of parameters to capture your views about the properties of the data-generating distribution. We will use here the Median and Quartile Ratio of the data-generating distribution to control the scale and shape respectively.

The scale of the distribution is determined by its Median. The larger the value of the Median, the larger the range of values over which the distribution is spread.

Similarly, the shape of the distribution is determined by its Quartile Ratio. The larger the value of the Quartile Ratio, the greater the positive skew of values of the distribution.

So we will consider here a distribution as a model for the data-generating process that has two fixed but unknown parameters. This elicitation is about gathering your understanding about this distribution, prior to observing the data. We will capture your knowledge about its Median and Quartile Ratio in a very specific way.

We now need to consider how we determine your prior beliefs about the costs when we use a distribution that only takes non-negative values and is skewed to the right as the model for the data-generating process and it's only (the values of) the Median and Quartile Ratio that are unknown.

Population

The population that we will be considering is the collection of the individual costs from each of those patients whose condition means that they will receive this treatment. We will denote the cost using the symbol Y.

An individual member of the population is subject to uncertainty about the values of both the parameters for the distribution from which it is drawn as well as the random effect involved in choosing that member of the population.

When we consider the population our only uncertainty is about the values of the parameters.

Parameters

The parameters are fixed but unknown, which also means that functions of the parameters, for example the Median, are fixed but unknown.

It is your lack of knowledge about these parameters which leads to your uncertainty about their values; there is no randomness involved.

We will capture your prior beliefs about the Median and Quantile Ratio by way of probability statements and then represent your prior beliefs in the form of a probability distribution. This will enable me, when the data is available, to combine these two sources of information to produce posterior views.

Elicitation

For a population the fixed but unknown parameters, and also functions of parameters, control properties of the distribution. The fixed but unknown quartiles describe important features of the distribution.

We will work here with the Median (M_Y) and Third Quartile (T_Y) and also the Quartile Ratio (R_Y). Each of these quantities is fixed, but unknown, for our population of costs and I want to capture your prior beliefs about them by way of probability statements and then represent your prior beliefs in the form of a probability distribution

Ratios

A ratio is defined as
ratio = numerator / denominator = num / den
and if num and den > 0 with num $>$ den then ratio > 1
and its value (only) tells us how many times num is greater than den
eg ratio $= 9$ could be $= 9/1$ or $= 18/2$ or $= 80.1/8.9$ etc etc.

If we were to analyse different volumes of air to determine their constituent volumes of oxygen and nitrogen then we would find that (approximately) the ratio of nitrogen to oxygen (by volume) $= 78/21$ or $= 156/42$ or $= 234/63$ $\simeq 3.7$.

Quartile ratio

We will be looking specifically at a Quartile Ratio defined as

$$R_Y = \frac{T_Y}{M_Y} = \frac{\text{Third Quartile}}{\text{Median}}$$

The larger the value of R_Y the larger, or more extreme, the positive skew.

The Quartile Ratio takes values in the range $(1, \infty)$ because num and also den > 0 with num $>$ den. As num and den are each costs measured in £ sterling then the quartile ratio is a value expressed as a (dimensionless) number.

Your prior beliefs

We will capture your prior beliefs about the Median, Third Quartile and also the Quartile Ratio by asking you about specific values as we will show in the example that follows for R_Y.

The value $l(R_Y)$ is the value such that $p\{R_Y \leq l(R_Y)\} = \frac{1}{3}$ and similarly the value $u(R_Y)$ is the value such that $p\{R_Y \leq u(R_Y)\} = \frac{2}{3}$. So in this case the values $l(R_Y)$ and $u(R_Y)$ divide the range of possible values into three sections with equal probability that the true value lies in any of the three sections.

We can also capture your beliefs about the value $m(R_Y)$ where this is the value such that $p\{R_Y \leq m(R_Y)\} = \frac{1}{2}$ and the value $m(R_Y)$ divides the range of possible values into two sections with equal probability that the true value lies in either of the two sections.

The Median and the Quartile Ratio

We want to determine whether knowing the value of the Median gives you any information about your value of the Quartile Ratio.

A practise

We want to use a positively skewed population, that you are not unfamiliar with, to practise the concepts involved.

The distribution of employee salary within the organisation that you work would appear to present a suitable opportunity.

The definition of salary that we wish to use here is individual gross pay for employees, which excludes any income earned outside the organisation, with our unit of measurement as £ sterling.

We will now conduct an elicitation following the stages outlined in the next section - The elicitation that we will follow.

The elicitation that we will follow

We will explain how we will determine the quantile values required in the four stages below as we deal with each stage in turn.

Population quartiles

In this first stage we are looking for values that are measured in £ sterling.

Whilst the Median, M_Y, and Third Quartile, T_Y, are fixed for the population you cannot be certain about their values, although they will be greater than 0. We want to capture your uncertainty about M_Y and T_Y as follows.

We want to elicit specific values for M_Y and T_Y. We want to ask you to

(a) determine the largest value for M_Y that you believe is possible

(b) determine $m(M_Y)$ such that M_Y is equally likely to be above or below this value

(c) determine the largest value for T_Y that you believe is possible

(d) determine $m(T_Y)$ such that T_Y is equally likely to be above or below this value

Quartile ratio

We will be looking at the Quartile Ratio, R_Y, where the larger the value of R_Y the larger, or more extreme, the positive skew which means that the probability of a very large value for Y is larger, as the figure below indicates

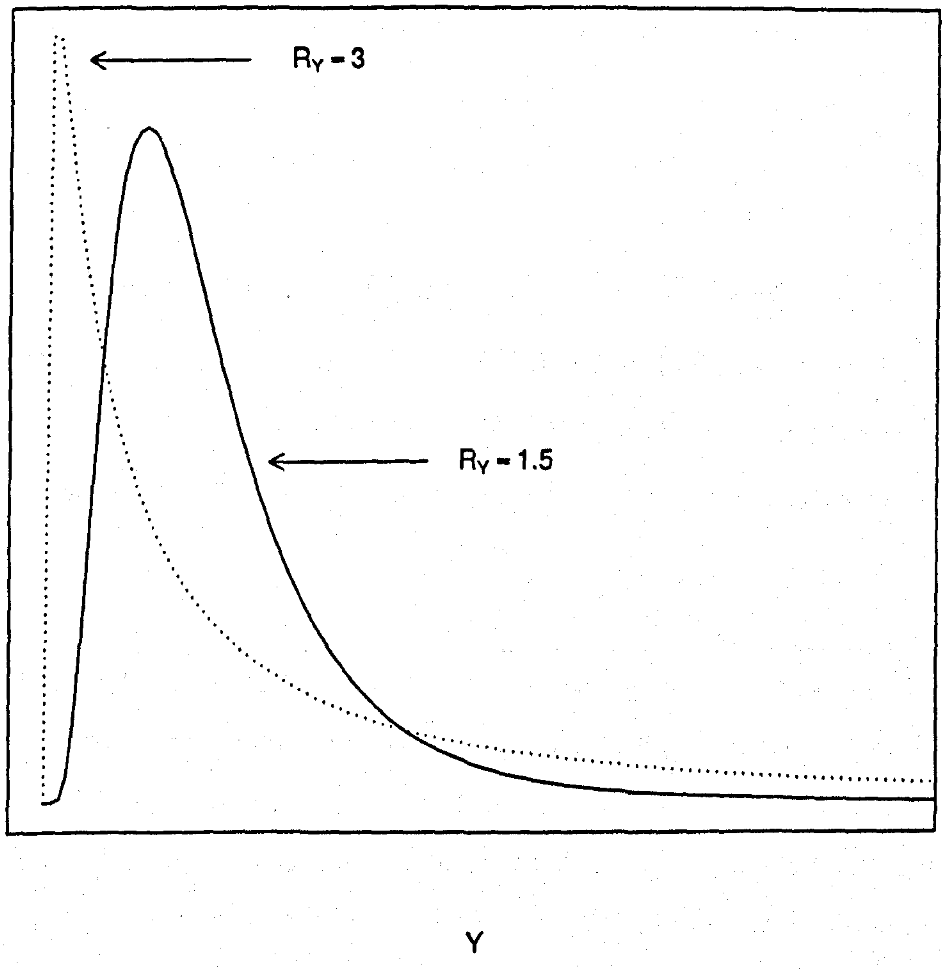


Figure 2: Distribution of Y for two values of the Quartile Ratio

The Quartile Ratio takes values in the range $(1, \infty)$ because num and also den > 0 with num $>$ den. As num and den are each costs measured in £ sterling then the quartile ratio is a value expressed as a (dimensionless) number.

So in this second stage we are looking for (dimensionless) numbers.

From the first stage we can determine a plausible value for the Quartile Ratio, R_Y, as $m(T_Y)/m(M_Y)$.

Whilst R_Y is fixed for the population you cannot be certain about its value, although it will be greater than 1. We want to capture your uncertainty about R_Y in the form of a probability distribution as follows.

We want to elicit specific values for R_Y. We want to ask you to

(a) determine the largest value that you believe is possible for R_Y

(b) determine $l(R_Y)$ and $u(R_Y)$ such that R_Y is equally likely to be below $l(R_Y)$, as above $u(R_Y)$, as between these two values

(c) determine $m(R_Y)$ such that R_Y is equally likely to be above or below this value

(d) we will then produce a distribution fitted to your elicited values and show this to you to ascertain if it represents your prior beliefs for R_Y. We will modify this distribution until you are satisfied with the final (visual) result. We will then be able to feedback to you some probability statements to confirm the final fitted distribution. If needs be we continue with this "fitted distribution and feedback" cycle until a satisfactory result has been obtained.

Median

The scale of a distribution is determined by its Median. The larger the value of the Median, the larger the range of values over which the distribution is spread, as the figure below indicates

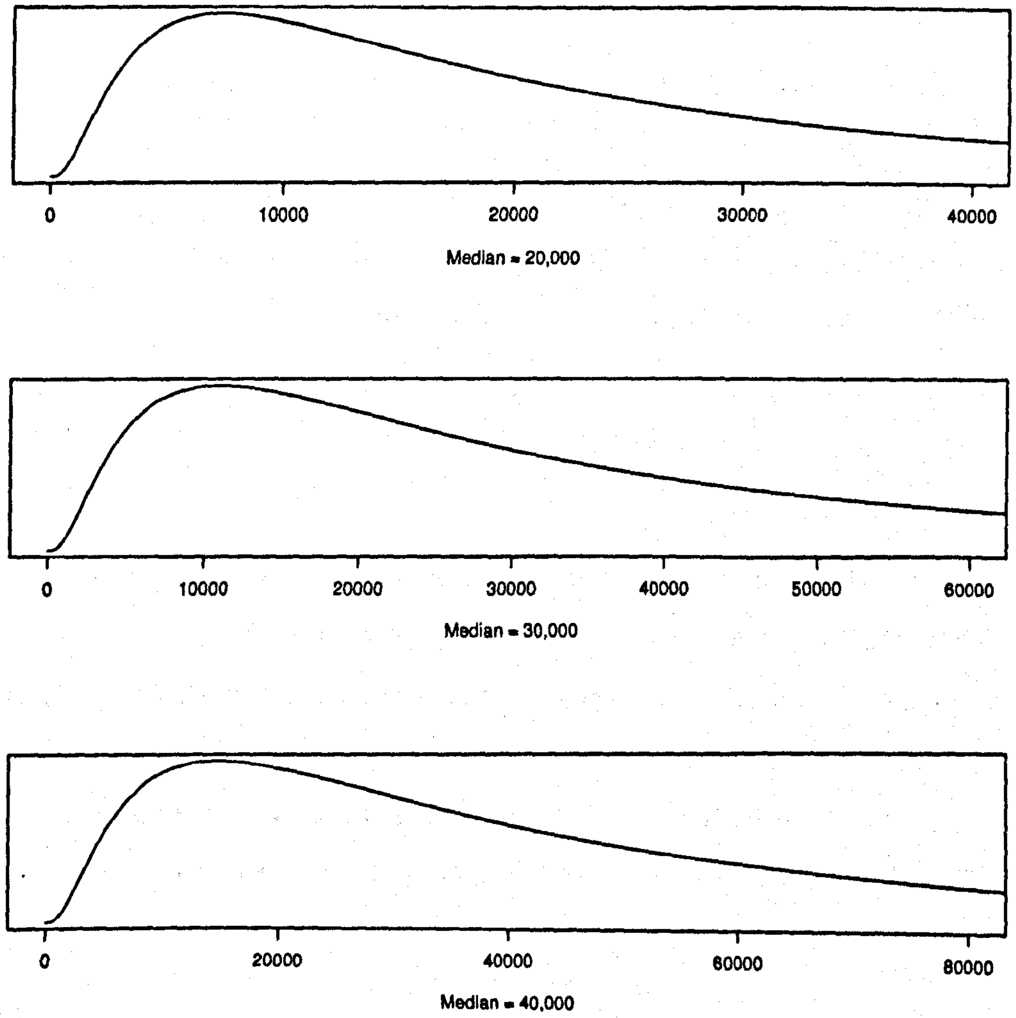


Figure 3: Distribution of Y for three values of the Median

Median

In this third stage we are looking for values that are measured in £ sterling. Whilst the Median, M_Y, is fixed for the population you cannot be certain about its value, although it will be greater than 0. We want to capture your uncertainty about M_Y in the form of a probability distribution as follows.

We would like to remind you that, from the first stage, you determined the largest value for M_Y that you believe is possible.

We want to elicit specific values for M_Y. We want to ask you to

(a) determine $l(M_Y)$ and $u(M_Y)$ such that M_Y is equally likely to be below $l(M_Y)$ as above $u(M_Y)$ as between these two values

(b) from the first stage you have determined $m(M_Y)$ such that M_Y is equally likely to be above or below this value

(c) we will then produce a distribution fitted to your elicited values and show this to you to ascertain if it represents your prior beliefs for M_Y. We will modify this distribution until you are satisfied with the final (visual) result. We will then be able to feedback to you some probability statements to confirm the final fitted distribution. If needs be we continue with this "fitted distribution and feedback" cycle until a satisfactory result has been obtained.

The Median and the Quartile Ratio

We want to determine whether knowing the value of the Median gives you any information about your value of the Quartile Ratio.

From the second stage we have been able to confirm the distribution for your prior beliefs for the Quartile Ratio and similarly for the Median from the third stage.

We will show you your confirmed prior distribution for the Quartile Ratio.

We will ask you if there is any particular value that could be chosen for the Median from your confirmed prior distribution, which is now shown as a reminder, that would that cause you to want to change your beliefs about your confirmed prior distribution for the Quartile Ratio.

Reflection

Having completed the practise elicitation, using the distribution of employee salary within the organisation that you work as our population, it will be opportune to reflect on the last two parts of the procedure, namely

Introduction to the concepts involved
The elicitation that we will follow, with practise

to ask if there are any aspects that we should return to or refresh ourselves about.

The elicitation

We are now ready to conduct our elicitation.