ANALYSIS OF UNBALANCED OCCUPATIONAL EXPOSURE DATA USING A BAYESIAN RANDOM EFFECTS MODEL

Justin Harvey [1]
Centre for Statistical Consultation, University of Stellenbosch
e-mail: *jharvey@sun.ac.za*

and
Abrie van der Merwe
University of the Free State

Key words: Occupational exposure limit (OEL), one-way random effects model, Bayes, non-informative prior.

Summary: Many authors have approached the analysis of occupational exposure data using a one-way random effects model. Krishnamoorthy and Mathew (2002) in particular proposed the use of generalized confidence intervals and a generalized p-value approach for the case of balanced data. The primary variable of interest was the occupational exposure limit (OEL). Harvey and van der Merwe (2014) proposed the use of an objective Bayesian approach to the problem. In this article we extend this Bayesian approach to include the case of unbalanced data. Again, Krishnamoorthy and Guo (2005), amongst others, have proposed non-Bayesian methods for modelling this case of 'incomplete'or unbalanced data. In this article the authors apply an objective Bayesian approach to the problem and evaluate the performance of several non-informative priors.

[1] Corresponding author.

AMS: 62GXX, 62FXX, 62DXX

1. Introduction

The presence of lognormally distributed data is a frequent occurrence in many analysis settings. Occupational health settings are among these. Many different methods have been presented for the analysis of data that is lognormally distributed and one such method is the one-way random effects model, as proposed, for example, by Krishnamoorthy and Mathew (2002). The primary parameter of interest was the occupational exposure limit (OEL) for lognormally distributed data. The interested reader is referred to the original articles by Krishnamoorthy and Mathew (2002) for a more complete description of the medical applications of this method as well as the texts by Rappaport, Kromhout and Symanski (1993), Heerderik and Hurley (1994), Lyles, Kupper and Rappaport (1997b) and Lyles, Kupper and Rappaport (1997a).

It is not just confidence intervals that are of interest, but the need also exists to test hypotheses concerning the primary parameter. To do this Krishnamoorthy and Mathew (2002) extended their previous work in this setting and attempted to analyse the data using generalized p-values and generalized confidence intervals. They were able to test specific hypotheses regarding the OEL.

The setting was analysed from a Bayesian perspective in Harvey and van der Merwe (2014). In the article they presented an objective Bayesian approach for modelling the arithmetic mean of the OEL using the one-way random effects model and compared the effect of several non-informative priors. It was shown that the Bayesian approach has several distinct advantages over the generalized confidence interval and p-value approach. The most evident advantage was the flexibility of the Bayesian approach that allowed for the modelling of mean exposure for individual workers.

The previous examples and articles considered the case of balanced data, whereby there are an equal number of observations for each observational unit, such as an individual worker, company or even groups thereof. Unfortunately, the case of balanced data is overly simplistic. Unbalanced data can arise due to a number of different factors. The analysis of unbalanced data will in turn require an analysis framework that accounts specifically for this setting.

Again, this situation has been approached by some authors (e.g. Krishnamoorthy and Guo (2005)), but the methods proposed involve generalized p-value approaches. The problem statement is nevertheless the same: we would like to estimate the proportion of exposure measurements exceeding a pre-specified limit (OEL) or perhaps the probability of an insurance claim exceeding a pre-specified boundary. According to Krishnamoorthy and Guo (2005) the one-way random effects model incorporates both within and between sources of variation in measurements. Since we are dealing with data that have a lognormal distribution (i.e. the logged exposure levels are normally distributed) we are interested in the overall mean effect and the two variance components associated with the random effects model.

In this article we extend the work presented by Harvey and van der Merwe (2014) regarding the one-way balanced random effects model to the unbalanced case from an objective Bayesian perspective. In order to complete the Bayesian specification of the model prior distributions have to be derived and this forms a large part of this article. The selection and determination of non-

informative priors in multi-parameter settings is not an easy task and it has been observed that the selection of a specific prior could have unexpectedly dramatic effects on the posterior distribution. In this article, the derivation of suitable priors will be considered, where the Reference prior (Berger and Bernanrdo (1992)) is one such option and the second is the Probability-Matching prior. A simulation study will also be presented to show the effectiveness of the proposed prior distributions.

2. Description of the setting

The setting for this article is similar to the setting described in Harvey and van der Merwe (2014). The point of departure, conceptually, is that the data now is unbalanced. The previous data referred to the amount of exposure to a particular agent and for each worker there were exactly the same number of observations. In the unbalanced case we have an unequal number of observations for each worker (the mechanism by which this "missing"data is generated is not of interest in this article). Even though this is a minor conceptual change all derivations of priors and posterior distributions would necessarily change.

Therefore, in the unbalanced case we have the following diagrammatic representation of the data:

Table 1: Representation of Shift Exposure Data.

	Shift-Long Exposure Measurements			
Worker	**1**	**2**	...	$\mathbf{n}_i$
1	x_{11}	x_{12}	...	x_{1n_1}
1	x_{21}	x_{22}	...	x_{2n_2}
...	...	...	...	...
k	x_{k1}	x_{k2}	...	x_{kn_k}

Let X_{ij} represent the j-th shift-long exposure measurement for the i-th worker, where $j = 1,\ldots,n_i$ and $i = 1,\ldots,k$. Therefore, there are n_i measurements for the i-th worker, which results in the "unbalanced"nature of the data. The X_{ij} are lognormally distributed and therefore $Y_{ij} = ln(X_{ij})$ are normally distributed. The situation can be represented by the following one-way random effects model:

$$Y_{ij} = \mu + \tau_i + e_{ij}, \qquad i = 1,\ldots,k; j = 1,\ldots,n_i.$$

where μ is the general mean, $\tau_i \sim N\left(0,\sigma_\tau^2\right)$ and $e_{ij} \sim N\left(0,\sigma_e^2\right)$. All the random variables are independent of each other and here τ_i represents the random effect due to the i-th worker.

Given the lognormal distribution of the X_{ij}

$$\mu_{x_i} = E\left(X_{ij}|\tau_i\right) = E\left(exp\left[Y_{ij}\right]|\tau_i\right) = exp\left(\mu + \tau_i + {}^{\sigma_e^2}/_2\right)$$

and μ_{x_i} is the mean exposure for the i-th worker. Let θ denote the probability that μ_{x_i} exceeds the OEL. Thus, $\theta = P(\mu_{x_i} > OEL) = P(ln(\mu_{x_i}) > ln(OEL)) = 1 - \Phi\left(\frac{ln(OEL) - \mu - \sigma_e^2/2}{\sigma_\tau}\right)$ where $\Phi(.)$ denotes the c.d.f. of the standard normal distribution. The kind of hypotheses that are going to be considered here are:

$$H_0 : \theta \geq A \quad vs \quad H_1 : \theta < A$$

where A is a specific quantity that is usually small, according to Krishnamoorthy and Guo (2005).

Table 2 is an example data set of simulated "styrene exposures"that will serve as a basis for discussion in this article and will help us define and illustrate the objectives of the article (it is the same data set as in Harvey and van der Merwe (2014), however, observations have been removed at random, resulting in an "unbalanced"design).

Table 2: Simulated Styrene Exposures.

	Shift-Long Exposures									
Worker	1	2	3	4	5	6	7	8	9	10
1	95.6	64.7	50.9	87.4	82.3	149.9	33.4	77.5	70.8	60.9
2	57.4	82.3	174.2	107.8	98.5	129	121.5	95.6	92.8	133
3	84.8	214.9	79.8	169	149.9	164	84.8	84.8	114.4	
4	68.7	77.5	54.1	41.3	64.7	46.5	59.1	45.2	54.1	
5	114.4	101.5	49.4	101.5	90	52.5	114.4	79.8	68.7	87.4
6	87.4	242.3	145.5	133	174.2	214.9	137	129	169	179.5
7	54.1	75.2	84.8	55.7	90	70.8	60.9	101.5	64.7	95.6
8	64.7	95.6	57.4	95.6	82.3	101.5	92.8	60.9	101.5	98.5
9	137	208.5	92.8	159.2	92.8	82.3	90			
10	125.2	87.4	121.5	90	154.5	107.8	117.9	179.5	129	129
11	42.5	73	50.9	59.1	49.4	66.7				
12	57.4	68.7	59.1	64.7	55.7	92.8	42.5			
13	101.5	149.9	111.1	77.5	111.1	84.8	64.7	62.8		
14	68.7	101.5	111.1	179.5	82.3	174.2	174.2	87.4	145.5	114.4
15	121.5	77.5	145.5	174.2	77.5	92.8	159.2	129	104.6	77.5

Table 2 represents the X_{ij} data points, from which the $Y_{ij} = ln(X_{ij})$ can easily be obtained.

From these data we have the following definitions and associated results (these results will be used in all subsequent applications and examples):

$k = 15$

$\nu_1 = \sum_{i=1}^{k} (n_i - 1)\,;\ \nu_2 = k - 1$

$\bar{Y}_{i\bullet} = \frac{1}{n_i} Y_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = [4.3\ \ 4.7\ \ 4.8\ \ 4.0\ \ 4.4\ \ 5.0\ \ 4.3\ \ 4.4\ \ 4.8\ \ 4.9\ \ 4.0\ \ 4.1\ \ 4.5\ \ 4.8\ \ 4.7]'$

$n_i = [10\ \ 10\ \ 9\ \ 9\ \ 10\ \ 10\ \ 10\ \ 10\ \ 7\ \ 10\ \ 10\ \ 6\ \ 7\ \ 8\ \ 10\ \ 10]'$

$$\bar{Y}_{\bullet\bullet} = \frac{1}{k}\sum_{i=1}^{k}\bar{Y}_{i\bullet} = 4.508 = \hat{\mu}$$

$$SS_e = \nu_1 m_1 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\bullet}\right)^2 = 10.492 = ''\textit{ within workers sum of squares}''$$

$$SS_\tau = \nu_2 m_2 = \sum_{i=1}^{k} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = 11.885 = ''\textit{ between workers sum of squares}''$$

3. Bayesian methodology

The basis for analyzing any situation from a Bayesian perspective is the following relationship, a well-known result of Bayes'theorem: $\textit{Posterior} \propto \textit{Likelihood} \times \textit{Prior}$. The likelihood function (in matrix form) is given by:

$$L\left(\mu, \boldsymbol{\tau}, \sigma_e^2, \sigma_\tau^2 | \mathbf{Y}\right) = \left(2\pi\sigma_e^2\right)^{-\frac{1}{2}\tilde{n}} exp\left\{-\frac{1}{2\sigma_e^2}\left(\mathbf{Y} - \mu\mathbf{1} - Z\boldsymbol{\tau}\right)'\left(\mathbf{Y} - \mu\mathbf{1} - Z\boldsymbol{\tau}\right)\right\} \times \left(2\pi\sigma_\tau^2\right)^{-\frac{1}{2}k} exp\left\{-\frac{1}{2\sigma_\tau^2}\boldsymbol{\tau}'\boldsymbol{\tau}\right\} \quad (1)$$

where

$$\tilde{n} = \sum_{i=1}^{k} n_i,\ \boldsymbol{\tau}' = \begin{bmatrix}\tau_1 & \tau_2 & \dots & \tau_k\end{bmatrix},\ \mu\mathbf{1} = \begin{bmatrix}\mu & \mu & \cdots & \mu\end{bmatrix}'$$

$$Z'_{\tilde{n}\times k} = \begin{bmatrix} 1_1 & 1_2 & \cdots & 1_{n_1} & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1_1 & 1_2 & \cdots & 1_{n_2} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 1_1 & 1_2 & \cdots & 1_{n_k} \end{bmatrix}$$

and

$$\mathbf{Y} = \begin{bmatrix}y_{11}\ y_{12}\ \cdots\ y_{1n_1}\ \cdots\ y_{k1}\ y_{k2}\ \cdots\ y_{kn_k}\end{bmatrix}'$$

Now, we already know, from the specification of the random effects model, that $\tau_i \sim N\left(0, \sigma_\tau^2\right)$ with $i = 1,\ 2,\ \dots,\ k$. Since this is the case we would therefore like to define prior distributions for μ, σ_e^2 and σ_τ^2. For the sake of convenience (since the posterior can then be expressed in hierarchical form) though we will define the quantity

$$\tilde{r} = \frac{\sigma_\tau^2}{\sigma_e^2}$$

and then define prior distributions for μ, σ_e^2 and $\tilde{r}$ instead. In order to derive prior distributions for

this though we first need to derive the integrated likelihood function, $L\left(\mu,\sigma_e^2,\sigma_\tau^2|\mathbf{Y}\right)$. Theorems (1) to (6) below all refer to the following model:

$$\mathbf{Y} = \mu\mathbf{1} + Z\boldsymbol{\tau} + \mathbf{e}$$

where $\mathbf{e} \sim N\left(\mathbf{0},\sigma_e^2 I_{\tilde{n}}\right)$ and $\boldsymbol{\tau} \sim N\left(\mathbf{0},\sigma_\tau^2 I_k\right)$.

All proofs of Theorems (1) to (7) can be found in Harvey (2012).

Theorem 1 The integrated likelihood function, $L\left(\mu,\sigma_e^2,\sigma_\tau^2|\mathbf{Y}\right)$ is given by the following:

$$L\left(\mu,\sigma_e^2,\sigma_\tau^2|\mathbf{Y}\right) \propto \left(\sigma_e^2\right)^{-\frac{1}{2}(\tilde{n}-k)} \prod_{i=1}^{k}\left(\frac{1}{n_i\sigma_\tau^2+\sigma_e^2}\right)^{\frac{1}{2}} exp\left\{-\frac{1}{2}\left[\frac{\nu_1 m_1}{\sigma_e^2}+\sum_{i=1}^{k}\frac{n_i(\bar{y}_{i\bullet}-\mu)^2}{n_i\sigma_\tau^2+\sigma_e^2}\right]\right\} \quad (2)$$

Now, if $\tilde{r} = \frac{\sigma_\tau^2}{\sigma_e^2}$ then it follows that

$$L\left(\mu,\sigma_e^2,\tilde{r}|\mathbf{Y}\right) \propto \left(\sigma_e^2\right)^{-\frac{1}{2}\tilde{n}} \prod_{i=1}^{k}\left(\frac{1}{n_i\tilde{r}+1}\right)^{\frac{1}{2}} exp\left\{-\frac{1}{2\sigma_e^2}\left[\nu_1 m_1+\sum_{i=1}^{k}\frac{n_i(\bar{y}_{i\bullet}-\mu)^2}{n_i\tilde{r}+1}\right]\right\} \quad (3)$$

Given this result we can prove the following theorems:

Theorem 2 $\bar{y}_{i\bullet}|\mu,\sigma_e^2,\sigma_\tau^2 \sim N\left(\mu\ ,\ \frac{n_i\sigma_\tau^2+\sigma_e^2}{n_i}\right)$

Theorem 3 The Fisher Information Matrix for the parameters $\left(\mu,\tilde{r},\sigma_e^2\right)$ is given by

$$F\left(\mu,\tilde{r},\sigma_e^2\right) = \begin{bmatrix} \frac{1}{\sigma_e^2}\sum_{i=1}^{k}\frac{n_i}{1+\tilde{r}n_i} & 0 & 0 \\ 0 & \frac{1}{2}\sum_{i=1}^{k}\frac{n_i^2}{(1+\tilde{r}n_i)^2} & \frac{1}{2\sigma_e^2}\sum_{i=1}^{k}\frac{n_i}{1+\tilde{r}n_i} \\ 0 & \frac{1}{2\sigma_e^2}\sum_{i=1}^{k}\frac{n_i}{1+\tilde{r}n_i} & \frac{\tilde{n}}{2}\left(\frac{1}{\sigma_e^2}\right)^2 \end{bmatrix}$$

In this article, two non-informative priors are compared, namely the Probability-matching and Reference priors. These priors often lead to procedures with good frequentist properties while retaining the Bayesian flavor. The fact that the resulting posterior intervals of level $1-\alpha$ are also good frequentist intervals at the same level is a very desirable situation. An in depth discussion of the nature and merits of the Reference and Probability-matching priors lies outside the scope of this article, but the interested reader should consult Berger and Bernanrdo (1992) as well as Datta and Ghosh (1995).

We derive the necessary prior distributions in the following theorems:

Theorem 4 The Probability-Matching Prior for the parameters $\left(\mu,\tilde{r},\sigma_e^2\right)$ is given by

$$P\left(\mu,\tilde{r},\sigma_e^2\right) \propto \frac{1}{\sigma_e^2}\left\{\sum_{i=1}^{k}\frac{n_i^2}{(1+\tilde{r}n_i)^2}-\frac{1}{n}\left(\sum_{i=1}^{k}\frac{n_i}{1+\tilde{r}n_i}\right)^2\right\}^{\frac{1}{2}} \quad (4)$$

Theorem 5 The Reference Prior for the parameter groupings $\left(\mu,\tilde{r},\sigma_e^2\right)$, $\left(\tilde{r},\ \mu,\sigma_e^2\right)$ and $\left(\tilde{r},\sigma_e^2,\ \mu\right)$ is given by

$$P_{R_1}\left(\mu,\tilde{r},\sigma_e^2\right)\ \propto\ \frac{1}{\sigma_e^2}\left\{\sum_{i=1}^{k}\frac{n_i^2}{\left(1+\tilde{r}n_i\right)^2}-\frac{1}{\tilde{n}}\left(\sum_{i=1}^{k}\frac{n_i}{1+\tilde{r}n_i}\right)^2\right\}^{\frac{1}{2}} \tag{5}$$

This is coincidentally the same as the Probability-Matching Prior and therefore the Probability-Matching Prior is also the Reference Prior.

Theorem 6 The Reference Prior for the parameter groupings $\left(\mu,\sigma_e^2,\tilde{r}\right)$, $\left(\sigma_e^2,\mu,\tilde{r}\right)$ and $\left(\sigma_e^2,\tilde{r},\mu\right)$ is given by

$$P_{R_2}\left(\mu,\ \sigma_e^2\ ,\ \tilde{r}\right)\ \propto\ \frac{1}{\sigma_e^2}\left\{\sum_{i=1}^{k}\frac{n_i^2}{\left(1+\tilde{r}n_i\right)^2}\right\}^{\frac{1}{2}} \tag{6}$$

It should be evident that if we substitute $n_1=\ n_2=\ \ldots=\ n_k=n$ and $\tilde{n}=kn$ in equations (5) and (6), i.e. assume we have the balanced case and if we transform $\tilde{r}$ back to σ_τ^2, then equations (5) and (6) become the Jeffreys Independence priors as used in Harvey and van der Merwe (2014).

3.1. Joint posterior distribution for μ , σ_e^2 and $\tilde{r}$

We are now able to examine the distribution of the posterior distribution of μ ,σ_e^2 and $\tilde{r}$. This is based on the previous derivations and theorems that have been stated. From the formulation of the Bayesian model we know the following:

$$p\left(\mu,\ \sigma_e^2\ ,\ \tilde{r}|\ \mathbf{Y}\right)\ \propto L\left(\mu,\ \sigma_e^2\ ,\ \tilde{r}\ |\ \mathbf{Y}\right)p\left(\mu,\ \sigma_e^2\ ,\ \tilde{r}\right)$$

where

$$L\left(\mu,\sigma_e^2,\tilde{r}|\mathbf{Y}\right)\ \propto\ \left(\sigma_e^2\right)^{-\frac{1}{2}\tilde{n}}\prod_{i=1}^{k}\left(\frac{1}{n_i\tilde{r}+\ 1}\right)^{\frac{1}{2}}exp\left\{-\frac{1}{2\sigma_e^2}\left[\nu_1m_1+\ \sum_{i=1}^{k}\frac{n_i(\bar{y}_{i\bullet}-\ \mu)^2}{n_i\tilde{r}+\ 1}\right]\right\}$$

If we use the Probability-Matching prior as defined by equation (4), which is the same as the Reference prior for the first ordering of parameters as described in equation (5), then the joint posterior distribution is given by:

$$P_{R_1}\left(\mu,\tilde{r},\sigma_e^2|\mathbf{Y}\right)\ \propto\ \left(\frac{1}{\sigma_e^2}\right)^{\frac{1}{2}(\tilde{n}+2)}\prod_{i=1}^{k}\left(\frac{1}{n_i\tilde{r}+\ 1}\right)^{\frac{1}{2}}\ \times exp\left\{-\frac{1}{2\sigma_e^2}\left[\nu_1m_1+\ \sum_{i=1}^{k}\frac{n_i(\bar{y}_{i\bullet}-\ \mu)^2}{n_i\tilde{r}+\ 1}\right]\right\}$$
$$\times\ \left\{\sum_{i=1}^{k}\frac{n_i^2}{\left(1+\tilde{r}n_i\right)^2}-\frac{1}{n}\left(\sum_{i=1}^{k}\frac{n_i}{1+\tilde{r}n_i}\right)^2\right\}^{\frac{1}{2}} \tag{7}$$

where

$$\nu_1 m_1 = SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})^2$$

From equation (7) it follows that the joint posterior can be expressed hierarchically as

$$P_{R_1}\left(\mu, \tilde{r}, \sigma_e^2 | \mathbf{Y}\right) = p\left(\mu | \mathbf{Y},\ \tilde{r}, \sigma_e^2\right)\ \times\ p\left(\sigma_e^2 | \tilde{r},\ \mathbf{Y}\right)\ \times\ p\left(\tilde{r} |\ \mathbf{Y}\right)$$

where

$$\mu | \mathbf{Y},\ \tilde{r}, \sigma_e^2\ \sim\ N\left(\hat{\mu}\ ,\ \sigma_e^2 \left(\sum_{i=1}^{k} \frac{n_i}{1+\tilde{r}n_i}\right)^{-1}\right). \tag{8}$$

and

$$\hat{\mu} = \frac{\sum_{i=1}^{k} \bar{y}_{i\bullet} \frac{n_i}{1+\tilde{r}n_i}}{\sum_{i=1}^{k} \frac{n_i}{1+\tilde{r}n_i}}$$

In addition,

$$P_{R_1}\left(\sigma_e^2 | \tilde{r},\ \mathbf{Y}\right) =\ K_1 \left(\frac{1}{\sigma_e^2}\right)^{\frac{1}{2}(\tilde{n}+1)} exp\left\{-\frac{1}{2\sigma_e^2}\left[\nu_1 m_1 + \sum_{i=1}^{k} \frac{n_i (\bar{y}_{i\bullet} - \hat{\mu})^2}{n_i \tilde{r} + 1}\right]\right\} \tag{9}$$

which is an inverse Gamma distribution. Furthermore, we know that

$$K_1 = \left\{\frac{1}{2}\left[\nu_1 m_1 + \sum_{i=1}^{k} \frac{n_i (\bar{y}_{i\bullet} - \hat{\mu})^2}{n_i \tilde{r} + 1}\right]\right\}^{\frac{1}{2}(\tilde{n}-1)}$$

and

$$P_{R_1}\left(\tilde{r} | \mathbf{Y}\right) \propto \prod_{i=1}^{k} \left(\frac{1}{n_i \tilde{r} + 1}\right)^{\frac{1}{2}} \times \left(\sum_{i=1}^{k} \frac{n_i}{1+\tilde{r}n_i}\right)^{-\frac{1}{2}} \times \left\{\sum_{i=1}^{k} \frac{n_i^2}{(1+\tilde{r}n_i)^2} - \frac{1}{n}\left(\sum_{i=1}^{k} \frac{n_i}{1+\tilde{r}n_i}\right)^2\right\}^{\frac{1}{2}} \times \left[\nu_1 m_1 + \sum_{i=1}^{k} \frac{n_i (\bar{y}_i - \hat{\mu})^2}{n_i \tilde{r} + 1}\right]^{-\frac{1}{2}(n-1)} \tag{10}$$

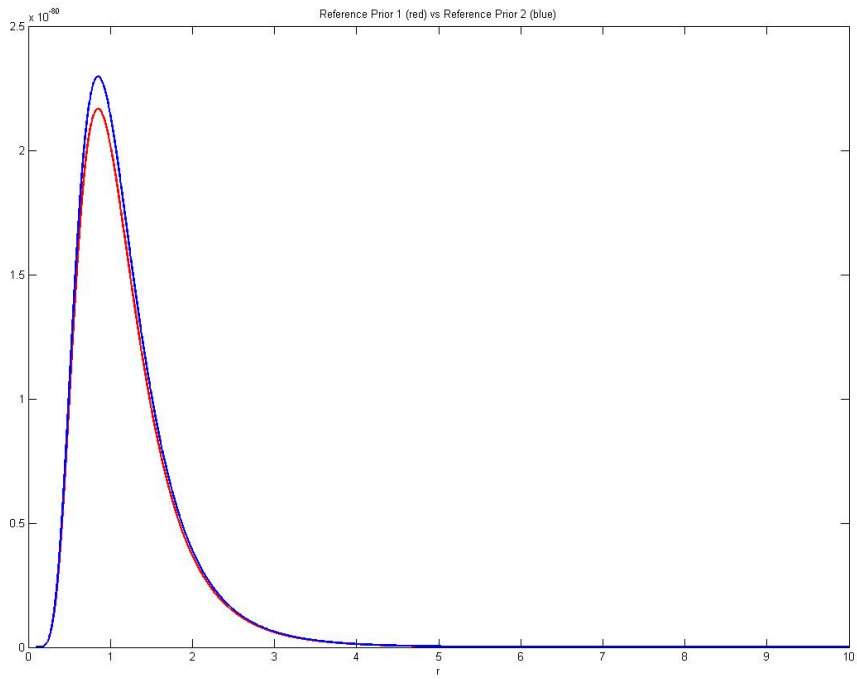
If we use the alternate ordering for parameters as described in equation (6) we find that the joint posterior distribution has the same hierarchical structure, except that

$$P_{R_2}(\tilde{r}|\mathbf{Y}) \propto \prod_{i=1}^{k}\left(\frac{1}{n_i\tilde{r}+1}\right)^{\frac{1}{2}} \times \left(\sum_{i=1}^{k}\frac{n_i}{1+\tilde{r}n_i}\right)^{-\frac{1}{2}} \times \left(\sum_{i=1}^{k}\frac{n_i^2}{(1+\tilde{r}n_i)^2}\right)^{\frac{1}{2}} \times \left[\nu_1 m_1 + \sum_{i=1}^{k}\frac{n_i(\bar{y}_{i\bullet}-\hat{\mu})^2}{n_i\tilde{r}+1}\right]^{-\frac{1}{2}(n-1)} \tag{11}$$

where $0 < \tilde{r} < \infty$.

Figure 1 depicts these two posterior distributions.

Figure 1: Two Reference Priors. Reference Prior 1 = Red; Reference Prior 2 = Blue.



**For further details see Harvey (2012) and van der Merwe, Pretorius and Meyer (2006).*

Theorem 7 The posterior distribution of $\mu + \tau_i$ given σ_e^2 and $\tilde{r}$ is normal with the following mean and variance:

$$E\left\{(\mu+\tau_i) \mid \mathbf{Y}, \sigma_e^2, \tilde{r}\right\} = \frac{\tilde{r}\,n_i}{1+\tilde{r}n_i}\bar{y}_{i\bullet} + \frac{1}{1+\tilde{r}n_i}\hat{\mu}$$

and

$$Var\left\{(\mu + \tau_i) \mid \mathbf{Y}, \sigma_e^2, \tilde{r}\right\} = \sigma_e^2 \left\{\tilde{r} + \frac{1}{1+\tilde{r}n_i}\left(\sum_{i=1}^{k}\frac{n_i}{1+\tilde{r}n_i}\right)^{-1}\right\}$$

From Theorem 7 it follows that

$$\mu + \tau_i + \tfrac{1}{2}\sigma_e^2 \mid \mathbf{Y}, \sigma_e^2, \tilde{r}$$

is distributed normally with mean

$$E\left\{\left(\mu + \tau_i + \frac{1}{2}\sigma_e^2\right) \mid \mathbf{Y}, \sigma_e^2, \tilde{r}\right\} = \frac{\tilde{r}\, n_i}{1+\tilde{r}n_i}\bar{y}_{i\bullet} + \frac{1}{1+\tilde{r}n_i}\hat{\mu} + \frac{1}{2}\sigma_e^2 \tag{12}$$

and variance

$$Var\left\{\left(\mu + \tau_i + \frac{1}{2}\sigma_e^2\right) \mid \mathbf{Y}, \sigma_e^2, \tilde{r}\right\} = \sigma_e^2 \left\{\tilde{r} + \frac{1}{1+\tilde{r}n_i}\left(\sum_{i=1}^{k}\frac{n_i}{1+\tilde{r}n_i}\right)^{-1}\right\} \tag{13}$$

Now, we are interested in the posterior distribution of

$$exp\left(\mu + \tau_i + {}^{\sigma_e^2}/_2\right) \tag{14}$$

for $i = 1, 2, \ldots, k$, in other words, for each worker.

Given σ_e^2 and $\tilde{r}$ we can now simulate from (14) by simulating from a Normal Distribution with mean and variance specified by equations (12) and (13) respectively. Using these results we are able to simulate and test hypotheses for individuals (e.g. individual workers). The results will be presented in later sections.

3.2. Procedure for simulation study

The purpose of this article is to describe the behaviour of the various prior distributions to the setting described earlier. Although detailed descriptions will be given in relevant sections, here we offer a broad description of the simulation of σ_e^2 and $\tilde{r}$ values from the distributions obtained in previous sections, including the final simulation of μ, which will ultimately enable the simulation of quantities such those as defined by equation (14). The simulation procedure can broadly be described as follows:

1. Simulate a value for $\tilde{r}$ using either equation (10) or (11), based on the choice of prior distribution. Since neither (10) nor (11) is a known distribution and cannot be solved in closed form the use of the Rejection method as described in Rice (1995) will be used.

2. Each value of $\tilde{r}$ simulated in the previous step will then be substituted into equation (9) to simulate a value of σ_e^2. In this case the distribution is of a known form, i.e. an Inverse Gamma distribution, and therefore we can simulate σ_e^2 by making use of the fact that:

$$\left\{ \frac{1}{\sigma_e^2} \left[\nu_1 m_1 + \sum_{i=1}^{k} \frac{n_i(\bar{y}_{i\bullet} - \hat{\mu})^2}{n_i\tilde{r} + 1} \right] \right\} \sim \chi^2_{n-1}$$

It follows that a simulated value of σ_e^2 can be obtained from the equation

$$\frac{1}{\chi^2_{n-1}} \left\{ \left[\nu_1 m_1 + \sum_{i=1}^{k} \frac{n_i(\bar{y}_{i\bullet} - \hat{\mu})^2}{n_i\tilde{r} + 1} \right] \right\} = \sigma_e^2$$

Using the values of σ_e^2 and $\tilde{r}$ simulated in the previous steps we can simulate values of μ (if desired) from equation (8). All the desired quantities are based on these variables in some manner.

4. An upper confidence bound and test for the overall mean exposure

In Krishnamoorthy and Guo (2005) one of the primary interests is testing the hypothesis of whether the occupational exposure in an individual (discussed previously in (12) and (13)) or group of workers exceeds a pre-specified or acceptable threshold. If we consider making inferences about the total group, we are interested in the distribution of the Overall Mean Exposure, which for this unbalanced case can be represented as:

$$\mu_x = exp\left\{ \mu + \frac{\sigma_e^2}{2}(\tilde{r} + 1) \right\} = e^{\theta} \tag{15}$$

Now we know from equation (8) that

$$\mu|\mathbf{Y}, \tilde{r}, \sigma_e^2 \sim N\left(\hat{\mu}, \sigma_e^2 \left(\sum_{i=1}^{k} \frac{n_i}{1 + \tilde{r} n_i} \right)^{-1} \right)$$

and therefore

$$\theta = \mu + \frac{\sigma_e^2}{2}(\tilde{r} + 1)|\mathbf{Y}, \tilde{r}, \sigma_e^2 \sim N(E(\theta), Var(\theta))$$

is distributed normally with the following mean and variance:

$$E(\theta) = E\left\{ \mu + \frac{\sigma_e^2}{2}(\tilde{r} + 1)|\mathbf{Y}, \tilde{r}, \sigma_e^2 \right\} = \hat{\mu} + \frac{\sigma_e^2}{2}(\tilde{r} + 1) \tag{16}$$

$$Var(\theta) = Var\left\{\mu + \frac{\sigma_e^2}{2}(\tilde{r}+1)|\mathbf{Y}, \tilde{r}, \sigma_e^2\right\} = \sigma_e^2\left(\sum_{i=1}^{k}\frac{n_i}{1+\tilde{r}n_i}\right)^{-1} \quad (17)$$

Thus, given $\tilde{r}$ and σ_e^2 we simulate θ from a normal distribution with mean and variance defined by (16) and (17) and substitute this into (15). We then repeat this process l $(= 10000)$ times.

Additionally one of the objectives of the work by Krishnamoorthy and Guo (2005) was to test hypotheses as to whether the overall exposure exceeds a certain limit. The authors also simulate the following statistic (and inference regarding this statistic will be made using the Bayesian methodology developed previously):

$$T = \mu + Z_{1-A}\sigma_\tau + \tfrac{1}{2}\sigma_e^2$$

where A is a suitably chosen parameter between 0 and 1 and $Z \sim N(0,1)$ is the density function of the standard normal distribution. Using a specific value of OEL the following hypothesis can be tested:

$$H_0: \mu + Z_{1-A}\sigma_\tau + \tfrac{1}{2}\sigma_e^2 \geq ln(OEL)$$

against the alternative hypothesis

$$H_1: \mu + Z_{1-A}\sigma_\tau + \tfrac{1}{2}\sigma_e^2 < ln(OEL)$$

For example, if our choice of A is 0.05 then essentially we are testing (one-sided) whether at least 5% of the workers have mean exposure levels in excess of the chosen OEL. In practice the OEL is chosen to be a clinically relevant value. The specific choice of OEL is not the primary concern of this research, but primarily a demonstration of the Bayesian methodology.

In order to replicate the methodology of Krishnamoorthy and Guo (2005) from a Bayesian perspective the following simulation study was undertaken for a range of both OEL and A values:

Let $T = \mu + \sigma_e^2\left(\frac{1}{2} + Z_{1-A}\tilde{r}\right)$

We know that $T \mid \mathbf{Y}, \tilde{r}, \sigma_e^2$ is distributed normally with:

$$E\left\{T \mid \mathbf{Y}, \tilde{r}, \sigma_e^2\right\} \sim \hat{\mu} + \sigma_e^2\left(\tfrac{1}{2} + Z_{1-A}\tilde{r}\right)$$

and

$$Var\left\{T|\mathbf{Y}, \tilde{r}, \sigma_e^2\right\} \sim \sigma_e^2\left(\sum_{i=1}^{k}\tfrac{n_i}{1+\tilde{r}n_i}\right)^{-1}$$

This procedure was performed for several choices of OEL (= [130; 140; 150; 160; 170; 180]) and for several choices of A (= [0.1; 0.05; 0.025; 0.001]), as was done in Harvey and van der Merwe (2014).

5. Results from the simulation study

Using the methodology derived previously a simulation study was conducted to simulate 10000 observations for each particular type of analysis. Using the unbalanced data provided in Table 2 we were able to simulate observations relating to occupational exposure in the workplace. Since two Reference priors were derived the simulations were repeated for each of these Reference priors. The results are presented in the following sections.

5.1. Results: Individual worker means

As mentioned it was possible to simulate observations from the posterior distribution for each of the 15 workers, using both Reference priors. Selected results will be shown here for the purposes of illustration. Simulation results for all 15 workers can be found in Harvey (2012).

From Tables 3 and 4 we can see that the results from the first and second Reference priors are comparable, with no large differences between the various Reference priors.

The effect of "unbalancing"has largely been minimized. For example, workers 4 and 11 both had comparable mean exposure levels (55.98 and 55.7 respectively), but were at the two extremes (in this hypothetical data set) with regards to unbalancing (worker 4 had 10 exposure observations, while worker 11 only had 6 observations). It is interesting to note though that in both cases the probability of exceeding the OEL of 130 was 0.0001 (based on 10000 simulated observations). It thus appears that the Bayesian methodology is stable with regards to unbalanced data, particularly at a worker-specific level.

Table 3: Simulation Summary Results: Reference Prior 1.

Worker	$P(\mu_{exposure} > 130)$	90% CI		95% CI		Mean	Median	Mode
		Low	High	Low	High			
Worker 4	0.0001	50.34	72.64	47.29	76.177	61.62	61.71	63.25
Worker 11	0.0001	51.48	74.27	48.33	77.76	63.17	63.19	61.75

Table 4: Simulation Summary Results: Reference Prior 2.

Worker	$P(\mu_{exposure} > 130)$	90% CI		95% CI		Mean	Median	Mode
		Low	High	Low	High			
Worker 4	0.0002	50.05	72.90	47.35	76.42	61.59	61.61	62.75
Worker 11	0.0001	51.24	74.07	48.18	77.50	63.15	63.28	64.25

5.2. Results: Overall mean exposure

The next result relates to the overall mean exposure, that is the exposure of the group of 15 workers as a whole. The results in Table 5 and Figures 2 and 3 were obtained for the two Reference prior distributions (the relevant information for each histogram is displayed in the Table 5).

Figure 2: Overall Mean Exposure: Reference Prior 1

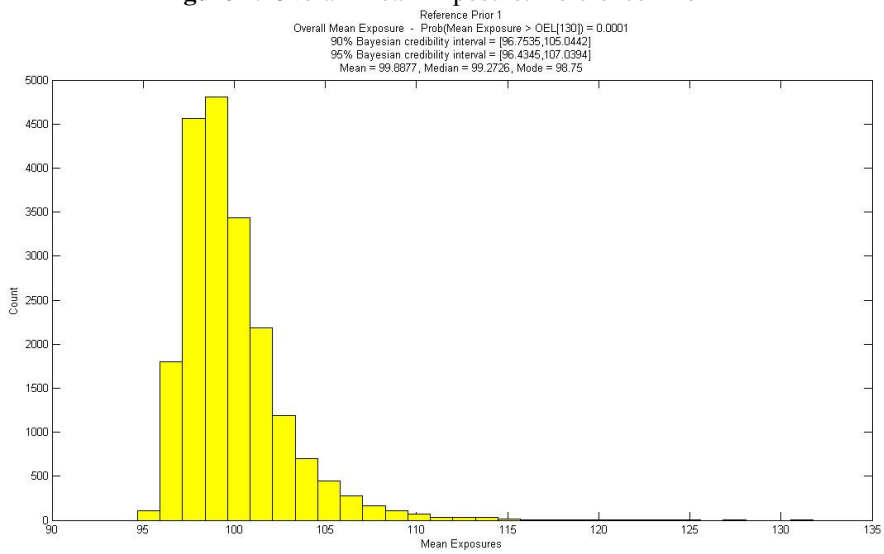


Table 5: Simulation Summary Results of Overall Mean Exposure

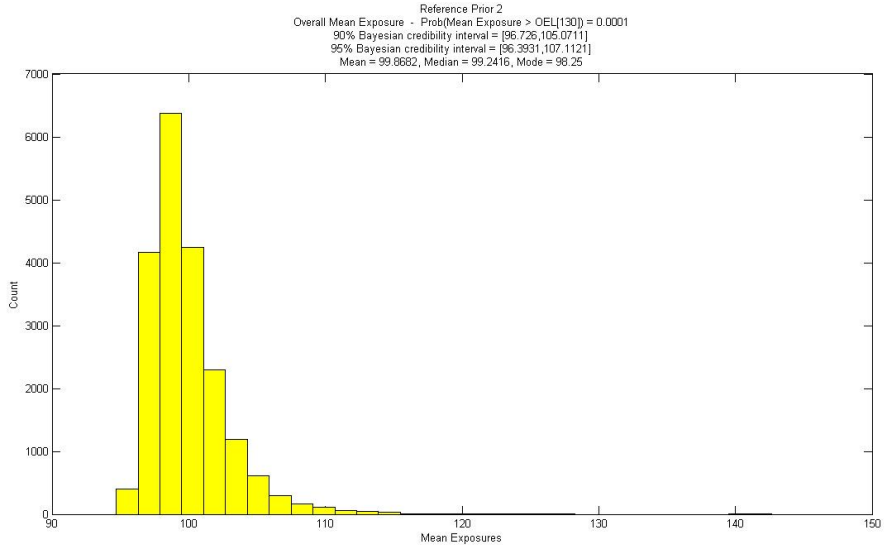
All Workers	$P(\mu_{exposure} > 130)$	90% CI		95% CI		Mean	Median	Mode
		Low	High	Low	High			
Reference Prior 1	0.0001	96.75	105.04	96.43	107.04	99.89	99.27	98.75
Reference Prior 2	0.0001	96.73	105.07	96.39	107.11	99.86	99.24	98.25

The results from Table 5 as well as Figures 2 and 3 are based on 20000 simulations. We see very little difference between the two Reference prior distributions.

5.3. Results: Hypothesis testing

Lastly, and perhaps most importantly, Krishnamoorthy and Guo (2005) tested hypotheses regarding the group of workers using the following measure:

Figure 3: Overall Mean Exposure: Reference Prior 2



$$T = \mu + Z_{1-A}\sigma_\tau + \tfrac{1}{2}\sigma_e^2$$

where A is a suitably chosen parameter between 0 and 1 and Z denotes the standard normal distribution. So, for example, if our choice of A is 0.05 then essentially we are testing (one-sided) whether at least 5% of the workers have mean exposure levels in excess of the chosen OEL.

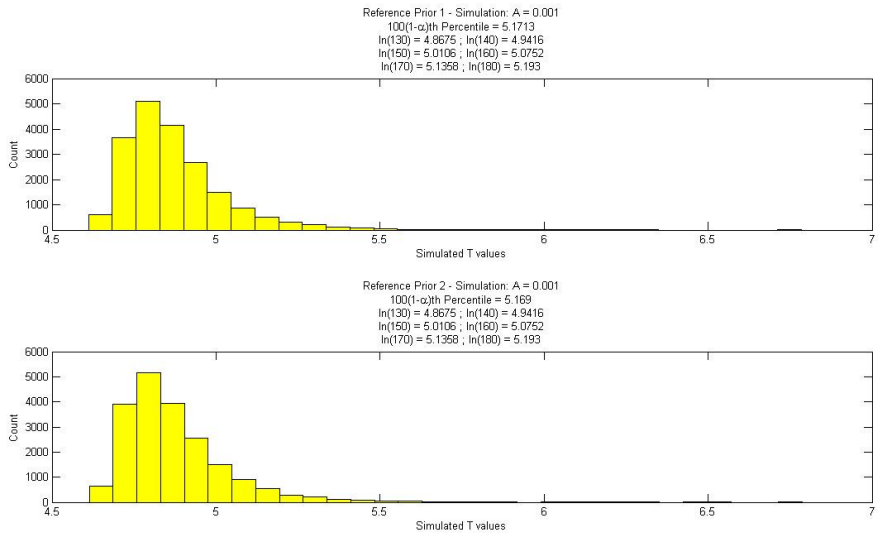
Several different values of A were chosen in addition to several different OEL limits. The results in Figure 4 are once again produced for both Reference prior distributions.

What is interesting to note is that compared to results obtained in Harvey and van der Merwe (2014), of which this article is merely an extension to the unbalanced case, the distributions in the unbalanced case are more skewed, with longer tails. From Figure 4 we can see that only 0.1% or more of workers had occupational exposure levels in excess of 5.1713 and 5.169 (99.9th percentile) respectively for Reference priors 1 and 2, which corresponds to an OEL of roughly 164.

6. Conclusion

In this article the usefulness of the Bayesian methodology to the proposed setting of occupational exposure data was examined, specifically for the case where there are an unequal number of observations for each worker. The one-way random effects model was adapted to account for unbalanced data using the chosen prior distributions. One of the advantages of the Bayesian model is that it is

Figure 4: Hypothesis Testing for Reference Priors 1 (above) 2 (below): $A = 0.001$



$100\,(1-\alpha)\,th$ *percentile = 5.1713 (Prior 1)*

$100\,(1-\alpha)\,th$ *percentile = 5.1690 (Prior 2)*

able to model results for individual workers and not simply for an unknown future worker. Only a few non-informative prior distributions have been derived in this article, but they do by no means represent an exhaustive list. The derivation and comparison of all possible prior distributions was not an objective of this research. However, the derivation and application of other non-informative priors could be used to refine the analysis and improve performance. Ultimately, if subjective prior information is available this could lead to significant improvements in prediction of future exposure for individual workers as well as for groups of workers.

References

BERGER, J. O. AND BERNANRDO, J. M. (1992). *On the development of the reference prior method*. Claredon Press, Oxford, pp. 35–60.

DATTA, G. S. AND GHOSH, J. K. (1995). On priors providing frequentist validity of bayesian inference. *Biometrika*, **82**, 37–45.

HARVEY, J. (2012). *Bayesian inference for the lognormal distribution*. Unpublished Ph.D dissertation, Bloemfontein, South Africa.

HARVEY, J. AND VAN DER MERWE, A. (2014). Modelling occupational exposure using a random effects model: a bayesian approach. *South African Statistical Journal*, **48**, 61–71.

HEERDERIK, D. AND HURLEY, F. (1994). Occupational exposure assessment: Investigating why exposure measurements vary. *Applied Occupational and Environmental Hygiene*, **9**, 71.

KRISHNAMOORTHY, K. AND GUO, H. (2005). Assessing occupational exposure via the one-way random effects model with unbalanced data. *Journal of Statistical Planning and Inference, Elsevier*, **128**, 219–229.

KRISHNAMOORTHY, K. AND MATHEW, T. (2002). Assessing occupational exposure via the one-way random effects model with balanced data. *Journal of Agricultural, Biological and Environmental Statistics, American Statistical Association and International Biometric Society*, **3**, 440–451.

LYLES, R. H., KUPPER, L. L., AND RAPPAPORT, S. M. (1997a). Assessing regulatory compliance of occupational exposures via the balanced one-way random effects anova model. *Journal of Agricultural, Biological, and Environmental Statistics*, **2**, 64–86.

LYLES, R. H., KUPPER, L. L., AND RAPPAPORT, S. M. (1997b). Lognormal distribution based exposure assessment method for unbalanced data. *Annals of Occupational Hygiene*, **41**, 63–76.

RAPPAPORT, S. M., KROMHOUT, H., AND SYMANSKI, E. (1993). Variation of exposure between workers in homogeneous exposure groups. *American Industrial Hygiene Association Journal*, **54**, 654–662.

RICE, A. (1995). *Mathematical Statistics and Data analysis*. Duxbury Press, Belmont, California.

VAN DER MERWE, A. J., PRETORIUS, A., AND MEYER, J. (2006). Bayesian tolerance intervals for the unbalanced one-way random effects model. *Journal of Quality Technology*, **38**, 280–293.

Manuscript recieved, 20XX-01-01, revised, 20XX-01-02, accepted, 20XX-01-03.