
DEPARTAMENT DE MATEMÀTIQUES

Doctorat en Matemàtiques

Curs Acadèmic 2014–2015

TESI DOCTORAL:

# Advanced Statistical Methods for Cytogenetic Radiation Biodosimetry

*Candidat:* Manuel Higueras Hernáez

*Directors de la tesi:*
Prof. Pere Puig Casado

Dr. Elizabeth Ann Ainsbury

Prof. Kai Rothkamm

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

The subject of this thesis was to develop novel solutions for statistical analysis of cytogenetic radiation biodosimetry data to correctly quantify uncertainty associated with dose estimates in exposed individuals. The chromosomal aberrations induced by ionising radiation can be explained in a probabilistic manner, leading to regression models for count data. The work done has led to distinct improvements in dosimetry and uncertainty calculation, which has potential impacts for both routine and emergency dosimetry.

In particular, this research focuses on cytogenetic dose estimation in the Bayesian framework. The chromosomal aberration data can be overdispersed and/or zero–inflated in several irradiation scenarios. These phenomena imply the management of compound Poisson and zero–inflated distributions.

The research presented encompasses the literature review of the cytogenetic Bayesian methods, the development of new Bayesian models for cytogenetic dose estimation, the software implementation for some of these systems, the empirical analysis of cytogenetic data, and the case–study of cytogenetic data for the comparison of different models to fit cytogenetic dose–response curves.

## 1.1 Funding

The work carried out at Public Health England Centre for Radiation, Chemical and Environmental Hazards was funded by the National Institute for Health Research. The work carried out at Universitat Autònoma de Barcelona was funded by the grant MTM2012-31118 by the Ministry of Economy and Competitiveness, and by the grant UNAB10-4E-378 co–funded by ERDF 'A way of making Europe'.

## 1.2 Outline of the thesis

The thesis format is *Thesis by publication.* Chapter 1 corresponds to the introduction, presenting biological dosimetry and the statistical limitations in cytogenetic dose estimation. Then, Chapter 2 is an overall presentation of the results and discussion of their novelty and relevance in the context of cytogenetic biodosimetry. Chapter 3 presents the general conclusions, reviewing and combining these results, and expecting the further research.

Chapters 4, 5, 6, 7 and 8 are the five scientific publications, Ainsbury *et al.* 2014 [4], Higueras *et al.* 2015a [39], Moriña *et al.* 2015a [65], Higueras *et al.* 2015b [40] and Oliveira *et al.* 2015 [70] respectively, that constitute the main part of this thesis.

Finally, Appendix A, Moriña *et al.* 2015b, shows a submitted scientific article which is complementary material in this thesis.

## 1.3 Biological dosimetry

Radiation exposure through radiotherapy treatments, nuclear accidents or terrorist acts are major current concerns for our global society. Biological dosimetry, IAEA 2011 [42], relies on quantifying the amount of damage induced by radiation at a cellular level, for example counting dicentrics, centric rings, or micronuclei. The frequency of these chromosome aberrations is an established biological indicator of radiation dose received. The quantification of the radiation dose absorbed is essential for predicting the derived health consequences in irradiated patients in the short, medium and long term. Dose estimation from chromosome damage biomarkers is necessary despite the physical measurement of dose, because it also takes into account the inter–individual variation in susceptibility. Biological sampling and dose estimation is sometimes the only method to be absolutely sure of the dose to the individual, for example classified radiation workers sometimes do not wear badges or workers who are not wearing them because they are never supposed to be near anything radioactive.

The ideal dosimeter, which obviously does not exist, would be specific to radiation, present low background, have low donor variability, present low doubling dose, dose response calibration, have persistent effect, allow ease of sampling, rapid analysis, low cost, and work as a 'risk meter'.

In cytogenetic biodosimetry irradiated blood samples are analysed to score chromosomal aberrations. The samples could be *in vitro*, they are taken from laboratory experiments to study the yield of chromosomal aberrations after fixed absorbed doses, or *in vivo*, the samples are taken from radiotherapy patients or nuclear accident victims to analyse the amount of absorbed dose and to predict their future health consequences. When the whole body is exposed in a homogeneous manner the scenario is whole body irradiation; by contrast, if the exposure is located in a fraction of the body, the scenario is partial body irradiation (PBI). Depending on the nature of the radiation, the linear energy transfer (LET) describes how much energy an ionising particle transfers to the material transverse per unit distance; e.g., $\gamma$–rays and $\alpha$–particles are considered low– and high–LET, respectively, IAEA 2011 [42]. Depending on the nature of the radiation chromosomal damage is caused by discrete energy deposition traces in time and space which are produced by ionising radiation. It is shown that low LET radiation can produce localised clusters of ionisations within a single electron track, meanwhile high LET radiation produces a larger number of ionizations that are close in spatial extent, IAEA 2011 [42]. Figure 1.1 displays the relation between the LET and the relative biological effectiveness (RBE), the ratio of doses producing equal biological effect; source Hall and Giaccia 2012 [36].

The most studied chromosomal aberrations are the dicentrics, which are the interchange between the fragments of two separate chromosomes, thus they are

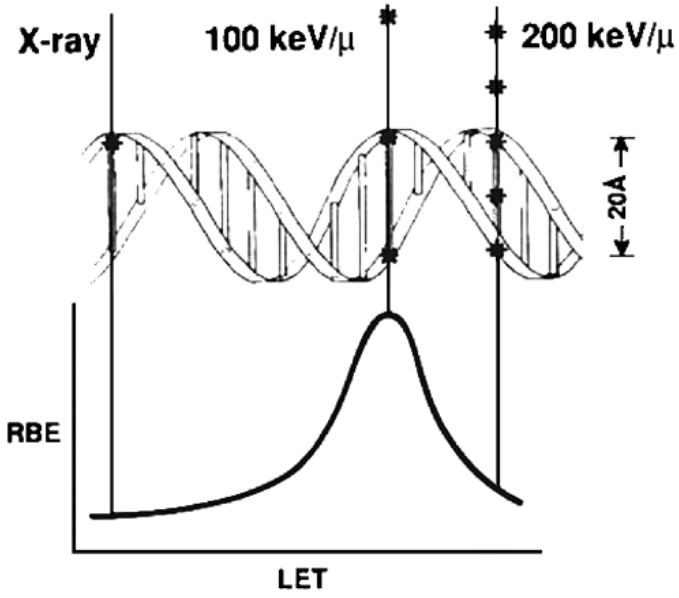


Figure 1.1: Diagram showing LET against RBE for cell killing, mutagenesis, or oncogenic transformation, Hall and Giaccia 2012 [36]. Note that 100 keV/$\mu$m has the greatest RBE.

chromosomes with two centromeres. Ring chromosomes, or centric rings, are also analysed, which are an exchange between two breaks on separate arms of the same chromosome and are also accompanied by acentric fragments (chromosomes without centromere). Another typical cytogenetic assay is based on micronuclei, which are lagging chromosomal fragments or whole chromosomes at anaphase that are not included in the nuclei of daughter cells. More emerging assays, not covered in this thesis are the H2AX, the H2A variant produced by the generation of double–stranded breaks, and the effects of ionising radiation on gene expression.

Figures 1.2, 1.3 and 1.4 show respectively the images of centric rings, dicentric chromosomes and micronuclei; source: International Atomic Energy Agency (IAEA) manual, IAEA 2011 [42].

Biological dosimetry has progressed from an initial research idea in the 1960s, to a varied and active field and has been based on the analysis of dicentric chromosome aberrations, which is the most frequent technique and for a long time was the only one. Since then, this field has progressed to become a standard in most of the radiation protection programmes and its application in radiation exposures cases has shown its importance. Chromosomal aberrations are a dosimeter providing very important information for nuclear or radiological accidents; all this biodosimetry information is compounded to get a trustable appraisal of the cases, IAEA 2011 [42]. Today there are a number of new and emerging assays being tested as markers of both radiation exposure and effect (e.g. individual cancer susceptibility).

Biological dosimetry has been applied in accidents like the Chernobyl disaster

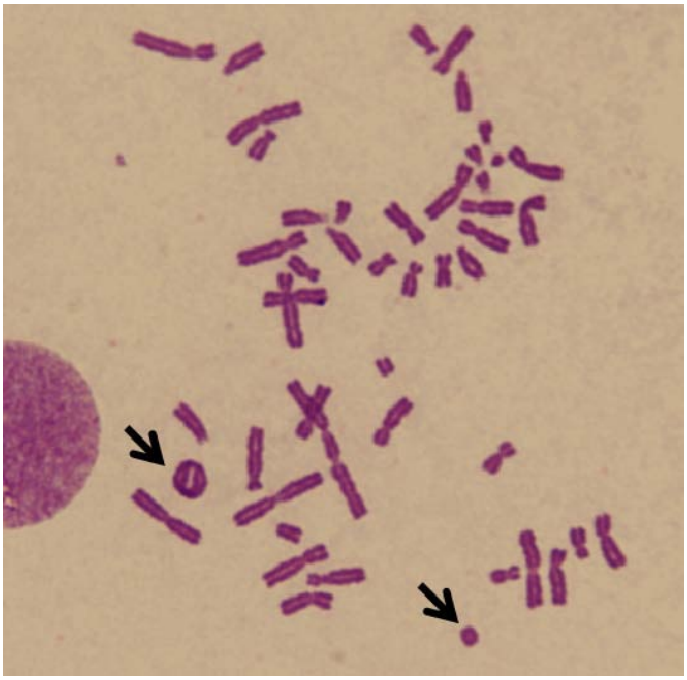

Figure 1.2: A metaphase spread with two rings (arrowed).

and most recently in the Fukushima occurrence (Suto *et al.* 2013 [91]), and for radiotherapy treatment patients, e.g. Serna *et al.* 2013 [86].

## 1.4 Statistical considerations

The number of chromosomal aberrations induced by ionising radiation among cells produces count data. The Poisson distribution is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event. If the expected number of occurrences in this interval is $\lambda$, then the probability that there are exactly $k$ occurrences ($k = 0, 1, 2, \ldots$) is $\lambda^k \mathrm{e}^{-\lambda}/k!$, Johnson *et al.* 2005 [44]. In cytogenetic biodosimetry $\lambda$ is related to the absorbed dose, i.e. the higher dose, the greater number of chromosomal aberrations per cell.

The IAEA manual, IAEA 2011 [42], states the Poisson distribution as the most widely recognised and commonly used probability distribution for cytogenetic data analysis. In fact, for dicentrics assay, irradiation with X– or $\gamma$–rays, low LET radiation, produces a distribution of damage which is very well represented by the Poisson distribution, IAEA 2011 [42].

A Poisson distribution with mean $\lambda$ has a variance equals to $\lambda$, and consequently the dispersion index, the ratio of the variance to the mean, is 1 (equidispersion), see e.g. Johnson *et al.* 2005 [44]. Distributions which dispersion indexes are greater (lower) than 1, are overdispersed (underdispersed).

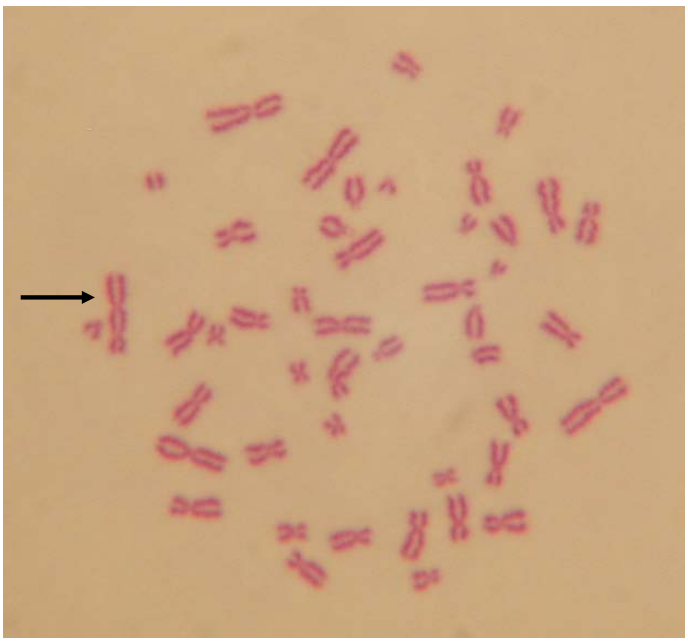

Figure 1.3: A dicentric chromosome (arrowed).

Given a sample $y = \{y_1, y_2, \ldots, y_n\}$ of counts Poisson distributed, representing here the number of chromosomal aberrations in $n$ blood cells, the lower and upper 95% confidence limits of the population mean are (see e.g. Johnson *et al.* 2005 [44])

$$y_L = \frac{\chi^2_{2X,0.025}}{2n}, \quad y_U = \frac{\chi^2_{2(X+1),0.975}}{2n},$$

respectively, where $\chi^2_{m,p}$ is the quantile function corresponding to a lower tail probability $p$ of the $\chi^2$–squared distribution with $m$ degrees of freedom, and $X$ is the sample sum of $y$.

To reject or not reject the Poisson assumption an informative test based on the property of equidispersion is performed, the $u$–test, a normalised unit of the dispersion unit, IAEA 2011 [42]. The $u$–test remains:

$$u = (s^2/\bar{y} - 1)\sqrt{\frac{n-1}{2(1-1/X)}},$$

where $\bar{y}$ and $s^2$ are respectively the sample mean and variance of $y$. Those $u$ values higher (lower) than (-)1.96 indicate overdispersion (underdispersion), with a significance level of 5%, Rao and Chakravarti 1956 [80].

The first step to estimate the absorbed dose is to do a calibration, that is, to construct a dose–response curve. The usual approach is to irradiate in the laboratory several blood samples from a healthy donor with several doses. For low LET radiation, 10 or more doses should be used in the range 0.25-5.0 Gy, IAEA 2011 [42]. The construction of the dose–response curve can be experimentally difficult according to the kind of radiation.

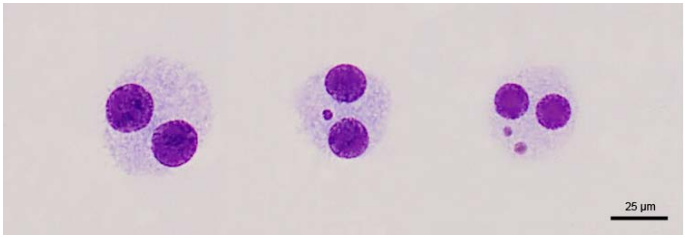


Figure 1.4: Examples of binucleated cells without (left) and with 1 (middle) and 2 (right) micronuclei.

There is a very strong empirical evidence that the yields (population mean) of chromosome aberrations $Y$ are related to the absorbed dose $D$ by the linear–quadratic equation, $Y = C + \alpha D + \beta D^2$, IAEA 2011 [42] and Hall and Giaccia 2012 [36]. In some scenarios (high LET radiation, for instance) the $\alpha$–term becomes large and eventually the $\beta$–term becomes biologically less relevant and also statistically 'masked' and the dose response is approximated by the linear equation, $Y = C + \alpha D$, IAEA 2011 [42].

Classical linear regression is not appropriate to fit the dose–response curve because the responses are the mean of Poisson counts and the assumption of equality of variances (homoscedasticity) is violated due to the biological process which lead to a dependence of the variance on the dose. Generalised Linear Models provides a unified approach for analysing the relationship between the response, e.g. Poisson and negative binomial (NB), and several explanatory variables, that can be numerical (regression problems) or categorical (ANOVA problems), McCullagh and Nelder 1989 [60]. In radiation biodosimetry, the responses are typically the number of dicentrics in each cell and they are classically considered Poisson distributed with mean $Y_j = C + \alpha D_j + \beta D_j^2$ (identity link) where $j = 1, 2, \ldots, m$ represents each analysed cell (a total of $m$ cells analysed).

Once $C$, $\alpha$, and $\beta$ are fitted by maximum likelihood estimation (MLE), the dose estimation leads to an inverse regression problem, because the exposed doses in the dose–response experiment are chosen by the laboratory researches, i.e. the dose is not a random variable in the calibration data collecting process. Supposing now that we have observed the above sample of chromosomal aberrations, $y$, because the MLE of the expected value for a count sample under the Poisson assumption is the sample mean, see e.g. Johnson *et al.* 2005 [44], the estimated absorbed dose is the solution $D$ for the equation $\bar{y} = C + \alpha D + \beta D^2$, Merkle 1983 [61], i.e.

$$D = \frac{-\alpha + \sqrt{\alpha^2 - 4\beta(C - \bar{y})}}{2\beta},$$

for the linear–quadratic model. For the linear model, $\beta = 0$, then

$$D = \frac{\bar{y} - C}{\alpha}$$

The 95% confidence region of the dose–response curve has the form (Merkle

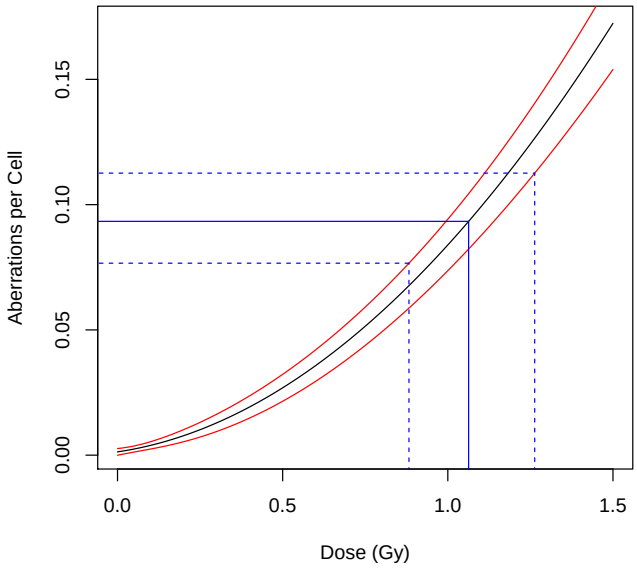


Figure 1.5: Fitted calibration curve (solid black line), its 95% confidence region curves (solid red lines) and the graphical solutions of the three equations (blue lines).

1983 [61])

$$Y_{\pm} = C + \alpha D + \beta D^2 \pm R \cdot \sqrt{\text{var}(C) + \text{var}(\alpha)D^2 + \text{var}(\beta)D^4 + 2\text{cov}(C,\alpha)D + 2\text{cov}(C,\beta)D^2 + 2\text{cov}(\alpha,\beta)D^3},$$

where $R^2$ is the 95% percentile of a chi-square distribution with 3 degrees of freedom (linear–quadratic model), or 2 (linear model; $\beta, \text{var}(\beta), \text{cov}(C,\beta), \text{cov}(\alpha,\beta) = 0$) degrees of freedom (df). For df $= 3$, $R = 2.795$ and for df $= 2$, $R = 2.448$. The variances and covariances of the parameters are obtained from the Poisson regression outputs. The 95% confidence lower and upper limits of the dose estimation are the solutions in $D$ of the equations $y_L = Y_+$ and $y_U = Y_-$ respectively, Merkle 1983 [61]. However, many researchers use the simplified version of the $Y_{\pm}$ curves, omitting the information of the covariances, i.e. all covariances are 0. Figure 1.5 shows the graphical solution of these equations in the example on Section 1.5.1.

The details presented so far represent the current accepted practice for the statistical analysis of whole body radiation induced chromosomal aberrations. Sections 1.4.1 and 1.4.2 introduce details of the limitations and proposed alternatives in the cytogenetic literature which led to the research of this thesis. Section 1.5.1 shows an application of this methodology and Section 1.5.1 shows its incorrectly measurement of the uncertainty through some simulations.

### 1.4.1 Limitations

For the current cytogenetic analysis methods the scenario upon which the methods (calibration curves, dose estimation) are based is an acute, whole body, homogeneous, external irradiation of a known low LET source, with dose in the region of between 0.1 and 5 Gy. The appearance of dicentric chromosomes in this scenario is very well represented by a Poisson process, IAEA 2011 [42]. In contrast, real life exposures present delayed blood sampling, protraction or fractionation, chronic exposures, inhomogeneity, very high or very low doses, high LET radiation, internally deposited radionuclides, mass casualty scenarios, and inter–scorer, –lab, –assay variation and other confounders, Vinnikov *et al.* [94]. For instance, the dicentric assay tends to produce overdispersed data (variance higher that the mean) for high LET radiation exposition, i.e. it is not under the Poisson assumption where the variance and the mean are equal. This is the case for the dicentric assay after $\alpha$–particles or neutrons irradiation, IAEA 2011 [42]. These overdispersed patterns are properly described by compound Poisson distributions, which provide alternative models which have been tested in the field, e.g. Virsik and Harder 1981 [96] and Puig and Barquinero 2011 [78].

The commonly used compound Poisson distributions in biodosimetry are the Neyman A (NA), Virsik and Harder 1981 [96], the NB, Brame and Groer 2002 [14], and more recently the family of $r^{\text{th}}$–order univariate Hermite distributions, Puig and Barquinero 2011 [78]. The $2^{\text{nd}}$–order univariate distribution is the Hermite distribution, Kemp and Kemp 1965 [47]. A random variable $Y$ follows a compound probability distribution if it can be represented by

$$Y = \sum_{i=1}^{N} \xi_i,$$

where $N$ is a count data random variable and $\xi_1, \xi_2, \ldots$ are independent, identically distributed random variables that are also independent of $N$. In the case where $N$ is Poisson, $Y$ is said to follow a compound Poisson distribution. The distribution of $\xi_i$ is called the generalizing distribution. In particular when the distribution of $\xi_i$ is Poisson, the distribution of $Y$ is a NA, when $\xi_i$ follows a logarithmic distribution, $Y$ is NB distributed, and when $\xi_i$ is distributed as a binomial with a number of trials equal to 2, then $Y$ follows a Hermite distribution, Johnson *et al.* 2005 [44]. The compound Poisson process considers that the number of particles traversing a cell nucleus follows a Poisson distribution, and for each particle, there is a probability (the generalising distribution) to produce a count of aberrations, Puig and Barquinero 2011 [78]. The MLE of the population mean for NA, NB and Hermite distributions is the sample mean, Johnson *et al.* 2005 [44].

Partial body exposures produce overdispersed distributions due to a 'zero–inflation' mechanism, represented by mixed Poisson distributions, Sasaki 2003 [83]. In these scenarios the distributions of chromosome aberrations are zero–inflated, $\omega + (1-\omega)Y$, where $\omega$ represents the non-irradiated cells and $(1-\omega)Y$ represents the irradiated cells. When $Y$ is Poisson distributed, this leads to the zero–inflated Poisson distribution. This distribution has one more parameter $\omega$, the proportion of extra zeros.

Let $Z = \omega + (1-\omega)Y$ where $Y$ is Poisson distributed and $0 < \omega < 1$, the

probability that there are exactly $k$ occurrences ($k = 0, 1, 2, \ldots$) is equal to

$$\mathrm{P}(Z = 0) = \omega + (1 - \omega)\mathrm{e}^{-\lambda}$$
$$\mathrm{P}(Z = k) = \frac{(1 - \omega)\lambda^k}{\mathrm{e}^{\lambda} k!}, \text{ for } k > 0$$

The population mean is just $(1 - \omega)\lambda$, Johnson *et al.* 2005 [44]. The MLE of $\lambda$ is obtained solving numerically the equation,

$$\lambda(n - n_0) = (1 - \mathrm{e}^{-\lambda})X \tag{1.1}$$

where $n_0$ is the number of zero counts and $X$ is the sample sum; the estimate of $\omega$ becomes (Johnson *et al.* 2005 [44])

$$\omega = \frac{n_0/n - \mathrm{e}^{-\lambda}}{1 - \mathrm{e}^{-\lambda}}. \tag{1.2}$$

An estimate of the variance of $\lambda$ can be obtained using the expression (Johnson *et al.* 2005 [44]),

$$\mathrm{V}(\lambda) \approx \frac{\lambda(1 - \mathrm{e}^{-\lambda})^2}{(n - n_0)(1 - \mathrm{e}^{-\lambda} - \lambda \mathrm{e}^{-\lambda})}.$$

From here 95% confidence limits are found in the usual way,

$$y_{U/L} = \hat{\lambda} \pm 1.96\sqrt{\mathrm{V}(\hat{\lambda})},$$

and the absorbed dose estimation 95% confidence limits are calculated like in the whole body irradiation scenarios. It is important to remark that $\omega$ is not the 'size of the irradiated body'. To estimate this is necessary to take into account a correction for the effects of interphase death and mitotic delay, IAEA 2011 [42], this is $d_0$ which is the 37% cell survival dose, with experimental evidence to be between 2.7 and 3.5 Gy, IAEA 2011 [42]. The fraction of the body irradiated is calculated by the following formula (IAEA 2011 [42]),

$$F = \frac{(1 - \omega)\mathrm{e}^{D/d_0}}{\omega + (1 - \omega)\mathrm{e}^{D/d_0}}. \tag{1.3}$$

Note that $\mathrm{e}^{-D/d_0}$ represents the proportion of cells which have survived. This is the so–called Dolphin's method. Section 1.5.2 shows an application of this methodology.

### 1.4.2 The Bayesian approach

Several authors have suggested that a Bayesian approach to uncertainty estimation may be suitable for analysing cytogenetic data. The methods of Bayesian inference provide a consistent framework for modelling and predicting these uncertain conditions. The Bayesian analysis framework focuses on the identification of several probability distributions involved in the process. Prior information is used in combination with experimental results to infer probabilities or the likelihood that a hypothesis is true. A Bayesian approach is highly applicable to ionising radiation dosimetry data. It has been shown that this approach improves both the accuracy and assurance of radiation dose estimates. Bayesian

framework allows the investigator to consider prior knowledge surrounding a system, and this type of data is often available in biodosimetry. A number of authors have begun to apply Bayesian methodology to analysis of cytogenetic data for the purposes of biodosimetry. Groer and Pereira 1987 [33] were the first to investigate the use of Bayesian models in chromosome dosimetry neutron exposure. A review of Bayesian methods in biodosimetry can be found in Ainsbury *et al.* 2013.

The Bayes' theorem in its continuous version establishes

$$\mathrm{P}(\Theta|y) = \frac{\mathrm{L}(\Theta|y)\mathrm{P}(\Theta)}{\int \mathrm{L}(\Theta|y)\mathrm{P}(\Theta)\mathrm{d}\Theta},$$

where $\Theta$ is the continuous parameter set, $y$ is the observed data set, $\mathrm{L}(\Theta|y)$ is the likelihood function, $\mathrm{P}(\Theta)$ is the prior probability density function of $\Theta$ and $\mathrm{P}(\Theta|y)$ is the posterior prior probability density function of $\Theta$ given data $y$. See Christensen *et al.* 2011 [16] and Sivia and Skilling 2006 [88], for instance.

The Bayesian framework considers that parameters are random variables, in the biological dosimetry case, the calibration coefficients ($C$, $\alpha$, $\beta$) and the absorbed dose $D$. The prior knowledge surrounding a system can be considered in the Bayesian inference.

The calibrative density is the solution of the Bayesian inverse regression problem, and its derivation is detailed in Section 4.4 which is the Appendix of Ainsbury *et al.* 2014 [4].

## 1.5 Examples of classical dose estimation

The classical methodologies to estimate the absorbed dose in cytogenetic biodosimetry presented in Sections 1.4 and 1.4.1 are extended here by reproducing two practical examples.

### 1.5.1 Whole body irradiation

Barquinero *et al.* 1995 [9] studied the dose–response curve for $\gamma$–rays induced dicentrics. shows the cell distribution of dicentrics for 11 different doses. This data set (Table 1.1) is based on blood samples which were irradiated *in vitro* using a cobalt source (Theratron–780) at a dose rate of 26.95 cGy/min simulating acute whole body exposure. The Poisson assumption is considered and supported for most of the samples, only the 0.25 and 2 Gy doses does the $u$–test value exceed 1.96 (overdispersion).

The 1 Gy sample is taken from the calibration data to be used as test data. The fitted calibration curve (omitting the 1 Gy sample) for a linear–quadratic model, $Y = C + \alpha D + \beta D^2$, results

$$\hat{C} = 1.312 \cdot 10^{-3}, \quad \hat{\alpha} = 1.971 \cdot 10^{-2}, \quad \hat{\beta} = 6.288 \cdot 10^{-2};$$

$$\Sigma = \begin{pmatrix} 2.281 & -10.734 & 4.407 \\ -10.734 & 279.772 & -146.915 \\ 4.407 & -146.915 & 158.404 \end{pmatrix} \cdot 10^{-7};$$

and $R = 2.795$. This is all the necessary information to build the curves $Y$ and $Y_{\pm}$. Figure 1.7 shows the plot of the fitted dose–response curve and the observed means of the calibration data.

Table 1.1: Dicentric distribution within cells, sample means and dispersion coefficients, and $u$ values for each distribution. Test data in italics.

| | Number of dicentrics | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Dose (Gy) | 0 | 1 | 2 | 3 | 4 | 5 | $\bar{y}$ | $d$ | $u$ |
| 0.00 | 4992 | 8 | | | | | 0.002 | 0.999 | -0.075 |
| 0.10 | 4988 | 14 | | | | | 0.003 | 0.997 | -0.135 |
| 0.25 | 1987 | 20 | 1 | | | | 0.011 | 1.080 | 2.610 |
| 0.50 | 1947 | 55 | | | | | 0.027 | 0.973 | -0.861 |
| 0.75 | 1736 | 92 | 4 | | | | 0.050 | 0.950 | -1.514 |
| *1.00* | *1064* | *99* | *5* | | | | *0.093* | *0.999* | *-0.017* |
| 1.50 | 474 | 76 | 12 | | | | 0.178 | 1.064 | 1.077 |
| 2.00 | 251 | 62 | 16 | 3 | | | 0.310 | 1.179 | 2.311 |
| 3.00 | 104 | 72 | 15 | 2 | | | 0.560 | 0.834 | -1.638 |
| 4.00 | 35 | 41 | 21 | 4 | 2 | | 1.000 | 0.882 | -0.844 |
| 5.00 | 11 | 19 | 11 | 9 | 6 | 3 | 1.814 | 1.150 | 0.811 |

Table 1.2: Distribution of dicentrics plus rings within cells, sample means and dispersion coefficients, and $u$ values for each distribution.

| | Number of Dic+CR | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Dose (Gy) | 0 | 1 | 2 | 3 | 4 | 5 | $\bar{y}$ | $d$ | $u$ |
| 0.00 | 8802 | 9 | | | | | 0.001 | 0.999 | -0.064 |
| 0.10 | 5034 | 14 | | | | | 0.003 | 0.997 | -0.134 |
| 0.25 | 1968 | 36 | 1 | | | | 0.019 | 1.034 | 1.097 |
| 0.50 | 1942 | 69 | 1 | | | | 0.035 | 0.993 | -0.212 |
| 0.75 | 1503 | 103 | 1 | | | | 0.065 | 0.954 | -1.301 |
| 1.00 | 1185 | 105 | 2 | | | | 0.084 | 0.953 | -1.198 |
| 1.50 | 582 | 93 | 7 | | | | 0.157 | 0.975 | -0.456 |
| 2.00 | 303 | 88 | 11 | 1 | | | 0.280 | 0.970 | -0.430 |
| 3.00 | 105 | 72 | 25 | 2 | 1 | | 0.644 | 0.921 | -0.799 |
| 4.00 | 71 | 73 | 41 | 16 | 3 | | 1.054 | 0.946 | -0.545 |
| 5.00 | 31 | 66 | 64 | 24 | 13 | 2 | 1.640 | 0.791 | -2.090 |

The 1 Gy sample presents a total of 109 dicentrics in 1168 cells analysed, following notation in Section 1.4 $X = 109$ and $n = 1168$. The sample mean is $\bar{y} = 0.093$ and the lower and upper confident limits are $y_L = 0.077$ and $y_U = 0.113$ respectively.

The best dose estimate is given by the solution of the equation $Y = \bar{y}$, 1.063 Gy. The limits for the 95% confidence region of the dose estimation are the solutions of $Y_+ = y_L$ and $Y_- = y_U$, giving a range of $(0.882, 1.263)$ Gy. Figure 1.5 shows the graphical solution of these equations.

### 1.5.2 Partial body irradiation

Barquinero *et al.* 1997 [10] studied the *in vitro* yield of dicentrics plus centric rings (Dic+CR) after $X$–rays exposure in partial body scenario. This calibration data set (Table 1.2) is based on blood samples from one healthy donor which was irradiated *in vitro* using a cobalt source (Theratron–780) at a dose rate ranged

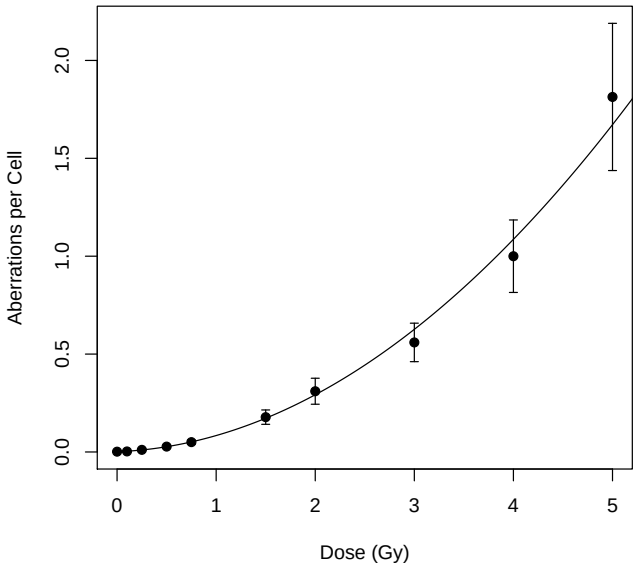


Figure 1.6: Observed means (dots), plus/minus twice their standard errors (error bars), and predicted means (solid line) of the number of dicentrics for Poisson fitting, based on the data in Table 1.1, omitting the 1 Gy test data.

from 117.5 cGy/min to 107 cGy/min simulating acute whole body exposure. The Poisson assumption is considered and supported for most of the $u$ values, only for 5 Gy the $u$–test value is lower than 1.96 (underdispersion).

The fitted calibration curve (omitting the 1 Gy sample) for a linear–quadratic model, $Y = C + \alpha D + \beta D^2$, results

$$\hat{C} = 9.054 \cdot 10^{-4}, \quad \hat{\alpha} = 3.431 \cdot 10^{-2}, \quad \hat{\beta} = 5.702 \cdot 10^{-2};$$

$$\Sigma = \begin{pmatrix} 9.721 & -38.347 & 12.916 \\ -38.347 & 2351.696 & -1012.336 \\ 12.916 & -1012.336 & 885.857 \end{pmatrix} \cdot 10^{-8};$$

and $R = 2.795$. This is all the necessary information to build the curves $Y$ and $Y_{\pm}$. Figure 1.7 shows the plot of the fitted dose–response curve and the observed means of the calibration data.

The test data in this example is going to be the sample comprised 3 Gy for 25%, 0.75 Gy equivalent whole body dose. This sample presents 493 cells free of dicentrics plus centric rings, 23 with 1, 3 with 2 and 1 with 3, this is a total of 36 dicentrics in 521 cells analysed, $n_0 = 493$, $X = 36$ and $n = 521$ following notation in Section 1.4.1. The sample mean and dispersion index and the $u$–test values are $\bar{y} = 0.069$, $s^2/\bar{y} = 1.601$ and 9.822 respectively (overdispersion). Solving Equation 1.1, $\lambda = 0.526$ and $\mathrm{V}(\lambda) = 0.032$.

The best dose estimate is given by the solution of the equation $Y = \lambda$, 2.747 Gy. The limits for the 95% confidence region of the dose estimation are the

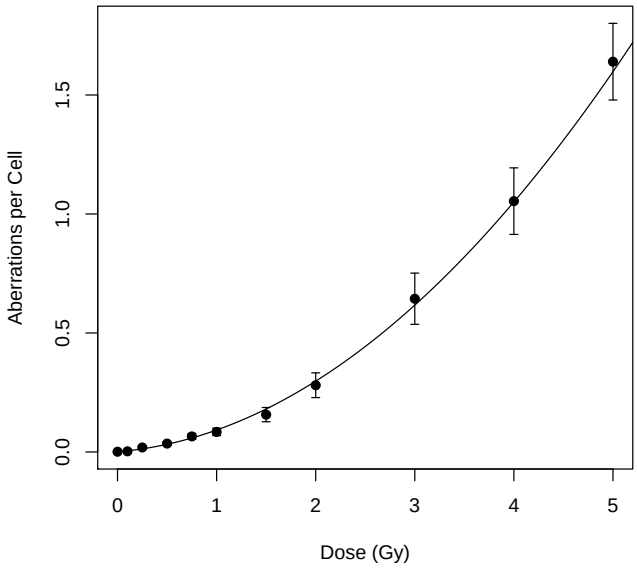


Figure 1.7: Observed means (dots), plus/minus twice their standard errors (error bars), and predicted means (solid line) of the number of dicentrics for Poisson fitting, based on the data in Table 1.2.

solutions of $Y_+ = \hat{\lambda} - 1.96\sqrt{\mathrm{V}(\hat{\lambda})}$ and $Y_- = \hat{\lambda} + 1.96\sqrt{\mathrm{V}(\hat{\lambda})}$, giving a range of $(1.403, 3.832)$ Gy. Figure 1.8 shows the graphical solution of these equations.

The proportion of non-irradiated cells $\omega$ is calculated through Equation 1.2, $\omega = 0.868$. Assuming $d_0 = 2.7$ the estimation fraction of the body irradiated (Equation 1.3) is $F = 0.295$.

## 1.6 Simulations

With the aim to check how the classical method to estimate cytogenetic dose (Section 1.4) measures the uncertainties of the dose estimates, different simulation practices are carried out.

### 1.6.1 Parametric bootstrap

Taking the calibration data of the example in Section 1.5.1, (Table 1.1), including the 1 Gy sample this time, the linear–quadratic curve $Y = C + \alpha D + \beta D^2$ is fitted, resulting

$$\hat{C} = 1.312 \cdot 10^{-3}, \quad \hat{\alpha} = 1.971 \cdot 10^{-2}, \quad \hat{\beta} = 6.288 \cdot 10^{-2};$$

$$\Sigma = \begin{pmatrix} 2.220 & -9.942 & 4.377 \\ -9.942 & 266.018 & -151.076 \\ 4.377 & -151.076 & 160.674 \end{pmatrix} \cdot 10^{-7};$$

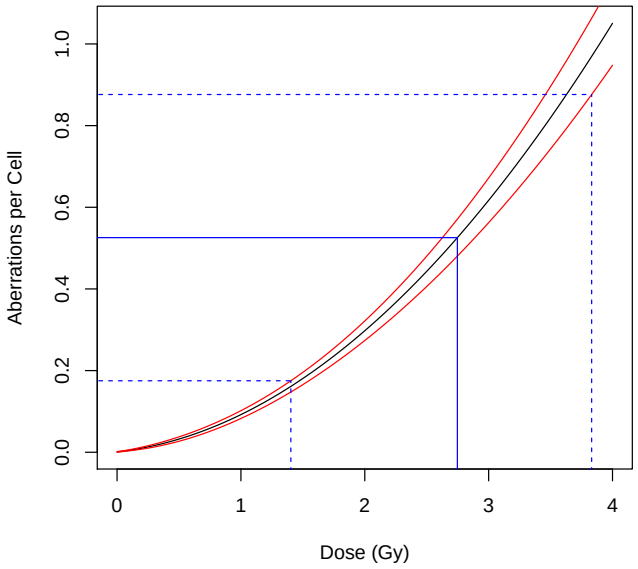


Figure 1.8: Fitted calibration curve (solid black line), its 95% confidence region curves (solid red lines) and the graphical solutions of the three equations (blue lines).

and $R = 2.795$. This is all the necessary information to build the curves $Y$ and $Y_{\pm}$. Figure 1.9 shows the plot of these curves and the observed means of the calibration data.

Samples are simulated for two fictitious scenarios (they do not exist in the Barquinero *et al.* 1995 [9] experiment), one simulating an irradiation of 0.9 Gy in 1500 blood cells and other of 2.5 Gy in 300 cells.

Two different simulations are performed to generate these data:

- Generated by the Poisson uncertainty: for the 0.9 Gy / 1500 cells scenario each simulation consists in 1500 random values simulated as a Poisson with expectation $\hat{C} + \hat{\alpha} \cdot 0.9 + \hat{\beta} \cdot 0.9^2 \approx 0.071$, and for the 2.5 Gy / 300 cells scenario each simulation consists in 300 random values simulated as a Poisson with expectation $\hat{C} + \hat{\alpha} \cdot 2.5 + \hat{\beta} \cdot 2.5^2 \approx 0.448$.

  The following algorithm produces this simulation:

  S1 Generate $y^*$ from Pois$(\hat{C} + \hat{\alpha} \cdot d + \hat{\beta} \cdot d^2)$, where $d$ is the irradiated dose;

  S2 Calculate $y_L$ and $y_U$ from $y^*$, and solve the equations $Y_+ = y_L$ and $Y_- = y_U$ to get the confidence dose range $(d_L, d_U)$, the confidence regions for $y_L/U$ and $Y_{\pm}$ change depending the desired one (e.g. 95%, 90%, 75%, 50%, 25% or 10%);

  S3 If $d \in (d_L, d_U)$ increase the counter by 1. Return to S1 until the desired number of repeats is done. The confidence region is the final

Table 1.3: Real confidence regions for some typical expected confidence regions performing simulations for the Poisson uncertainty.

| Dose / #cells | 95% | 90% | 75% | 50% | 25% | 10% |
|---|---|---|---|---|---|---|
| 0.9 Gy / 1500 | 99.96% | 99.54% | 96.66% | 87.70% | 61.19% | 39.99% |
| 2.5 Gy / 300 | 99.93% | 99.42% | 97.09% | 83.85% | 58.51% | 40.14% |

Table 1.4: Real confidence regions for some typical expected confidence regions performing simulations for the Poisson and the dose–response curve uncertainties.

| Dose / #cells | 95% | 90% | 75% | 50% | 25% | 10% |
|---|---|---|---|---|---|---|
| 0.9 Gy / 1500 | 99.79% | 98.80% | 95.22% | 79.80% | 54.84% | 37.20% |
| 2.5 Gy / 300 | 99.69% | 99.01% | 94.76% | 81.24% | 57.54% | 35.10% |

number of the counter divided by the total number of repeats.

- Generated by the Poisson and the dose–response uncertainty: for the 0.9 Gy / 1500 cells scenario each simulation consists in 1500 random values simulated as a Poisson with expectation $C + \hat{\alpha} \cdot 0.9 + \beta \cdot 0.9^2$, and for the 2.5 Gy / 300 cells scenario each simulation consists in 300 random values simulated as a Poisson with expectation $C + \alpha \cdot 2.5 + \beta \cdot 2.5^2$, where $(C, \alpha, \beta)$ are simulated as trivariate normal with expectation $(\hat{C}, \hat{\alpha}, \hat{\beta})$ and covariance matrix $\Sigma$. These simulations includes the uncertainty of the dose–response curve.

  The following algorithm produces this simulation:

  S1 Generate $(C^*, \alpha^*, \beta^*)$ from $\mathrm{N}\left((C, \alpha, \beta), \Sigma\right)$

  S2 Generate $y^*$ from $\mathrm{Pois}(C^* + \alpha^* \cdot d + \beta^* \cdot d^2)$;

  S3 Solve the equations $Y_+ = y_L$ and $Y_- = y_U$ to get the confidence dose range $(d_L, d_U)$;

  S4 If $d \in (d_L, d_U)$ increase the counter by 1. Return to S1 until the desired number of repeats is done. The confidence region is the final number of the counter divided by the total number of repeats.

To measure the real confidence region for a expected confidence region given of this methodology, for each experiment 10000 samples are simulated checking how many of them return a confidence region which cover the real dose.

Tables 1.3 and 1.4 show the confidence regions obtained applying the classical methodology performing the simulations for the Poisson uncertainty and the simulations for the Poisson and the dose–response curve uncertainties, respectively. These results clearly show that this methodology does not provide an accurate measure of the uncertainty of the dose estimations.

### 1.6.2 Non–parametric bootstrap

Taking the calibration data of the example in Section 1.5.1, (Table 1.1), omitting the 1 Gy (1168 blood cells analysed) sample which is going to be the test data, and assuming a linear–quadratic dose–response curve, $Y = C + \alpha D + \beta D^2$, 10000

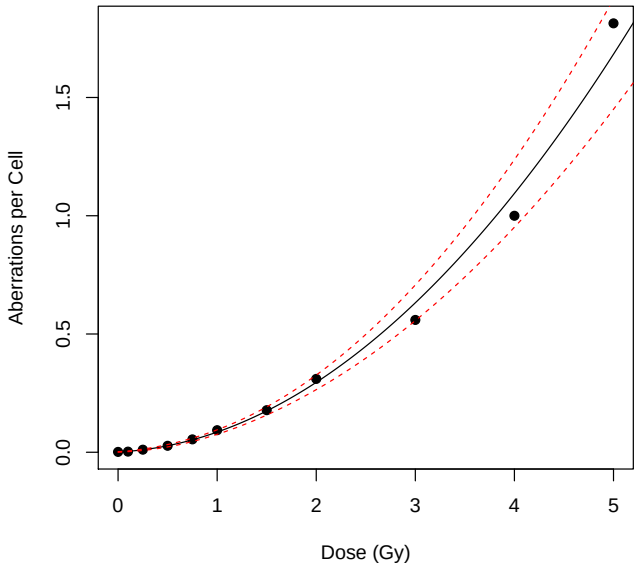


Figure 1.9: Observed means (dots), and predicted means (solid line) of the number of dicentrics for Poisson fitting, and its 95% confidence region (delimited by the red/dashed lines), based on the data in Table 1.1 including the 1 Gy test data.

simulations are carried out by using the frequency distribution of each sample, e.g. each sample (the calibration and the test data) is simulated by samples with replacement from the original one with the same size. For each step the dose–response is fitted for the "new" calibration data and the 95% confidence region of the absorbed dose is estimated for the "new" test data. It is checked how many of them return a confidence region which cover the real dose, 1 Gy, to measure the real confidence region.

The following algorithm produces this simulation:

S1 Generate $\theta^* = (\theta_1^*, \theta_2^*, \ldots, \theta_m^*)$ from each sample $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$, where $\theta_i^*$ is random selected from $\{\theta_1, \theta_2, \ldots, \theta_m\}$ with replacement, $i = 1, 2 \ldots, m$.

S2 Calculate the calibration coefficients $\hat{C}$, $\hat{\alpha}$ and $\hat{\beta}$, and the covariance matrix $\Sigma$ from the simulated calibration data;

S3 Calculate the confidence dose interval from the simulated test data and the fitted dose–response curve;

S4 If the confidence dose interval covers the real dose, increase the counter by 1. Return to S1 until the desired number of repeats is done. The confidence region is the final number of the counter divided by the total number of repeats.

Table 1.5: Real confidence regions for some typical expected confidence regions performing a non–parametric bootstrap.

| Dose / #cells | 95% | 90% | 75% | 50% | 25% | 10% |
|---|---|---|---|---|---|---|
| 1 Gy / 1168 | 97.13% | 93.39% | 81.45% | 60.01% | 37.93% | 21.92% |
| 2 Gy / 332 | 98.45% | 96.64% | 87.82% | 69.18% | 46.13% | 28.94% |

Table 1.5 shows the confidence regions obtained applying the classical methodology performing a non–parametric bootstrap. The same simulation is analogously performed taking the 2 Gy (332 blood cells analysed) sample as test data.

These results also show that the classical methodology does not provide an accurate measure of the uncertainty of the dose estimations. This is one of the reasons to explore the Bayesian approach for cytogenetic dose estimation in the research of this Thesis.

# Chapter 2

# Presentation and discussion of the results

This thesis collects six scientific works, five of them published/accepted by the date of the submission of this document and the other one in the revision stage.

## 2.1 Review of Bayesian methods in biodosimetry

This work is reproduced in Chapter 4 and corresponds to reference Ainsbury *et al.* 2014 [4].

This is a review of the Bayesian models for radiation cytogenetics proposed in the literature. Indeed, there is a practical overview of Bayesian cytogenetic dose estimation including some application examples.

Throughout the literature the Bayesian analysis have been applied showing its usefulness for cytogenetic data analysis. The Bayesian framework provide results in form of probability densities which could give more accurate conclusions, like the probability for an absorbed dose to be in a concrete range. Indeed the Bayesian techniques gives a fuller and more rigorous consideration of the associated uncertainties for dose estimation. The Bayesian approach has a number of obvious advantages for cytogenetic radiation dose estimation, where high quality prior information is generally available and where the estimated dose is more correctly represented by a distribution of possible values.

## 2.2 New calibration model applied for biodosimetry

This work is reproduced in Chapter 5 and corresponds to reference Higueras *et al.* 2015a [39].

In this project new Bayesian–type count data inverse regression methods are introduced for situations where responses are Poisson or two–parameter compound Poisson distributed with application in cytogenetic radiation biodosimetry and in radiotherapy. These models can be calculated in a closed form, in

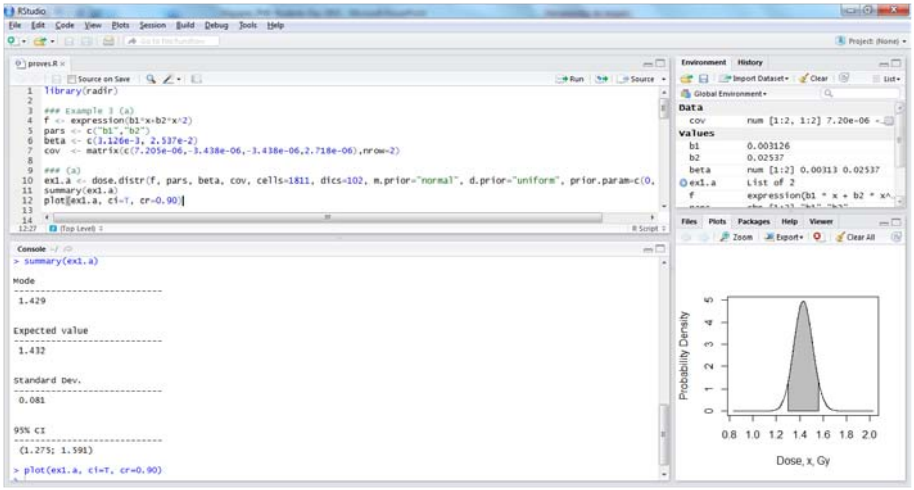


Figure 2.1: Screenshot of `radir` in the `RStudio` interface.

the case of the compound Poisson only the simplified ones and they allow any dose–response curve with very mild conditions.

The applied examples given in this work demonstrate that the approach described is accurate and informative for practical cytogenetic dosimetry.

## 2.3 Package `radir`

This work is reproduced in Chapter 6 and corresponds to reference Moriña *et al.* 2015a [65].

The Bayesian framework has been shown to be very useful in cytogenetic dose estimation. This paper describes the new `R` package `radir`, which implements a the Bayesian–type dose estimation methodology in Higueras *et al.* 2015a [39] for the Poisson assumption of the chromosomal aberrations yield and the required dose-response curve (typically linear or linear-quadratic). The package is able to calculate and plot the calibrative dose density for a given set of inputs and provides its most relevant summary statistics including the best dose estimate, the expected value, the standard deviation and credibility interval. Figure 2.1 shows a screenshot of an application of `radir`.

Several examples of application are provided. The package is useful for a quick and easy examination of patients after unplanned exposures, like accidental overexposures in radiology services at hospitals, occupational exposures or to follow up people affected by major nuclear accidents, such as Chernobyl or Fukushima.

## 2.4 New model for partial body irradiation

This work is reproduced in Chapter 7 and corresponds to reference Higueras *et al.* 2015b [40].

A new Bayesian zero-inflated Poisson model is derived for dose and fraction of the body irradiated estimation in partial body irradiation scenarios inside the Bayesian framework. In the current study, Bayes factors are applied to identify whether a sample of chromosomal aberrations in blood cells has been partially or wholly body irradiated, contrasting the zero–inflated Poisson against the Poisson assumption. Estimates are given in form of probability densities and bivariate histograms, providing accurate and informative inference results. The methods are tested and validated using data from a range of simulated exposure scenarios, irradiated fractions and doses up to 20 Gy to the irradiated fraction.

The examples show that this methodology is highly promising for practical biological dosimetry; indeed, the results are more accurate and more appropriate to analysis of cytogenetic data than the classical methods currently in use. Further, the data required to implement this analysis (dose response curve parameters and covariance matrices) are readily available.

## 2.5 Zero–inflated regression models for radiation–induced chromosome aberration data

This work is reproduced in Chapter 8 and corresponds to reference Oliveira *et al.* 2015 [70].

For radiation induced chromosome aberrations data, the Poisson distribution is the most widely recognised and commonly used distribution and constitutes the standard framework for explaining the relationship between the outcome variable and the dose. However, in practice, the assumption of equidispersion implicit in the Poisson model is often violated due to unobserved heterogeneity in the cell population, which will render the variance of observed aberration counts larger than their mean, and/or the frequency of zero counts greater than expected for the Poisson distribution. The goal of this work is to study the performance of zero-inflated models for modelling such data. For that purpose, a substantial analysis is performed, where zero-inflated models are compared with other models already used in biodosimetry, such as Poisson, negative binomial, Neyman type A, Hermite, Polya–Aeppli and Poisson-inverse Gaussian. Several real data sets obtained under different scenarios, whole and partial body exposure, and different types of radiation are considered in the study.

## 2.6 Package `hermite`

This work is reproduced in Chapter A and corresponds to reference Moriña *et al.* 2015b [66].

Generalized Hermite distributions are a family of two–parameter count distributions. These distributions can be useful for modelling count data that presents multi–modality or overdispersion (the variance greater than the mean), situations that appear commonly in practice in many fields. These distributions are closed under convolution and their maximum likelihood estimator of the population mean is the sample mean. A Generalized Hermite distribution of order $m$ is represented by $X_1 + mX_2$, where $X_1$ and $X_2$ are Poisson distributed independent random variables, and $m$ is an integer greater or equal to 2. The second order Generalized Hermite distribution is the classical Hermite

distribution. In this work a new R package is presented which allows the user to work with the probability density, cumulative density, quantile and random generation functions of the Generalized Hermite distributions. When one (or both) of the population means of and is (are) greater than 20, the distribution function is approximated using an Edgeworth expansion, the probability mass function is calculated from this approximation of the distribution function, and the quantile function is approached by a Cornish–Fisher expansion. The hermite package also allows the user to perform the likelihood ratio test for Poisson assumption and to estimate parameters using the maximum likelihood method.

Practical examples of the usage of these distributions can be found in biology and economy fields, like cytogenetic biological dosimetry.

# Chapter 3

# Summary, conclusion and further work

Cytogenetic dose estimation statistical models have been developed and applied to the entire process of cytogenetic biological dosimetry to get accurate inferences and quantifications of their uncertainties. This has resulted in production of statistical models for dose estimation which have a relevant role in the different situations where overexposure irradiation is suspected, IAEA 2011 [42].

This thesis has been focused on investigating statistical models with application to cytogenetic biodosimetry data analysis in several different exposure scenarios. This research could lead to establishment of an alternative accepted process for the estimation of absorbed dose, including consideration of more count data distributions like the two–parameters compound Poisson distributions and their respective zero–inflated models, and allowing a wide range of dose–response curves (not only the typical linear and linear–quadratic). The Bayesian Information Criterion and the Bayes factor can be applied to determine the radiation scenario, instead the frequentist $u$–test.

First, in Chapter 4, Ainsbury *et al.* 2014 [4], Bayesian models for cytogenetic data analysis over the field literature are reviewed, with special focus in dose–estimation methodology.

After this, in Chapter 5, Higueras *et al.* 2015a [39] new Bayesian–like inverse regression models for Poisson and two parameters compound Poisson responses are developed for dose estimation in whole body irradiation. These models are very flexible, allowing any kind of dose–response curves with very mild conditions. Some of these models derive the usage of compound Hermite distributions, for which an expression of the probability mass function is provided.

Chapter 6, Moriña *et al.* 2015a [65], describes the new software implementation within the `R` framework for the Poisson models presented in Chapter 5, the `radir` package. This package have dependency on the `R` package entitled `hermite`, Appendix A, Moriña *et al.* 2015b [66], which provides utilities for the Generalised Hermite distributions.

In Chapter 7, Higueras *et al.* 2015b [40], a new Bayesian model for partial body irradiation is introduced, leading to zero–inflated Poisson models.

Finally, Chapter 8, Oliveira *et al.* 2015 [70], presents a comparative study

for the case of cytogenetic data to the study of different count models to fit dose–response curves. Here, zero-inflated models to describe the number of chromosome aberrations in biological dosimetry are compared with the Poisson, and some two–parameter compound Poisson models (negative binomial, Neyman A, Hermite, Pólya-Aeppli and Poisson-inverse Gaussian) under different irradiation scenarios.

## 3.1 Further work

### 3.1.1 Zero–inflated compound Poisson models

Following the models in Chapter 7, Higueras *et al.* 2015b [40], and the compound Poisson models in Chapter 5, Higueras *et al.* 2015a [39], zero–inflated negative binomial, zero–inflated Neyman A and zero–inflated Hermite models could be derived in the Bayesian framework for dose and fraction of the body irradiated estimation in partial body exposure scenarios which the chromosomal aberration yield in the irradiated fraction is overdispersed, e.g. high–LET sources like $\alpha$–particles. Indeed, this methodology could be applied to all zero–inflated responses to whom the underlying distribution is a two–parameter compound Poisson.

### 3.1.2 Dose estimation in gradient exposure scenarios

Gradient exposure is an irradiation scenario in which a subject is exposed to different fractions of the body. This is a situation which provides great difficulties for biological dosimetry and clinicians, and as yet no viable biodosimetry solutions have been proposed. The dose estimations in these situations could lead to finite mixture models, see e.g. Frühwirth–Schnatter 2006 [24].

### 3.1.3 Software

The `radir R` package software introduced in Chapter 6, Moriña *et al.* 2015a [65], is dedicated only to the Poisson responses, so it mainly covers the low–LET whole body irradiation scenarios. Enhancements in this software could lead to include the negative binomial, Neyman A and Hermite models in Chapter 5, Higueras *et al.* 2015a [39], and the zero–inflated Poisson models in Chapter 7, Higueras *et al.* 2015b [40].

The current calculations of the posterior density in these zero–inflated Poisson models applying the acceptance–rejection technique is computative intensive and sometimes require a couple of hours. Laplace approximations could be explored before the implementation of these models in the `radir` system, to get quicker calculations.

In general, most of the advances forwarding this research lines for cytogenetic dose estimation could be included in `radir`.

### 3.1.4 Other assays

A chromosomal translocation is a chromosome aberration caused by rearrangement of parts between different chromosomes. These count data is managed analogously to dicentrics, dicentrics plus centric rings and micronuclei data and

taking into account individual factors like the age and the gender of the exposed subjects. In a similar way the H2AX assay also leads to count models taking into account some inter–individual factors

Indeed, the methods could be applicable for markers of late radiation effects, the biomarkers of effect, an area of increasing interest within the low dose radiation research community. Markers to be investigated would include those for radiation induced cancers, for instance, and a number of interesting challenges with data interpretation are foreseen.

The gene expression assay is not represented by count data; thus the models developed in this thesis are not applicable.

# Chapter 4

# Review of Bayesian statistical analysis methods for cytogenetic radiation biodosimetry, with a practical example

This chapter corresponds to the contents of [4].

**Abstract:** Classical methods of assessing the uncertainty associated with radiation doses estimated using cytogenetic techniques are now extremely well defined. However, several authors have suggested that a Bayesian approach to uncertainty estimation may be more suitable for cytogenetic data, which are inherently stochastic in nature. The Bayesian analysis framework focuses on identification of probability distributions (for yield of aberrations or estimated dose), which also means that uncertainty is an intrinsic part of the analysis, rather than an 'afterthought'. In this chapter Bayesian, as well as some more advanced classical, data analysis methods for radiation cytogenetics are reviewed that have been proposed in the literature. A practical overview of Bayesian cytogenetic dose estimation is also presented, with worked examples from the literature.

**Keywords:** chromosome–aberrations, biological dosimetry, critically accident, Poisson distribution, confidence limits, model, distributions, credibility, uncertainty, software.

## 4.1 Introduction

The classical methods for cytogenetic radiation dose estimation are now extremely well defined. In brief, dose–response calibration data are collected and fitted to a linear or linear quadratic model, the coefficients of which are then used to calculate the dose from the chromosome aberration yield scored in peripheral blood lymphocytes from a potentially exposed individual. The Poisson

model is generally used to estimate the uncertainty on the yield of aberrations, and this is combined with uncertainty on the fitted calibration coefficient(s) using standard methodology in order to give an overall estimate of uncertainty associated with the evaluated dose. Full details of this procedure can be found in the International Atomic Energy Agency manual [42] and the International Organization for Standardization (ISO) standard 19238 [43].

In general, within the field of cytogenetic dose estimation, chromosome aberration yields are considered as something 'fixed' according to the frequentist or classical assumptions. This allows a deterministic estimate of the radiation dose and associated confidence limits to be made. Note, however, that the classical statement that one is 95% confident that the unknown true value of the parameter is contained in the interval defined by the confidence limits really means that this interval has been defined using a procedure that is successful in giving correct results 95% of the time. This interpretation is quite different from the Bayesian confidence interval (or credible interval), understood as an interval in the domain of a posterior probability. Following a measurement, the data are known and fixed, and the parameter values are unknown. So, the natural probability to be considered is the probability distribution of the parameter values, given the data. Using the elementary rules of conditional probability, this brings in the prior probability distribution of true values, and so the proper probabilistic interpretation of a measurement will depend on the prior information. However, in classical cytogenetic dose estimation, assignment of a probability to an event is usually based solely on the observed frequency of occurrence of the event without any prior information, and the confidence limits are calculated based on the assumed distribution of data around the observed mean frequency. It is also important to note that, using appropriate priors, examples can be constructed where the posterior probability distribution is entirely outside the classical limits.

For instance, if one dicentric has been scored in 1000 cells, the classical probability of observing a dicentric in the 1001st cell is 1/1000. This 'frequentist' approach takes into account only the chromosomal damage which has been identified in a particular sample, and as such does not fully consider the intrinsically stochastic nature of aberrations or any previous knowledge of the process. In reality, the observed number of aberrations (e.g. in a calibration data set or in a single measurement) only represents a snapshot of the true situation which can only be realistically expressed as a probability distribution. The overall outcome of a chromosomal analysis is a combination of the probability of induction of aberrations by radiation and the chance of their detection. In other words, the classical 'probability' is based on unknown quantities and is therefore not useful in defining the true probability.

The alternative to the classical methodology is a Bayesian approach to data analysis. In the Bayesian framework, probability of an event is described in terms of previous beliefs and uncertainty. Previously existing, or prior, information is used in combination with experimental results to infer the probability that a hypothesis is true, in accordance with the Bayes theory

$$P(D_p|D_n) = \frac{P(D_n|D_p)P(D_p)}{P(D_n)}$$

where, $D_p$ is previously existing or prior data; $D_n$ refers to newly collected experimental results; $P(D_p)$ is the probability of occurrence of $D_p$; $P(D_n)$ is

the probability of occurrence of $D_n$; $P(D_p|D_n)$ is the conditional probability of $D_p$ given the occurrence of $D_n$, and $P(D_n|D_p)$ is the conditional probability of occurrence of $D_n$ given the occurrence of $D_p$. The details of Bayes theory and Bayesian analysis are presented elsewhere (e.g. [88], [16]) and will not be reproduced here. Worked examples of the use of Bayes theorem in a radiation cytogenetics setting are, however, given in the Section 4.4.

A number of authors have suggested that a Bayesian approach to analysis of cytogenetic data may be preferable to the classical, frequentist techniques and the potential of the former has been demonstrated in several sets of specific circumstances. Bayesian methods are eminently suitable for analysis of chromosome aberrations, which are inherently stochastic in nature. Bayesian statistics considers that aberrations can be induced in a probabilistic manner, but will not necessarily be detected. This is a much more realistic reflection of the true scenario: the classical method of representing the mean dose and associated confidence limits allows consideration only of a single "point" estimate of suspected dose, ignoring the fact that the suspected or likely dose is much more accurately represented as a distribution of possible values. In contrast, the Bayesian method gives the probability that the dose was in a certain dose range or above or below a certain value. Additionally, prior knowledge is often available in biodosimetry in the form of background levels of aberrations, calibration curves, or expected adherence to the Poisson distribution.

A further important advantage of Bayesian over classical methodology is that in Bayesian statistics, the use of distributions for all parameters means that the uncertainties associated with the parameters are intrinsically included in the analysis. Uncertainties are therefore an integral part of the dose calculations, rather than being considered separately as an 'afterthought' which is typically how uncertainty is incorporated using the classical approach. As a result of this, it has been shown that the Bayesian approach increases quality assurance as, in order to create and use a statistical model, it is necessary to completely understand the measurement that is being performed and this necessarily minimises false assumptions and so leads to optimisation [87].

In practice, the Bayesian framework is based on the continuous modification of the understanding of a model by newly collected data. At any one time, the understanding of the model is characterised by what is termed the prior information. Each parameter in a given model has a prior distribution associated with it, and in this way the uncertainty surrounding each parameter is an inherent part of the analysis. When new data are collected, the data are combined with the prior information or model(s) in order to form a posterior model, which then represents the complete, up to date, knowledge regarding the system that is under investigation. Bayesian analysis methods can be used to calculate, for instance, the probability that a radiation exposure occurred or that a specific radiation dose was received.

In this chapter we review Bayesian, as well as some more advanced classical, data analysis methods for radiation cytogenetics that have been proposed in the literature. We also present in the appendix a practical overview of Bayesian cytogenetic radiation dose estimation, with a worked example from the literature, which more fully demonstrates the usefulness of Bayesian techniques in this setting.

## 4.2 Bayesian methods for cytogenetic biodosimetry

### 4.2.1 Model selection

There are several forms and classes of distributions that can be used to model the probability of occurrence of events. The type of distribution chosen is very important for accurate data analysis, and several models have been proposed and implemented for the assessment of cytogenetic data. The most commonly used models and their applicability to radiation cytogenetics are discussed below.

In radiation cytogenetics, the Poisson distribution has long been the model of choice. In general, counts of chromosome aberrations are low and Edwards and colleagues showed that chromosome aberrations can therefore be modelled extremely effectively by the Poisson model [22]. In the wider literature, the Poisson distribution has been used in the Bayesian framework by many authors. Despite this, Bayesian Poisson modelling in radiation cytogenetics is surprisingly limited, perhaps as, to date, most forays into the field of cytogenetic data analysis have focused on analysing data that violate the assumptions associated with the Poisson model, in particular that the variance and mean should be statistically equivalent. The classical method of dealing with observed overdispersion is to increase the standard error associated with the measurement; moreover, significant overdispersion would ideally be dealt with by adding an additional parameter to the Poisson model, for example through implementation of the negative binomial, Neyman A or Hermite models. However, other than the negative binomial, only a limited number of attempts have been made to develop a Bayesian methodology in radiation cytogenetics using other overdispersed distributions like the Neyman A or the Hermite. Intuitively, the Bayesian approach is certainly more correct than the classical approach, as it leads to a much more rigorous interpretation of the uncertainty associated with the resulting dose estimate.

In 2003, Sasaki [83] presented a method of analysis for chromosome aberration data, in an attempt to deal with the problems of inappropriate estimation of average dose which result from inhomogeneity of exposure. In such a scenario, the cell population consists of a mix of sub populations, each exposed to a different dose, causing a different amount of damage. The distribution of chromosome damage in cells can therefore be expressed in terms of a mixed Poisson distribution, and 'unfolding' of this creates a dose distribution profile. Although formally classical in nature, this 'Bayesian-like' approach produces a final distribution containing information on dose inhomogeneity as well as the 'prior' information of variability induced by having a spectrum of charged particles and multiple ionisation events in the cell nuclei. The model was demonstrated to provide adequate fits for the linear-quadratic dose response for simulated accidents and real overexposure data.

Morand and others [64] published a technical note describing the NETA computer program, which can be used to calculate the 95% confidence limits of Neyman A distributed events. This distribution was first proposed by Neyman in 1939 [69], who introduced it as a way to test the difference between means of two samples of count data with different variances. This is in contrast to

other standard tests such as the z– and t–tests, for example, which are based on normally distributed data with known and unknown population standard deviations respectively. The Neyman A distribution tends towards the Poisson distribution when its theoretical dispersion index tends to 1. Morand and colleagues found that the confidence limits calculated using the Neyman distribution were smaller than those calculated using the traditional Poisson-based method for small sample sizes (numbers of cells) [64].

Stiratelli et al. [90] consider the beta–binomial model. The beta distribution embodies a family of continuous probability distributions, which are defined on the interval $[0, 1]$ by two shape parameters, usually referred to as alpha and beta. Although beta distributions are used extensively in Bayesian statistics as they provide conjugate prior distributions (defined as being in the same family as the posterior distribution) for binomial (including negative binomial) and geometric distributions, the approach of Stiratelli et al. [90] is classical in nature. The Dirichlet distributions are an extension of the beta distribution for multiple ($>2$) parameters. The beta–binomial distribution arises when one considers the $p$ parameter in a binomial distribution as being randomly drawn from a beta distribution. Stiratelli and colleagues [90] compared the Poisson and binomial distributions for chemically induced chromosomal damage with the beta–binomial, negative binomial and correlated–binomial distributions. In contrast to the Poisson and simple binomial distributions, these models do not rely on independence of cellular response. The authors found that all the beta–binomial distribution based models showed improved fits with respect to the Poisson and binomial models (as tested by the $\chi^2$ test). The beta–binomial model provided the best fit with respect to the authors' data set.

It is well documented that chromosome aberration data produced as a result of high energy and high LET radiation can be over–dispersed, that is the variance of the data is greater than the mean, violating the Poisson assumption that the variance and mean are equal. Brame and Groer considered over–dispersion from a Bayesian standpoint [14]. The usual assumptions are of an underlying statistical relation between dose, yield and number of cells scored. This can be compared with a previously formed calibration curve for Poisson distributed yield which is linear or linear quadratic in nature. A Bayesian approach using a negative binomial model is applicable when over–dispersion is suspected: the negative binomial is characterised by a parameter, $\Psi$, which measures the degree of overdispersion. With this model, overdispersion becomes independent of dose as the expected number of dicentrics increases and as $\Psi$ tends to 0, the negative binomial tends towards the Poisson. The authors also used gamma priors, amongst others, and demonstrated the use of Bayes factors for model comparisons.

One of the criticisms of the Bayesian framework is that a successful outcome relies heavily on the initial choice of model(s). Kottas and others [49] and Krnjajic and colleagues [50] recently presented details of Bayesian non–parametric models which can be used for data which would traditionally be analysed using a Poisson-based parametric model. The authors tested the Bayesian approach with in vitro and real overdose data for radiation induced micronuclei, and found that, in many cases, the non–parametric model produced more accurate predictions than the parametric, Poisson, models. Mukhopadhyay [67] discusses the use of the Dirichlet prior for non–parametric Bayesian inference of dose levels. A numerical example of the described method is given and its particular use in

calculating percentiles is demonstrated.

Pereira and Stern [73] presented details of a full Bayesian significance test for hypothesis testing based on credible sets. The method presented by the authors samples from the "parameter space" rather than from the 'sample space' which may be much more appropriate for many different types of data, including, in the opinion of the authors of this manuscript, cytogenetic data. In 2001, the application of the full Bayesian significance test for model selection was discussed by the same authors, which relies on testing of significance of individual parameters of models. An example was given using the multiple linear regression model [74].

Probability density functions (PDFs) are a very important concept in Bayesian uncertainty analysis. PDFs represent the distribution of probabilities of a quantity, and are used to formally define prior knowledge. In 2006, van Dijk presented details of numerical methods for calculating uncertainty in personal dosimetry. Monte Carlo sampling was used to construct a probability density function of dose from personal dosimeter measurements. This allows the uncertainty and confidence intervals to be calculated: PDFs were assigned to all input quantities and then these were combined to produce a single output PDF. Although classical rather than Bayesian in nature, this method demonstrates the importance of correctly forming the PDF in order to accurately assess uncertainties [93].

### 4.2.2 Detection limit and decision threshold

Miller and colleagues [62] presented details of a Bayesian method of determining detection limits, i.e. whether results of biological measurements should be deemed positive or negative. The authors discuss the advantages of the Bayesian approach, which allows knowledge of the results of previous measurements to be incorporated into the decision making process. In 1995, Miller and colleagues followed this work with quantitative assessment of the methodology, showing that Bayesian methods are much more suitable for detection limit analysis than the corresponding classical methods, as they allow the prior knowledge of the population to be included in the calculations [63].

In 2002, Groer demonstrated the applicability of Bayesian techniques in statistical analysis of Poisson distributed radiation net counting rates. The author highlighted the fact that the Bayesian methods involving probability densities allow uncertainties to be fully characterised and also represented pictorially [31]. The following year, Groer and Carnes describe the application of Bayesian statistical methods for threshold estimation for radiation–induced lung cancer in mice. A Weibull based proportional hazards model was used, which allows the dose threshold to be calculated. The authors used a Bayesian approach to estimate the parameters of the model and characterise the uncertainty of the estimates with probability distributions, allowing calculation of probability based confidence intervals. Uniform and improper priors were used to account for a lack of initial knowledge of the distributions. The results show differences between the two types of radiation investigated, Co-60 gamma rays and neutrons: the gamma estimates of threshold were approximately normally distributed with a median on the order of 0.5 Gy; the neutron distribution was exponential with thresholds only becoming likely at 0.2 Gy or less [32]. In 2006, Weise and colleagues described the concepts of decision threshold and detection limit using Bayesian methodology. The detection limit is defined as the value of

the upper 95% confidence limit of the distribution of possible doses given a true dose of zero; it is therefore the smallest true value that is statistically detectable with the method used. The decision threshold is defined as the limit for which the lower 95% confidence threshold is equivalent to the detection limit; values above this level are statistically indicative of at least 95% probability of a true event. Bayesian methods for calculating both were presented, together with confidence limits. Again, the ISO standard (ISO 2004) was considered, and the authors concluded that Bayesian methods are most appropriate for evaluation of uncertainties [99].

### 4.2.3 Dose estimation using the dicentric assay

In 1987, Groer and Pereira [33] investigated the application of Bayesian statistical methods for chromosome dosimetry in neutron exposure. Bayesian calibrative density functions were formed, which are defined by incorporating data from physical dose estimates, calibration experiments for neutrons of the same dose, and 'new' information from the dicentric chromosome aberration assay. The authors demonstrated the use of the density function to produce an analytical estimate of dose, assuming the Poisson form of the calibration data, and applying a gamma prior for dose and for the alpha (linear) term of the calibration curve $y = \alpha D$.

As dicentrics are not stable over time, it is generally accepted that the dicentric assay can only be used for short time periods after radiation exposure. Straume and Bender, for example, discussed the reliability of techniques in cytogenetic biodosimetry including the dicentric assay, with respect to complex and protracted radiation exposures such as those experienced by astronauts in space. The authors point out that simple proportional yields are not sufficient to fully describe the protracted and complex exposure conditions that are experienced in space [89]. Limitations of the dicentric assay include the fact that it is not sufficient for dosimetry on missions of periods greater than a few months, due to the large uncertainties associated with back–extrapolation of dicentric frequencies. However, in 1988, Bender and colleagues investigated the problem of detection of very low doses (<0.1 Gy) for individuals sampled decades after exposure, using a Bayesian statistical approach. The authors concluded that Bayesian methods could be used to provide a formal statement of likelihood that observed chromosome damage is due to radiation exposure at time points far later than is conventionally possible [12].

The negative binomial distribution has also been used as an example to demonstrate a Bayesian approach to parameter and dose estimation for chromosome aberrations caused by neutron exposure. Brame and Groer [14] used a negative binomial PDF for yield to characterise uncertainty in the dose. The negative binomial model was shown to give a better fit for data from high energy (14.7 MeV) neutrons and to be very similar to the Poisson model results for fission neutrons at 0.7 MeV.

Criticality accidents are those in which a combination of neutron and gamma irradiation contributes to the overall dose. Present methods for dose estimation in this situation rely on the availability of physical calculations or estimates of the ratio of neutron:gamma doses. Brame and Groer also described a Bayesian approach to dose estimation in a criticality accident. Posterior probability densities for the total and the neutron and gamma doses were derived, allowing

the uncertainty in dose ratio to be included in the calculations. The Bayesian approach led to an increase in uncertainty associated with neutron, gamma and total dose estimates but did not affect the calculated values of the doses. The Bayesian criticality method was found to give very similar results to the classical iterative approach in a simulated accident situation ([14], [97]).

In 2003, Maznik and colleagues investigated the distribution of initial biological dose estimates for Chernobyl cleanup workers. The authors used a Bayesian approach to analyse cytogenetic data, constructing probability distributions for yields of dicentrics and incorporating a classical conversion to dose. The authors discuss the advantages of Bayesian methodology in cases of low doses and small numbers of scored cells or observed aberrations [57]. The method was also applied for evacuees from the vicinity of Chernobyl [58] and for retrospective cytogenetic dosimetry for cleanup workers [59]. In all cases, the Bayesian approach was shown to be robust and relatively simple to implement, allowing the calculation of mean doses and associated confidence intervals as well as the most probable doses and dose limits for the probability density intervals outlined by the researchers.

There has been a large amount of interest in the potential for automation of the dicentric assay, indeed this is the logical next step for this biodosimetry technique. As early as 1992, Piper and Sprey were investigating the potential of parametric Bayesian methods for automatically classifying centromere candidates [75]. The Bayesian classification system was found to give considerable improvements in terms of the false positive rate, in comparison to a traditional 'box' classifier.

Most recently, DiGiorgio and Zaretzky used a Bayesian approach to present the uncertainty on a biological dose estimate for a radiation overexposure patient in Latin–America [21]. A Poisson model with a Jeffrey's prior was used and it was further demonstrated that the Bayesian approach allows presentation of probabilities for dose ranges, which leads to a much more intuitive interpretation of the biological dosimetry results.

### 4.2.4 Dose estimation using the micronucleus assay

Low LET/low dose irradiation has been shown to lead to a Poisson distribution of micronuclei (MN) [22]. A highly variable spontaneous frequency of MN is observed across individuals and this produces challenges for modelling damage yields and producing calibration curves. In 1994, Madruga and colleagues presented a Bayesian method of analysing micronuclei to give dose–response curves, using the log–odds transformation presented by Aitchison and Shen [5]. The information contained in the experimental data is used to produce the a–priori calibration (again with the Dirichlet model), which can then be used to give the posterior distribution [54]. Madruga and colleagues further developed the above methods for calculation of Bayesian credibility intervals, and measured credibility, associated with estimated dose [55]. The authors demonstrate that credible intervals can be created for unknown doses, solely based on observed frequencies of micronuclei. Matthews also presented details of a Bayesian approach to credibility assessment, though in the context of clinical trial outcomes. As with other applications, the Bayesian methodology allows quantitative use of prior information in credibility calculations: the critical prior interval is a measure of the confidence in the prior information that is required for the outcome of the

trial to be considered plausible [56].

Serna and colleagues used Bayesian techniques to calculate uncertainties for micronuclei found in thyroid cancer patients after treatment with Iodine–131. The analysis was based on the fact that the total number of micronuclei incorporates both the background distribution and the radiation induced Poisson distribution. The method facilitates the inclusion of uncertainties in the confidence limit calculations. The Bayesian approach was demonstrated to be particularly useful in low dose situations as, even where the counts of radiation induced micronuclei were equal to or less than the background counts, the Bayesian dose calculated in this situation was always positive due to the prior information that a radiation dose was received [86].

## 4.3 Discussion and conclusion

Where methods have been proposed and tested, the Bayesian framework has been shown to be useful for analysis of cytogenetic data. The results of a Bayesian analysis are provided in terms of probability distributions, which can then be manipulated to give more logical conclusions, for instance the probability that a radiation dose received by an individual was in a certain range. In general the Bayesian analyses have been shown to give conceptually better results than the classical counterparts, i.e. a fuller and more rigorous consideration of the associated uncertainties. A Bayesian approach also requires a more complete understanding of the system which is under investigation, which is advantageous for both analysis and interpretation of the results. The only negative is perhaps that full Bayesian analysis can sometimes be mathematically and computationally intensive, and given the extremely well defined nature of currently available classical methods, it may be difficult to persuade those who have many years of experience with classical data analysis that a move towards the Bayesian scheme could be advantageous, in order to overcome some of the limitations in the existing standardised statistical methodology [94]. Nevertheless, numerous texts are available to guide the user (e.g. [88]) and numerous statistical software packages are available, for instance several packages made in R [81] and WinBUGS [53] and one recent cytogenetic biodosimetry–driven program [3], which allow the appropriate calculations to be carried out. The advantages of the Bayesian framework are clear, and therefore more work in this area is required, both to publicise the potential for Bayesian analysis in biological radiation dose estimation and to develop the methodology further.

Section 4.4 presents a worked example for the application of Bayesian methods to radiation cytogenetics. In addition, the authors are currently developing a software tool which facilitates application of a number of the Bayesian methods discussed in this chapter.

## 4.4 Appendix: Methodology of Bayesian cytogenetic radiation dose estimation

As discussed in the main body of text, the principal of Bayesian analysis relies on the use of pre–existing or 'prior' information in order to make judgements about 'future' experimental measurements. In terms of cytogenetic biodosimetry, the

'prior' information may be the distribution of background rates of chromosome aberrations, e.g. dicentrics, and/or previously collected calibration data regarding yields of aberrations at different doses that have been used to set up reference curves for different types of radiation [42]. The word 'future' here can be taken to refer to all experimental data collected following the initial calibration experiments.

Although it is possible to find numerous basic texts on the subject of Bayesian statistical methods(e.g. [16, 88]), for completeness, the basic terminology and methodology of Bayesian inference is explained below. This is followed by a detailed example (from the literature) which demonstrates how to use this method for cytogenetic radiation dose estimation.

### 4.4.1 Method of Bayesian inference of posterior predictive distribution of new data and calibration data

Given a prior model (distribution), $f(x|\theta)$, for previously collected data with $x$ one or more parameters $\theta$ and a prior distribution for this parameter, $p(\theta)$, the 'joint' (combined) distribution of the prior data model and the prior parameter model, $h(\theta|x)$, is defined by Equation (4.1):

$$h(\theta|x) = f(x|\theta)p(\theta). \tag{4.1}$$

The 'marginal distribution' of this information, $m(x)$ (also called the 'prior predictive distribution'), is defined as the integral of $h(\theta|x)$, over all values of the parameter(s) $\theta$:

$$m(x) = \int h(\theta|x)\mathrm{d}\theta = \int f(x|\theta)p(\theta)\mathrm{d}\theta. \tag{4.2}$$

The posterior distribution of the model for the calibration data, $p(\theta|x)$, is the distribution which emerges when the prior information regarding the parameter $\theta$ and the model of the calibration data are combined as in equation Equation (4.1) and normalised by all possible values of the distribution – given by Equation (4.3):

$$p(\theta|x) = \frac{h(\theta|x)}{m(x)} = \frac{f(x|\theta)p(\theta)}{\int f(x|\theta)p(\theta)\mathrm{d}\theta}. \tag{4.3}$$

The posterior distribution $p(\theta|x)$ can be viewed as a collation of all the previously collected information, i.e. the calibration data, and as discussed above, also includes all the uncertainty information.

When new data, $y$, are collected, the information contained in Equation (4.1) can be used as prior information for the calculation of the distribution of the new data given the calibration data, as follows. Given a model distribution: $f(y|\theta)$, for with one or more parameters $\theta$, the posterior predictive distribution $f_p(y|x,d)$ for data $x$ given new data $y$ and a fixed dose $d$, is defined as the integral of the joint distribution of the model $f(y|\theta,d)$ and $p(\theta|x)$ over all possible values of $\theta$:

$$f_p(y|x,d) = \int f(y|\theta,d)p(\theta|x)\mathrm{d}\theta. \tag{4.4}$$

Note that in the posterior predictive distribution $d$ acts like a parameter. In order to calculate the distribution of dose, the calibrative density $f(d|x,y)$

is calculated using Equation (4.5), which incorporates prior information for the dose distribution $p(d)$:

$$f_p(d|x,y) = p(d) f_p(y|x,d). \tag{4.5}$$

The derivation for Equation (4.5) is given in Groer and Pereira [33] and thus is not repeated here.

### 4.4.2 Worked example for the Poisson model

Here we demonstrate how the above method can be applied to estimation of radiation dose, given data $x$ from a new measurement and data from a previously defined calibration curve.

In order to simplify the example as far as possible, we will assume a linear calibration relationship between the number of dicentric chromosome aberrations per cell $y$ and dose $d$, $y = \alpha d$; as is common for high LET radiations, for example fission neutrons [33]. In accordance with the IAEA manual [42], a linear dose response should also be applied for cases of protracted exposure to low LET radiation.

The first and perhaps most important step in any Bayesian analysis is model selection. As discussed, the Poisson model has long been the model of choice for cytogenetic biodosimetry. Thus we will assume a Poisson model for both the distribution of each observation of the calibration data $x_i$ and for the distribution of the future data $y$:

$$f(x_i|\alpha) = \frac{(\alpha d_i)^{x_i} \mathrm{e}^{-\alpha d_i}}{x_i!}. \tag{4.6}$$

$$f(y|\alpha, d) = \frac{(\alpha d)^{y} \mathrm{e}^{-\alpha d}}{y!}. \tag{4.7}$$

We assume that each observation $x_i$ follows a Poisson distribution with expectation equal to $\alpha d_i$ where $d_i$ is the dose corresponding to this observation. The probability function, which also known as the likelihood function, for the complete calibration data model for $N$ measurements with $x_i$ aberrations in each cell at dose $d_i$ becomes:

$$f(x|\alpha) = \prod_{i=1}^{N} \frac{(\alpha d_i)^{x_i} \mathrm{e}^{-\alpha d_i}}{x_i!} \propto \alpha^{\sum_{i=1}^{N} x_i} \mathrm{e}^{-\alpha \sum_{i=1}^{N} d_i} \tag{4.8}$$

The likelihood function (4.8) can be summarized using the total number of aberrations $z_i$ observed for each different dose $d_i$ $(i = 1, 2, \ldots, k)$ in $n_i$ cells, obtaining the following expression:

$$f(x|\alpha) \propto \alpha^{\sum_{i=1}^{k} z_i} \mathrm{e}^{-\alpha \sum_{i=1}^{k} n_i d_i}. \tag{4.9}$$

It should be noted that Equation (4.9) arises because the sum of the observed aberrations is a sufficient statistic for the parameter of the Poisson distribution. Such a simplification is not possible for other distributions like the negative binomial where the likelihood function has to be calculated using the observations cell by cell, if we do not want to lose information.

To calculate the calibrative density in this example, the parameter $\theta$ is thus replaced by $\alpha$ in Equations (4.1)–(4.4). In order to completely include all possible uncertainty information and begin the procedure of calculating the posterior predictive distribution of the dose, we must also specify a prior model for the model parameter $\alpha$, $p(\alpha)$ and for the dose, $p(d)$. For this example, we will use the gamma distribution for both $p(\alpha)$ and $p(d)$. The gamma distribution is the 'conjugate prior' for Poisson distributed data:

$$p(\alpha) = \mathrm{Ga}(\alpha|a,b) = \frac{b^a}{\Gamma(a)}\alpha^{a-1}\mathrm{e}^{-b\alpha}. \tag{4.10}$$

$$p(d) = \mathrm{Ga}(d|A,B) = \frac{B^A}{\Gamma(A)}d^{A-1}\mathrm{e}^{-Bd}. \tag{4.11}$$

In mathematical terms, the use of the conjugate prior means that when the prior and data are combined, the posterior will be of the same or similar form, thus we are justified in choosing this 'Poisson like' model to represent the distribution of $\alpha$. The practical advantage of this is that analytical treatment of Equations (4.1)–(4.5) will be very much simplified. $\mathrm{Ga}(\alpha|a,b)$ in Equation (4.10) is the gamma distribution of the parameter $\alpha$, with distribution parameters $a$ and $b$ which refer to the 'shape' and 'rate' of the distribution, respectively. $\Gamma(a)$ is the gamma function which is related to the factorial function in that $\Gamma(a) = (a-1)!$ for positive integers. Similarly for Equation (4.11), $A$ and $B$ are the shape and rate parameters of the gamma prior distribution of the dose, $d$.

Once the distributions are defined, the next task in the method outlined above is to calculate the joint distribution of the prior $p(\alpha)$ and the calibration data $f(x|\alpha)$. Using Equation (4.1):

$$h(\alpha|x) \propto \alpha^{\sum_{i=1}^k z_i}\mathrm{e}^{-\alpha\sum_{i=1}^k n_i d_i}\frac{b^a}{\Gamma(a)}\alpha^{a-1}\mathrm{e}^{-b\alpha} \propto \alpha^{\sum_{i=1}^k z_i + a - 1}\mathrm{e}^{-\alpha(\sum_{i=1}^k n_i d_i + b)}. \tag{4.12}$$

The properties of the conjugate prior mean that $h(\alpha|x)$ is in fact in the form of a gamma distribution with modified parameters:

$$\mathrm{Ga}\left(\alpha \,\middle|\, a + \sum_{i=1}^k z_i, b + \sum_{i=1}^k n_i d_i\right).$$

In order to calculate the posterior distribution of the calibration data it is not necessary to calculate the marginal distribution $m(x)$ expressed in Equation (4.2). In fact, $m(x)$ only introduces in Equation (4.12) the corresponding normalizing constant such that Equation (4.12) becomes a proper density of $\alpha$. Consequently, the posterior distribution, $p(\alpha|x)$, becomes a gamma distribution:

$$p(\alpha|x) = \frac{h(\alpha|x)}{m(x)} = \frac{(b + \sum_{i=1}^k n_i d_i)^{a+\sum_{i=1}^k z_i}}{\Gamma(a + \sum_{i=1}^k z_i)}\alpha^{a+\sum_{i=1}^k z_i - 1}\mathrm{e}^{-\alpha(b+\sum_{i=1}^k n_i d_i)}. \tag{4.13}$$

If the newly collected data in the form of $y$ total aberrations in $n_y$ cells, for dose $d$, is represented by a Poisson distribution:

$$f(y|\alpha,d) = \frac{(\alpha n_y d)^y \mathrm{e}^{-\alpha n_y d}}{y!}, \tag{4.14}$$

according to Equations (4.4), the posterior predictive distribution, $f_p(y|x,d)$, becomes:

$$f_p(y|x,d) = (n_y d)^y \frac{(b+\sum_{i=1}^k n_i d_i)^{a+\sum_{i=1}^k z_i}}{\Gamma(a+\sum_{i=1}^k z_i)y!} \int_0^\infty \alpha^{y+a+\sum_{i=1}^k z_i - 1} \mathrm{e}^{-\alpha(b+n_y d+\sum_{i=1}^k n_i d_i)} \mathrm{d}\alpha.$$

In order to evaluate the integral in the simplest possible way, we use a property of the gamma function: $\int_0^\infty z^{p-1}\mathrm{e}^{-qz}\mathrm{d}z = \Gamma(p)/q^p$, for all values $p, q > 0$. So, the part of the preceding expression that we wish to integrate:

$$\int_0^\infty \alpha^{y+a+\sum_{i=1}^k z_i - 1} \mathrm{e}^{-\alpha(b+n_y d+\sum_{i=1}^k n_i d_i)} \mathrm{d}\alpha = \frac{\Gamma(y+a+\sum_{i=1}^k z_i)}{(b+n_y d+\sum_{i=1}^k n_i d_i)^{y+a+\sum_{i=1}^k z_i}},$$

and therefore,

$$f_p(y|x,d) = (n_y d)^y \frac{(b+\sum_{i=1}^k n_i d_i)^{a+\sum_{i=1}^k z_i}}{\Gamma(a+\sum_{i=1}^k z_i)y!} \frac{\Gamma(y+a+\sum_{i=1}^k z_i)}{(b+n_y d+\sum_{i=1}^k n_i d_i)^{y+a+\sum_{i=1}^k z_i}}. \tag{4.15}$$

Equation (4.15) has the form of the probability function of a negative binomial distribution. The distribution of the dose (calibrative density) can then finally be calculated using Expression (4.16):

$$\begin{aligned} f(d|y,x) = &\quad p(d) f_p(y|x,d) \\ = &\quad \frac{B^A}{\Gamma(A)} d^{A-1}\mathrm{e}^{-Bd} (n_y d)^y \frac{(b+\sum_{i=1}^k n_i d_i)^{a+\sum_{i=1}^k z_i}}{\Gamma(a+\sum_{i=1}^k z_i)y!} \\ &\quad \frac{\Gamma(y+a+\sum_{i=1}^k z_i)}{(b+n_y d+\sum_{i=1}^k n_i d_i)^{y+a+\sum_{i=1}^k z_i}} \\ \propto &\quad \frac{d^{A+y-1}\mathrm{e}^{-Bd}}{(b+n_y d+\sum_{i=1}^k n_i d_i)^{y+a+\sum_{i=1}^k z_i}}. \end{aligned} \tag{4.16}$$

which is equivalent to Equation (4) in Groer and Pereira (1987) [33]:

$$f(d_f|D) \propto d_f^{B+y_f-1} \left( d_f + \frac{a+\sum_{i=1}^N n_i d_i}{n_f} \right)^{-(\sum_{i=1}^N x_i + y_f + b)} \mathrm{e}^{-A d_f}. \tag{4.17}$$

where $x_i$ are the calibration yields (our $z_i$), $D$ refers to all 'prior' information (i.e. the prior distributions of the calibration parameter and dose), $y_f$ is the measured yield in $n_f$ cells for the observed dose $d_f$, for which values of $f(d_f|D)$ are calculated, and $N$ is the different number of doses (our superindex $k$). Also note that in the notation of Groer and Pereira [33] the parameters $a$, $b$ and $A$, $B$ have been exchanged.

In order to use this equation, it is also necessary to insert appropriate values for $a$, $b$, $A$ and $B$ where $a$ and $b$ are the parameters of the prior gamma distribution of the calibration parameter $\alpha$ and $A$ and $B$ are the parameters of the prior gamma distribution of the dose. For a dosimetry situation where the type of radiation exposure is either known or at least suspected, the fact that the mean of a gamma distribution is calculated as $a/b$ and its standard deviation as $\sqrt{a}/b$, it can be used with a maximum likelihood procedure to estimate the parameters. The maximum likelihood estimator (MLE) of $\alpha$ is obtained

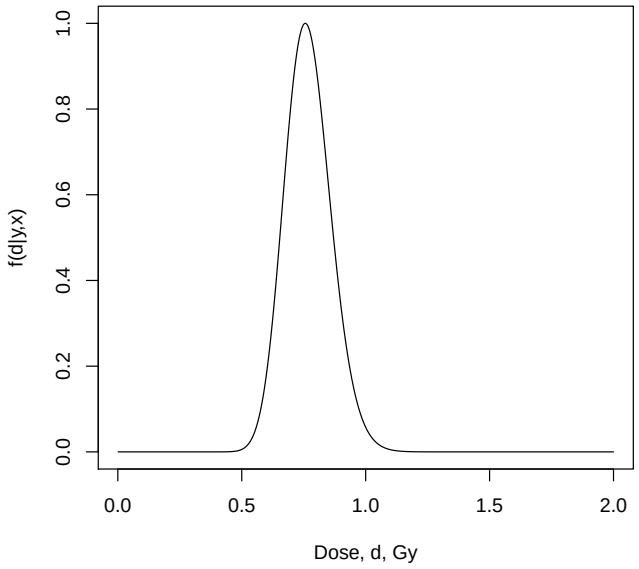


Figure 4.1: Normalised calibrative density for neutron exposure, based on Poisson distributed calibration data with gamma distributed prior information regarding the calibration coefficient, $\alpha$, and the distribution of doses, $d$, for an observed number of 64 dicentrics in 104 cells. Data taken from Groer and Pereira [33].

maximizing Expression (4.9) with respect to $\alpha$, obtaining the simple estimator, $\hat{\alpha} = \sum_{i=1}^{k} z_i / \sum_{i=1}^{k} n_i d_i$. Using the classical theory of MLE we can calculate an estimator of the standard deviation of $\hat{\alpha}$, by means of the reciprocal of the Fisher information. This estimator is $\hat{\sigma}(\hat{\alpha}) = \sqrt{\sum_{i=1}^{k} z_i} / \sum_{i=1}^{k} n_i d_i$.

For the calibration data of Groer and Pereira [33], these estimates are, $\hat{\alpha} = 0.833$ and $\hat{\sigma}(\hat{\alpha}) = 0.031$. Therefore, the use of a gamma distribution as a prior for $\alpha$, with $a = 722$ and $b = 867$, would agree with the MLE of $\alpha$ and its standard deviation estimate. However, following Groer and Pereira for this example, we take the values $a = 10$ and $b = 10$ [33]. Likewise for the dose parameters, $A$ and $B$, it is possible to use the expected distribution of dose, according to the irradiation scenario, to assign values to these parameters of $A = 10$ and $B = 10$ [33]. Inserting these values into Equation (4.16) gives the calibrative density distribution that is shown in Figure 4.1.

Using this method, the best estimate of dose (the modal dose) is found to be 0.756 Gy. The density itself can then be used to calculate probabilities, for instance there is a 95% probability that the dose was between 0.597–0.961 Gy and, given the calibration data and the gamma priors for dose and the alpha coefficient and an observation of 64 dicentrics in 104 cells, there is a 98.97% probability that the dose was <1 Gy.

It is important to remark that, in general, the influence in the calibrative

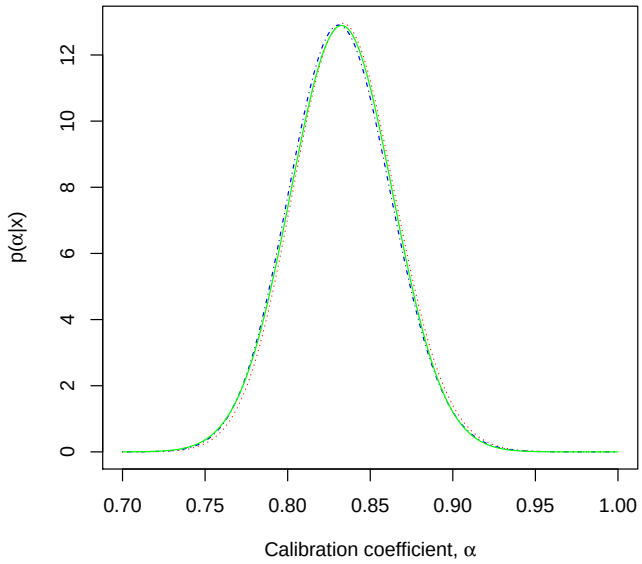


Figure 4.2: Posterior densities $p(\alpha|x)$ for a $\text{Ga}(\alpha|10, 10)$ prior (red/dotted line), a Jeffreys prior (blue/dash–dot line) and the normal approximation $\text{N}(0.833, 0.031)$ (green/solid line). Data taken from Groer and Pereira [33].

density of the prior dose density $p(d)$ is greater than the influence of the prior density of the parameter $\alpha$. In fact, for large sample size calibration data the posterior distribution of $\alpha$ described in Equation (4.1) and (4.12), tends to a Gaussian distribution with mean equal to the MLE $\hat{\alpha}$ and standard deviation equal to $\hat{\sigma}(\hat{\alpha})$ [26]. This fact is independent of the prior density considered. So for large sample size data, the normal approximation of the posterior distribution is always a good option, and in this situation it is not necessary to choose a prior density for $\alpha$. This is the case for the example analyzed here. Figure 4.2 shows the posterior density obtained using a prior, the Jeffreys prior $p(\alpha) \propto 1/\sqrt{\alpha}$ and the normal approximation: The three densities are almost indistinguishable. Note that the posterior for the Poisson model with a Gamma prior or a Jeffrey's prior can be calculated analytically.

In contrast, the choice of prior dose density $p(d)$ has a relatively large influence on the calibrative density. For example, if the prior dose had been assumed to be in the region of 2 Gy, the parameters of the gamma prior could have been assigned values of $A = 10$ and $B = 10$. In this case, the best estimate of dose becomes 0.868 Gy.

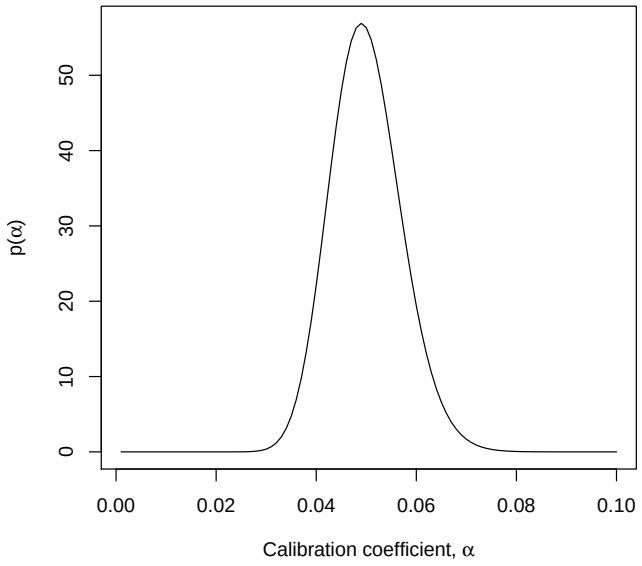


Figure 4.3: Prior gamma distribution for alpha coefficient: $\mathrm{Ga}(\alpha|50, 1000)$ with theoretical mean alpha coefficient $\sim 0.050$.

**An example using classical and Bayesian methods of cytogenetic dose estimation**

The below example is presented using calibration data published by Lloyd and colleagues in 1984 [51], for an assumed suspected dose of 1 Gy, a measured yield of dicentric chromosome aberrations of 50 in 1000 cells scored. The authors fitted dose response curves for chromosome aberration data formed following protracted exposures to Cobalt 60 gamma irradiation. The 12 hr exposure time data were fitted to a linear quadratic dose response, however the first four data points, up to approximately 2 Gy, give a nice fit to a linear model, making this data set suitable for the calibrative density analysis described above.

The calibration data are thus as follows: Doses, $d_i$ (Gy)= (0,0.28,0.534,0.994, 2.04); numbers of cells, $n_i = (10000, 1033, 500, 600, 700)$; total count of dicentrics, $z_i = (5, 8, 12, 30, 81)$. Note that the zero Gy data point is not given explicitly in the reference, although it is stated that a 0 Gy dose point was used for each fit. The yield at zero Gy is given in the text, and thus the numbers of cells and dicentrics given here for the 0 Gy data point are estimated values which allow this yield to be reproduced.

The first step is to estimate the gamma priors for the alpha coefficient and for dose. For the alpha coefficient, the prior can be constructed using the MLE, obtaining $\hat{\alpha} = 0.052$ and $\hat{\sigma}(\hat{\alpha}) = 0.0045$. The Dose Estimate program [1] was used to investigate the fitting of these five data points from a classical point of view. The $z$–test $p$ value for the alpha coefficient was 0.012, indicating its significance in determining the fit, and the $p$ value for the quadratic coefficient

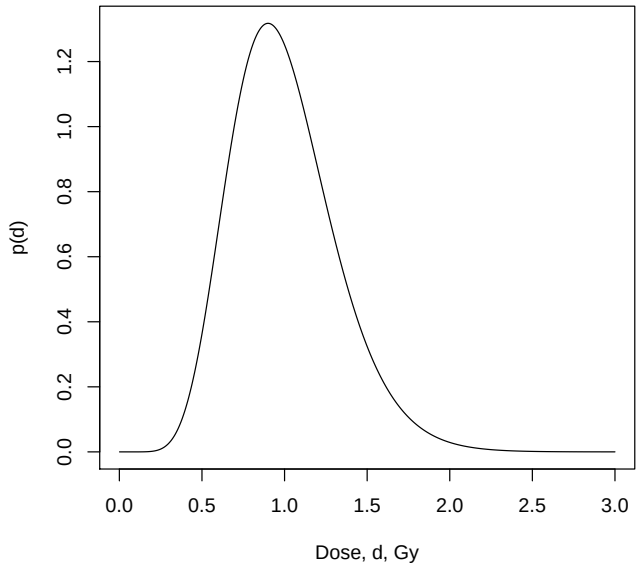


Figure 4.4: Prior gamma distribution for dose: $\mathrm{Ga}(\alpha|10, 10)$ with theoretical mean dose of 1 Gy.

in the case of a linear quadratic fit was 0.916, indicating that the quadratic part of the curve has no significance with these data. The gamma prior is constructed following the procedure described before: A gamma shape coefficient of 50 and a gamma rate coefficient of 1000 give an appropriate theoretical mean (0.050) very close to the MLE, as well as a theoretical standard deviation which is slightly greater than $\hat{\sigma}(\hat{\alpha})$: 0.007. Inspection of the form of the gamma distribution $\mathrm{Ga}(\alpha|50, 1000)$ reveals it is distributed fairly widely around the MLE of 0.053 (Figure 4.3). Furthermore, several examples in the literature are available (including Groer and Pereira 1987 [33]), indicating values of the gamma coefficients on this order. Therefore values of gamma $a = 50$ and $b = 1000$ can be used to create an appropriate prior distribution for the alpha coefficient of the dose response curve.

If a dose of 1 Gy is suspected, then it is reasonable to assume that the true dose might lie within a gamma distributed region around a dose of 1 Gy. Inspection of the gamma distribution $\mathrm{Ga}(\alpha|50, 1000)$ (Figure 4.4) reveals a sensible range for the estimated dose, therefore these values can be used to create the gamma prior for a suspected dose of 1 Gy.

As described above, the priors can then be used with the calibration data to create the normalised calibrative density as per Equation (4.16), with values of $a = 50$, $b = 1000$, $A = 10$, $B = 10$, $y = 50$, $n_y = 1000$, and the calibration data $d_i$, $n_i$ and $z_i$ given above. The resulting calibrative density is illustrated in Figure 4.5.

The results reveal that the modal dose value is 0.954 Gy, with 95% credibility

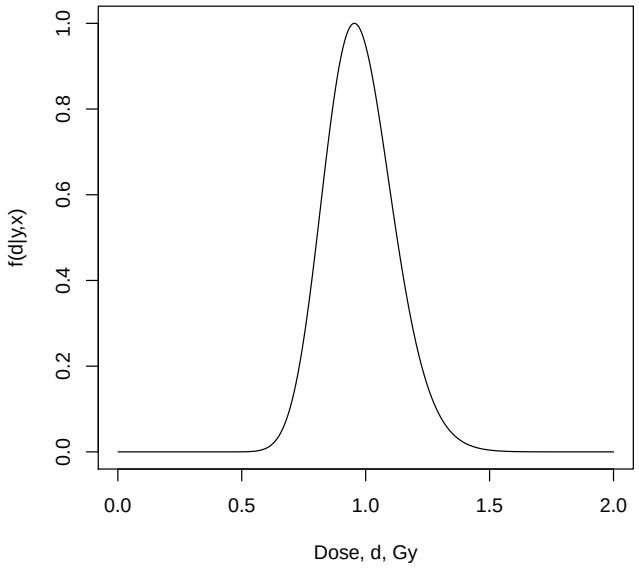


Figure 4.5: Normalised calibrative density, $f(d|y,x)$, for a measured number of 50 dicentrics in 1000 cells, for gamma prior for dose Ga$(10,10)$ (mean $\sim$1 Gy) and gamma prior for alpha coefficient Ga$(50,1000)$ (mean $\sim$0.050).

interval between 0.726 and 1.272 Gy. There is a 58.54% chance that the dose was $<$1 Gy.

If the classical method of dose estimation is used, the calibration data would first be fitted using standard maximum likelihood or weighted linear regression techniques (e.g. using the Dose Estimate software; Ainsbury and Lloyd 2010) to give a relationship between yield, $y$, and dose, $D$, of: $y = 0.0005(\pm 0.0003) + 0.0503(\pm 0.0003)D$. Note that this is a linear model with intercept. This line can then be used to calculate an estimated dose of $0.984 \pm 0.137$ by means of the inverse regression technique. The associated 95% confidence limits, which are calculated following the recommended approximation to combine the error on the measurement with the error on the curve [42] are 0.715 to 1.253. It is important to remark that this is a confidence interval, conceptually different from the credibility interval provided by the Bayesian method. The Bayesian method gives the probability that the dose was in a certain dose range (aside from the confidence limits) or above or below a certain value, e.g. the 1 Gy suspected dose given here. Moreover, the Bayesian method of dose calculation allows full incorporation of any 'prior' information, e.g. the suspected dose given in this example. A classical method that has been developed to deal with this is the odds ratio approach of Szluińska, which allows the suspected dose and calculated dose to be compared [92]. However, this approach only allows consideration of a single "point" estimate of suspected dose, ignoring the fact that the suspected or likely dose is much more accurately represented as a

distribution of possible values.

# Chapter 5

# A new inverse regression model applied to radiation biodosimetry

This chapter corresponds to the contents of [39].

**Abstract:** Biological dosimetry based on chromosome aberration scoring in peripheral blood lymphocytes enables timely assessment of the ionising radiation dose absorbed by an individual. Here, new Bayesian-type count data inverse regression methods are introduced for situations where responses are Poisson or two-parameter compound Poisson distributed. Our Poisson models are calculated in a closed form, by means of Hermite and Negative Binomial distributions. For compound Poisson responses, complete and simplified models are provided. The simplified models are also expressible in a closed form and involve the use of compound Hermite and compound negative binomial distributions. Three examples of applications are given that demonstrate the usefulness of these methodologies in cytogenetic radiation biodosimetry and in radiotherapy.

**Keywords:** calibrative density, compound Poisson distribution, Hermite distribution, Panjer recursion.

## 5.1 Introduction

In spite of strict safety measures and regulations, radiation accidents or unplanned exposures occur, for instance in radiology services and radiotherapy departments at hospitals, or using radiography cameras in industry. There have also been some major radiation/nuclear accidents, such as Chernobyl or Fukushima, that have affected many people [91]. In the event of a radiation accident, biological dosimetry is essential for the timely determination of the radiation dose to which an individual has been exposed. On the other hand, radiotherapy is commonly used to treat cancerous tumors, and it is important to know the total absorbed blood dose to prevent possible complications or side effects. Biological dosimetry relies on quantifying the amount of damage induced by radiation at a cellular level, for instance by counting dicentrics or micronuclei.

These aberrations appear because when cells are exposed to radiation, breaks are induced in the chromosomal DNA, and the broken fragments may rejoin incorrectly. Therefore, the frequency of chromosome aberrations increases with the amount of radiation and is a reliable and very well established biological indicator of radiation absorbed dose. Such information supports the clinical management of a patient, enables rapid triage in the case of a large scale radiation incident and reassures the 'worried well' that they have not received a severe radiation dose. At high acute whole body doses above 2 Gy, haematopoietic failure (or myelodysplasia) is the primary threat associated with acute radiation syndrome which can be supported by early treatment with cytokines or, at very high doses, bone marrow transplants [19]. To estimate the dose absorbed by an individual, dose-effect calibration curves are required which are produced by irradiating peripheral blood lymphocytes to a range of doses. The protocol and methodology for such calibration experiments is described in a recent manual of the International Atomic Energy Agency [42].

The usual approach for constructing the calibration curve is to irradiate $n$ blood samples from various healthy donor with several doses $x_i$, $i = 1, \ldots, n$. Then, for each irradiated sample, $n_i$ cells are examined and the numbers of observed chromosomal aberrations $y_{ij}$, $j = 1, \ldots, n_i$ is recorded. For the dicentric assay it is usually assumed that the counts $y_{ij}$ follow a Poisson distribution [22] or a compound Poisson distribution [68] whose mean is a function of $x_i$ and a set of parameters $\beta$, i.e. $\mathrm{E}(y_{ij}) = \mathrm{f}(x_i, \beta)$. From the point of view of [42] $\beta$ are the calibration coefficients and $\mathrm{f}(x_i, \beta)$ is the mean of aberrations per cell (called yield or frequency of aberrations per cell, in the cytogenetics field). The parameters of this regression model are usually estimated by maximum likelihood [23], and the MLE and its estimated variance–covariance matrix are calculated and recorded. Therefore, in the case of an irradiated patient, a blood sample is taken and $m$ lymphocytes are scored obtaining the counts $\tilde{y}_1, \ldots, \tilde{y}_m$. The classical approach to estimate the absorbed dose $x$ and its confidence limits is to use the inverse regression method of [61], also described as a standard procedure in [42]. An improved classical inverse regression method applied to Electron Paramagnetic Resonance (EPR) dosimetry is found in [17].

Bayesian approaches allow simple incorporation of prior information concerning the circumstances of the exposure. Groer and Pereira [33] were the first to investigate the use of Bayesian models in chromosome dosimetry, for neutron exposure, and since then several researchers have used Bayesian methods in radiation biodosimetry. For instance, DiGiorgio and Zaretzky [21] used a Bayesian approach to present the uncertainty on a biological dose estimate for a radiation overexposed patient in Latin–America: a Poisson model with a Jeffrey's prior was used and it was further demonstrated that the Bayesian approach allows presentation of probabilities for dose ranges, which leads to a much more intuitive interpretation of the biological dosimetry results. A review of these methods can be found in [4]. There is also one recent program, CytoBayesJ [3], which provides some basic software tools for Bayesian analysis of cytogenetic radiation dosimetry data.

In this chapter we present a new Bayesian–type method to use cytogenetic data to estimate the dose to which a patient has been exposed. This method uses dose–effect calibration curves estimated by the classical (frequentist) approach suggested in the IAEA manual. Therefore, our new method has the advantage that allows reanalysis of many of the published examples of radiation exposures

that were studied using the classical methods. In addition, the method is in fact a general inverse regression model for count responses that could also be applied in contexts other than radiation biodosimetry.

An R package called 'radir' [65], which implements the Poisson response models presented here, is available in CRAN repository: `http://cran.r-project.org/web/packages/radir/index.html`.

## 5.2 A Bayesian-type inverse regression model

The Poisson distribution is usually used to describe the distribution of dicentric chromosomes per cell when the patient has been irradiated with small doses and with a low linear energy transfer (low–LET radiation). However, after exposure to high–LET, acute radiation, the distribution of dicentrics per cell often presents overdispersion and therefore compound Poisson distributions are preferred. The commonly compound Poisson distributions in biodosimetry are the Neyman A (NA) [96], the negative binomial (NB) [14], and recently the family of r*th*–order univariate Hermite distributions [78]. These compound Poisson distributions, also known as stopped–Poisson distributions, can be justified by a simple physical model of chromosomal aberration formation: the particles traverse the cell nucleus following a Poisson process and, for each particle, there is a probability (the generalizing distribution) to produce $k$ aberrations. Then the number of aberrations follows a compound Poisson distribution. In other words, a random variable $Y$ follows a compound probability distribution if it can be represented by

$$Y = \sum_{i=1}^{N} \xi_i, \tag{5.1}$$

where $N$ is a count data random variable and $\xi_1, \xi_2, \ldots$ are independent, identically distributed random variables that are also independent of $N$. In the case where $N$ is Poisson, $Y$ is said to follow a compound Poisson distribution. The distribution of $\xi_i$ is called the generalizing distribution. In particular when the distribution of $\xi_i$ is Poisson, the distribution of $Y$ is a Neyman A, when $\xi_i$ follows a Logarithmic distribution, $Y$ is negative binomial distributed, and when $\xi_i$ is distributed as a binomial with a number of trials equal to 2, then $Y$ follows a Hermite (Herm) distribution [44]. This can be expressed according to the Gurland's notation ([44], [35]) as $N \bigvee \xi$. In particular, parameterising with respect to the population mean $\mu$ and dispersion index $\delta$ (the ratio of the variance to the mean $\sigma^2/\mu$) we have the symbolic representation,

- $\text{NA}(\mu, \delta) \sim \text{Pois}\left(\frac{\mu}{\delta - 1}\right) \bigvee \text{Pois}(\delta - 1)$
- $\text{NB}(\mu, \delta) \sim \text{Pois}\left(\frac{\mu \log(\delta)}{\delta - 1}\right) \bigvee \text{Log}\left(\frac{\delta - 1}{\delta}\right)$
- $\text{Herm}(\mu, \delta) \sim \text{Pois}\left(\frac{\mu}{2(\delta - 1)}\right) \bigvee \text{Bin}(2, \delta - 1)$

Properties, formulae and algorithms to calculate the probabilities of these distributions can be found in [44]. In brief, they are partially closed under addition [77], the maximum likelihood estimator of the population mean is the sample

mean and they are also members of the discrete exponential dispersion family of distributions. These properties are shared with other distributions potentially useful in biodosimetry, such as Polya Aeppli or Poisson-Inverse Gaussian. See [77] for more properties and characterizations of these distributions. In particular, given a random variable $Y$ (with mean $\mu$ and dispersion index $\delta$) belonging to one of these models, the sum of $n$ independent copies of $Y$ also belongs to the same model having the same dispersion index and a mean equal to $n\mu$. Moreover, if $\delta$ is known, the sum of the observations is a sufficient statistic for $\mu$, containing all the information of the model. This is an important property that will be used in section 4.

Let $D = \{(x_i, y_{ij})\}$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$ be a calibration data set where each $y_{ij}$ represents a count data observation which will be assumed to follow a Poisson distribution or a two–parameter compound Poisson distribution. Here $x_i$ are the values of the independent variable, dose in the case of cytogenetic radiation biodosimetry. The number of different exposed doses is $n$ and $n_i$ is the sample, the number of blood cells for the $i^{th}$ dose. For all the models we define the regression function $\mathrm{E}(y_{ij}) = \mathrm{f}(x_i, \beta)$, $\beta \in \mathbb{R}^p$. Moreover, for compound Poisson modeling, we assume that the dispersion index is a constant ($\delta$). In practice, this assumption could be verified by plotting the empirical values of the dispersion index ($s^2_{y_i}/\bar{y}_{i\cdot}$) against the $x_i$. However, we could assume another relationship between the independent variable and the dispersion index. Therefore, from now, we will consider the dispersion coefficient $\delta$ not to depend on $x_i$, and then the domain of the parameters is $\Theta = \{\beta, \delta\}$. Note that for the Poisson model $\delta = 1$ and the domain of the parameters is just $\Theta = \{\beta\}$.

Let $\mathrm{p}(y_{ij} = k) = \mathrm{p}(k|\mu, \delta)$ be the probability mass function of the model, parameterized in terms of its population mean and dispersion index. It is clear that $\mathrm{p}(y_{ij} = k) = \mathrm{p}(k|f(x_i, \beta), \delta) = \mathrm{p}(k|x_i, \Theta)$, and then the likelihood function of the calibration data $D$ becomes:

$$\mathrm{L}(D|\Theta) = \prod_{\substack{i=1,\ldots,n \\ j=1,\ldots,n_i}} \mathrm{p}(y_{ij}|x_i, \Theta). \tag{5.2}$$

According to the IAEA manual, the parameters are estimated by maximizing the likelihood function (5.2), obtaining $\hat{\Theta} = \{\hat{\beta}, \hat{\delta}\}$. It is well known that for large data samples, the distribution of $\Theta \in \mathbb{R}^{p+1}$ can be approximated by a multivariate Gaussian distribution $\mathrm{N}_{p+1}(\hat{\Theta}, \hat{\Sigma}_{\hat{\Theta}})$ where $\hat{\Sigma}_{\hat{\Theta}}$ is its estimated variance-covariance matrix, that is, the inverse of the estimated Fisher information matrix of the model. Note, however, that in the frequentist framework $\hat{\Theta} \sim \mathrm{N}_{p+1}(\Theta, \hat{\Sigma}_{\hat{\Theta}})$. It is important to remark that the laboratory providing the outputs of the calibration curve, that is $\hat{\Theta}$ and $\hat{\Sigma}_{\hat{\Theta}}$, could be different from the one analysing the patient sample; even though for a consistent assay, the calibration curve should be constructed with the data provided by the same laboratory that will analyze the patient data to guarantee that the scoring criteria applied for the construction of the curve are the same as those applied for patient analysis.

From here, the distribution of the expected count of dicentrics and dispersion index for a given dose of $x$, $(\mu, \delta)|x$ can be approximated by a bivariate normal distribution. This is a straightforward consequence of the multivariate delta method [85],

$$(\mu, \delta)|x \sim \mathrm{N}_2\left((\mathrm{f}(x, \hat{\beta}), \hat{\delta}), \nabla \cdot \hat{\Sigma}_{\hat{\Theta}} \cdot \nabla^t\right), \tag{5.3}$$

where $\nabla$ denotes the derivative of $(\mathrm{f}(x,\beta),\delta)$ at $(\hat{\beta},\hat{\delta})$, that is,

$$\nabla = \begin{pmatrix} \dfrac{\partial \mathrm{f}}{\partial \beta_0} & \cdots & \dfrac{\partial \mathrm{f}}{\partial \beta_p} & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Following these arguments, note that for the Poisson model the distribution of $\mu|x$ is approximated by a univariate Normal distribution with expectation $\mathrm{f}(x,\hat{\beta})$ and variance equal to $\mathrm{v}(x,\hat{\beta}) = \nabla \cdot \hat{\Sigma}_{\hat{\Theta}} \cdot \nabla^t$, where $\nabla$ is now the gradient of $\mathrm{f}(x,\beta)$ at $\hat{\beta}$. The bivariate normal density in (5.3) will be denoted as $\phi(\mu,\delta|x)$ and $\phi(\mu|x)$ will be the normal univariate density used for the Poisson model. In some situations, the use of a bivariate or univariate normal could be incompatible with the fact that $\mu > 0$, and in general $\delta > 1$. Then, some approximations have to be carried restricting the parameters' domain. For the univariate normal distribution, one solution is to replace it by a Gamma density with the same mean and variance. It is well known that a larger gamma distribution shape parameter (i.e., the ratio of the square of the mean to the variance) implies a better Normal approximation. As we will see in the next sections, the Normal approximation can be used in a wide range of situations, and it also will be compared with the Gamma approximation. For our purposes $\mu|x$ will be called the *mean prior distribution*, because it will act as a prior for the inverse regression estimation problem.

Consider the test (patient) data $\tilde{y} = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_m\}$, formed by $m$ count data observations depending on an unknown regressor $x$ that we aim to estimate. The likelihood function of the test data becomes,

$$\mathrm{L}(\tilde{y}|\mu,\delta) = \prod_{i=1}^{m} \mathrm{p}(\tilde{y}_i|\mu,\delta). \tag{5.4}$$

Note that, because the knowledge of $\mu$ implies the knowledge of $x$, then we can write $\mathrm{L}(\tilde{y}|\mu,\delta) = \mathrm{L}(\tilde{y}|\mu,\delta,x)$. Therefore, an application of Bayes' theorem shows the expression of the posterior density of the parameters given the test data,

$$\mathrm{f}(\mu,\delta,x|\tilde{y}) = \frac{\mathrm{L}(\tilde{y}|\mu,\delta)\mathrm{p}(\mu,\delta,x)}{\displaystyle\int \mathrm{L}(\tilde{y}|\mu,\delta)\mathrm{p}(\mu,\delta,x)\mathrm{d}\mu\mathrm{d}\delta\mathrm{d}x},$$

where $\mathrm{p}(\mu,\delta,x)$ is the joint prior density of $\mu$, $\delta$ and $x$. But, $\mathrm{p}(\mu,\delta,x) = \phi(\mu,\delta|x)\mathrm{p}(x)$, where $\mathrm{p}(x)$ summarizes the prior information for $x$. This prior information can come from the characteristics of the radiation accident, such as the source and the duration of the exposure, etc.

Therefore, marginalizing over $\mu$ and $\delta$ we obtain the *calibrative density* of $x$, that it is the solution of the inverse regression problem:

$$\mathrm{f}(x|\tilde{y}) \propto \mathrm{p}(x) \int \mathrm{L}(\tilde{y}|\mu,\delta)\phi(\mu,\delta|x)\mathrm{d}\mu\mathrm{d}\delta\,. \tag{5.5}$$

As it will be shown in Section 5.3, this calibrative density can be exactly calculated for the Poisson model, solving completely the problem of the absorbed dose estimation in the most frequent situation.

However, for the two–parameter compound Poisson models the integral in (5.5) does not have a closed form, thus some approximations are required such as numerical integration or simulation methods. For this reason, the model will be simplified in Section 5.4.

## 5.3 The Poisson model

When data is Poisson distributed, the likelihood function of the test data has the form,

$$\mathrm{L}(\tilde{y}|\mu) \propto \prod_{i=1}^{m} \mathrm{p}(\tilde{y}_i|\mu) \propto e^{-m\mu}\mu^{\sum_{i=1}^{m}\tilde{y}_i}.$$

Because the sum of the observations is a sufficient statistic for the parameter of Poisson data, and the sum of independent Poisson random variables is also Poisson distributed, this likelihood function is equivalent to the probability function of one Poisson observation evaluated at $s$, that is,

$$\mathrm{L}(\tilde{y}|\mu) \propto \mathrm{p}(s|m\mu) \propto e^{-m\mu}(m\mu)^s,$$

where $s = \sum_{i=1}^{m} \tilde{y}_i$. Therefore, the calibrative density (5.5) remains,

$$\mathrm{f}(x|\tilde{y}) = \mathrm{p}(x)\mathrm{q}_s(x), \tag{5.6}$$

where,

$$\mathrm{q}_s(x) = \int_{-\infty}^{\infty} \mathrm{p}(s|m\mu)\phi(\mu|x)\mathrm{d}\mu\,. \tag{5.7}$$

Note that (5.7) represents the probability function of a mixed Poisson-normal distribution evaluated at $s$. Of course, strictly speaking, it is not possible to mix a Poisson with a Normal distribution because the Poisson parameter always has to be positive. However, understanding this mixture as a purely formal operation, Kemp and Kemp [48] showed that this mixed Poisson distribution, provided the population mean of the mixing Normal is greater than its variance, is just the Hermite distribution. Specifically, using Gurland's notation ([44], [35]) we have the symbolic representation,

$$\mathrm{Pois}(m\mu) \bigwedge_{\mu} \mathrm{N}(a, b^2) \sim \mathrm{Herm}(ma, 1 + mb^2/a).$$

This notation means that the $\mu$ parameter in the Poisson distribution (left part) is normally distributed (right part). This representation is valid only for $a \geq mb^2$.

Consequently, (5.7) is the probability that a Hermite random variable takes a value equal to $s$. Specifically, it can be directly shown that the probability (5.7) can be obtained from the Hermite probability recursion described in [47],

$$(r+1)\mathrm{q}_{r+1}(x) = (m\mathrm{f}(x,\hat{\beta}) - m^2\mathrm{v}(x,\hat{\beta}))\mathrm{q}_r(x) + m^2\mathrm{v}(x,\hat{\beta})\mathrm{q}_{r-1}(x),$$

with $\mathrm{q}_0(x) = \exp(-m\mathrm{f}(x,\hat{\beta}) + m^2\mathrm{v}(x,\hat{\beta})/2)$ and defining $\mathrm{q}_{-1}(x) = 0$, provided that $\mathrm{f}(x,\hat{\beta}) - m\mathrm{v}(x,\hat{\beta}) \geq 0$. This last inequality is achieved for most of the studied examples, for the range of interest of the absorbed dose $x$. In a hypothetical situation where this inequality was not achieved, that is $\mathrm{f}(x,\hat{\beta}) - m\mathrm{v}(x,\hat{\beta}) < 0$, expression (5.7) mathematically does not make sense (the dispersion coefficient cannot be greater than 2) and it is therefore better to replace the mean prior normal density $\phi(\mu|x)$ by a Gamma density $\Gamma(\mu|x)$ with the same mean $\mathrm{f}(x,\hat{\beta})$ and variance $\mathrm{v}(x,\hat{\beta})$. Then, expression (5.7) would remain,

$$\mathrm{q}_s(x) = \int_{0}^{\infty} \mathrm{p}(s|m\mu)\Gamma(\mu|x)\mathrm{d}\mu\,. \tag{5.8}$$

Because mixing a Poisson with a Gamma produces a negative binomial distribution, it can be shown that $\mathrm{q}_s(x)$ in (5.8) is the probability that a negative binomial random variable, with mean $m\mathrm{f}(x,\hat{\beta})$ and variance $m^2\mathrm{v}(x,\hat{\beta})+m\mathrm{f}(x,\hat{\beta})$, takes a value equal to $s$.

The method presented here for the Poisson model, using the Gamma distribution as a mean prior, is exactly the same as the full Bayesian method of Groer and Pereira [33] for the simple case where $\mathrm{f}(x,\beta)=\beta x$. However for other dose–response curves both methods differ. For this simple linear dose–response case, considering a Uniform dose prior, direct calculations show that,

$$\mathrm{f}(x|\tilde{y}) = \frac{m^{s+1}(\sum n_i x_i)^{\sum y_i}}{\mathcal{B}(s+1,\sum y_i - 1)} \frac{x^s}{(mx+\sum n_i x_i)^{s+\sum y_i}};$$

with mean, mode and variance of,

$$\mathrm{M}_{|\tilde{y}} = \frac{s}{m}\frac{\sum n_i x_i}{\sum y_i},$$

$$\mathrm{E}_{|\tilde{y}} = \frac{\sum n_i x_i}{m}\frac{\mathcal{B}(s+2,\sum y_i - 2)}{\mathcal{B}(s+1,\sum y_i - 1)},$$

$$\mathrm{V}_{|\tilde{y}} = \frac{\mathrm{E}_{|\tilde{y}}}{m}\cdot\left[\left(\sum n_i x_i - 2\right)\mathcal{B}\left(s+2,\sum y_i - 2\right) + 2\mathcal{B}\left(s+3,\sum y_i - 3\right)\right];$$

according to notation in Section 5.2, where $\mathcal{B}(\cdot)$ denotes Euler's Beta function. The distribution function of this calibrative density can be expressed in terms of the hypergeometric function.

The following example illustrates how this methodology is applied to a real data set.

### 5.3.1 Example: Cobalt–60 gamma rays irradiation

Here we consider data from an inter–laboratory comparison for the semi–automated dicentric assay undertaken as part of the Multibiodose project (a large scale European biodosimetry project) [82]. This data set (Table 5.1) is based on blood samples from 8 healthy donors which were irradiated *in vitro* with Cobalt–60 gamma rays at a high dose rate of 0.27 Gy/min simulating acute whole body exposure. The data presented here were collated and analysed using the Metafer 4 automated analysis system (MetaSystems, Altlussheim, Germany) at a single participating laboratory, using the 'BfS' image analysis classifier (system settings – further information in Romm et al., [82]).

The $u$ figures shown in Table 5.1 are the values of the $u$–test statistic of Rao and Chakravarti [80], which is a normalized sample dispersion index,

$$u = (d-1)\sqrt{\frac{n-1}{2(1-1/z)}},$$

where $d = s_y^2/\bar{y}$ is the sample dispersion coefficient, $n$ the sample size (number of cells), and $z = n\bar{y}$ the total number of count events (number of dicentrics). When $d$ is close to 1 then the data follow an equidispersed distribution. If the value of the $u$ statistic is higher (lower) than (-)2, the distribution can be considered over– (under–) dispersed. The $u$–test is suggested by the IAEA [42]

Table 5.1: Frequency distributions of the number of dicentrics after exposure to 6 doses of gamma–rays, and the sample means, dispersion coefficients and $u$ values for each distribution. Test data in italics.

| | Number of dicentrics | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dose (Gy) | 0 | 1 | 2 | 3 | 4 | $\bar{y}$ | $d$ | $u$ |
| 0.25 | 2185 | 8 | | | | 0.004 | 0.997 | -0.113 |
| 0.75 | 2550 | 44 | 1 | | | 0.018 | 1.026 | 0.952 |
| 1.00 | 2231 | 54 | 2 | | | 0.025 | 1.044 | 1.503 |
| *1.50* | *1712* | *96* | *3* | | | *0.056* | *1.003* | *0.092* |
| 2.50 | 1196 | 123 | 7 | 1 | | 0.105 | 1.038 | 0.985 |
| 3.00 | 1070 | 320 | 41 | 6 | 1 | 0.295 | 1.012 | 0.334 |

Table 5.2: BIC values using a second degree polynomial dose–response curve without constant term for the different models.

| Model | NB | Hermite | NA | Poisson |
|---|---|---|---|---|
| BIC | 4088.834 | 4085.594 | 4085.524 | 4079.639 |

and in fact it is equivalent to the classical Fisher dispersion test. According to the $u$ values shown in Table 5.1, equidispersion of the calibration data can be assumed, thus justifying the use of a Poisson regression model.

The 1.5 Gy row was removed from the calibration data set to be used as test data. This means that the true dose is known and it is possible to compare it with the resulting calibrative density. Following notation in Section 5.3, $s = 102$ and $m = 1811$, i.e. 102 scored dicentrics in 1811 blood cells.

In this example, for high dose rate gamma–radiation exposure, an appropriate dose–response curve, i.e. the regression model, is a second degree polynomial without intercept [42], $\mathrm{f}(x, \beta) = \beta_2 x^2 + \beta_1 x$ (Figure 5.1). In biodosimetry this is called the *linear–quadratic* dose–response curve. The intercept has been removed because we assume that for a dose $x = 0$ the expected number of dicentrics will be zero (for the 0 Gy sample there was only 1 dicentric in a total of 2592 blood cells). In general regression modelling, to analyze count data using a second degree polynomial mean response is not common, and a log–link mean response is the usual approach. However, in biodosimetry, the linear–quadratic dose–response curve has a biophysical interpretation [42] and is one of the most frequently employed in practice. Some problems could occur maximizing the likelihood function because $\beta_1$ and $\beta_2$ have to be necessarily positive. To ensure this, it is sometimes necessary to use numerical algorithms allowing constrains in the parameter domain.

Table 5.2 shows the Bayesian Information Criterion (BIC) values for the four different response distributions treated in this work from the calibration data. These values support the use of the Poisson model. So for a Poisson response the maximum likelihood parameter estimates and their estimated covariance matrix are the following:

- Fitted coefficients:

$$\hat{\beta}_1 = 3.126 \cdot 10^{-3}, \quad \hat{\beta}_2 = 2.537 \cdot 10^{-2}.$$

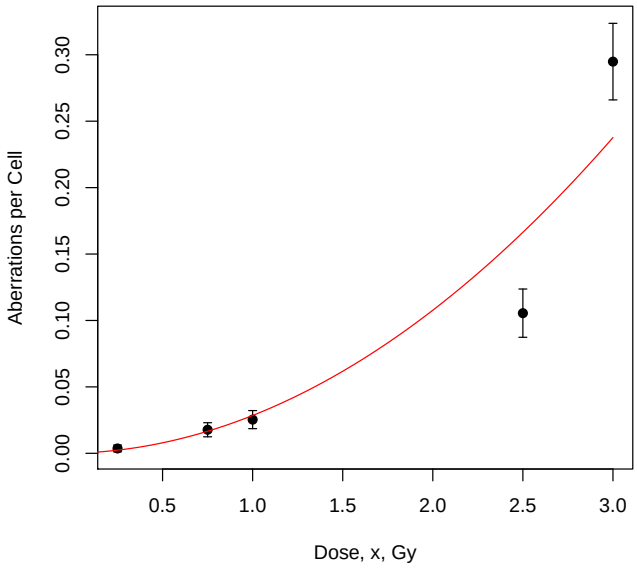


Figure 5.1: Observed means (dots), plus/minus twice their standard errors (error bars), and predicted means (red/solid line) of the number of dicentrics for Poisson fitting, based on the data in Table 5.1, omitting the 1.5 Gy test data.

- Estimated covariance matrix:

$$\hat{\Sigma}_{\hat{\beta}} = \begin{pmatrix} 7.205 & -3.438 \\ -3.438 & 2.718 \end{pmatrix} \cdot 10^{-6}.$$

As has been commented in section 5.2, $\mu|x$ will follow a Normal or a gamma distribution with mean $\mathrm{f}(x, \hat{\beta}) = \hat{\beta}_2 x^2 + \hat{\beta}_1 x$ and variance $\mathrm{v}(x, \hat{\beta}) = \nabla \cdot \hat{\Sigma}_{\hat{\beta}} \cdot \nabla^t$, where:

$$\nabla = \left( \frac{\partial \mathrm{f}}{\partial \beta_1}, \frac{\partial \mathrm{f}}{\partial \beta_2} \right) = \left( x, x^2 \right),$$

and therefore $\mathrm{v}(x, \hat{\beta}) = \hat{\Sigma}_{22} x^4 + 2\hat{\Sigma}_{21} x^3 + \hat{\Sigma}_{11} x^2$.

According to (5.7) and (5.8), for a normal or a gamma mean prior, the predictive posterior distribution $\mathrm{q}_{102}(x)$ represents the probability of a Hermite or negative binomial random variable taking a value of 102 counts, both with same mean $45.939x^2 + 5.661x$ and variance $8.913x^4 - 22.553x^3 + 69.571x^2 + 5.661x$.

Despite the real dose being known, firstly, a non-informative prior dose distribution is chosen in order to not take advantage of this fact, so $\mathrm{p}(x) \propto 1$. Secondly, for our purposes of comparing results, we define an informative prior dose distribution assuming we do not know the real dose of the test data, but we observe a mean of 0.056 dicentrics per cell, then by comparison with Table 5.1 it can reasonably be estimated that the dose is between 1 and 2.5 Gy. A simple

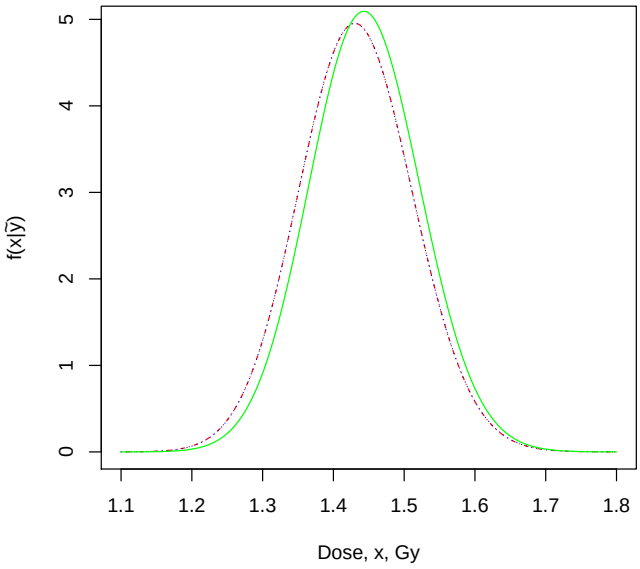


Figure 5.2: Calibrative densities of the 1.5 Gy test data calculated from a normal (blue/dotted line) and a gamma (red/dash–dot line) mean prior with non–informative prior dose distribution, and for a gamma mean prior with informative prior dose distribution (green/solid line). Red and blue curves are indistinguishable.

informative prior could be a gamma whose mean is in the midpoint of this interval, i.e. 1.75, and whose standard deviation is in the halfway from the mean to cover this interval, i.e. 0.375. For a gamma distribution with this mean and standard deviation, the 95.67% of the values fall in the region of $1.75 \pm 2 \cdot 0.375$.

Figure 5.2 shows the plot of the three densities of the estimated dose for the data test. Note how these results incorporate the real dose (1.5 Gy) and show the similarities found using both mean priors. Note that the gamma mean prior is moderately more conservative.

To use the normal mean prior (5.7) for this calibration set, the following condition must be satisfied: $\mathrm{f}(x, \hat{\beta}) - m\mathrm{v}(x, \hat{\beta}) \geq 0$. It holds when $x \leq 3.337$ Gy, and this could also be used as prior information about the dose, that is $\mathrm{p}(x) \sim \mathcal{U}(0, 3.337)$. For the range of the likely doses studied, the minimum value of the shape parameter of the mean prior gamma is 328.616, so the gamma or normal mean priors are practically indistinguishable.

The statistics of the three calibrative densities calculated in this example are shown in Table 5.3.

Table 5.3: Statistics summary of the calibrative densities for a normal (a) and a gamma (b) mean prior with non–informative prior dose distribution, and for a gamma mean prior with informative prior dose distribution (c).

| Model | Mode | Expected | SD | 95% CI |
|---|---|---|---|---|
| (a) | 1.430 | 1.432 | 0.081 | (1.277, 1.594) |
| (b) | 1.430 | 1.432 | 0.081 | (1.277, 1.593) |
| (c) | 1.443 | 1.445 | 0.078 | (1.294, 1.602) |

### 5.3.2 Example: Analysis of doses in thyroid cancer patients

This example illustrates how our methodology can be applied having only the fitted parameters of the dose–response curve, without knowing the calibration points. Serna et al. [86] studied chromosomal damage in lymphocytes of thyroid cancer patients after radioiodine treatment. The authors did a micronuclei assay in binucleated cells of blood samples from 25 patients 3 days after Iodine–131 (3.7 GBq) exposure.

The *in vitro* calibration curve was fitted by a linear–quadratic model with intercept, $\mathrm{f}(x,\beta) = G\beta_2 x^2 + \beta_1 x + \beta_0$ according to Poisson's law, and the estimate of $\beta_0$ was not taken into account, because the authors in [86] argued that the intercept could change for each patient. Constant $G$ is the Lea–Catcheside generalized dose–protraction factor, that modifies the quadratic term according to the temporal pattern of exposure, being $G = 1$ for the *in vitro* assay. The authors calculated the following parameter estimates $(\hat{\beta}_i \pm SE(\hat{\beta}_i))$,

$$\hat{\beta}_1 = (13.6 \pm 5.5) \cdot 10^{-3}, \quad \hat{\beta}_2 = (3.7 \pm 1.6) \cdot 10^{-2}, \quad \rho = -0.89,$$

where $\rho$ is the correlation coefficient for $\hat{\beta}_1$ and $\hat{\beta}_2$. The patients were subjected to ablative radioiodine treatments for post–surgical thyroid remnants. Consequently, they had a prolonged exposure lasting several days and which means, the temporal pattern of exposure was different than that of the *in vitro* assay. Taking into account the exposure profile of the Iodine–131 treatment, the authors in [86] found the factor G to be close to 0.1.

Then $\beta_0$, the background for each patient, can be estimated counting the micronuclei of the patient from a blood sample taken before the treatment, information provided in [86]. This leads to the fitted regression model $\mathrm{f}(x,\hat{\beta}) = G\hat{\beta}_2 x^2 + \hat{\beta}_1 x + \hat{\beta}_0$ with a covariance matrix that incorporates the variance of $\hat{\beta}_0$ without correlation with $\hat{\beta}_1$ and $\hat{\beta}_2$.

To illustrate our techniques we are going to estimate the absorbed dose for Patient 1, but the same can be done for the others. Patient 1 presented 487 normal cells and 13 cells with just one micronucleus each. Before the treatment 5 micronuclei where found in 500 blood cells, thus $\hat{\beta}_0 = (10 \pm 4.450) \cdot 10^{-3}$. The $u$–statistic of this test data is $-0.395$, so this is compatible with the Poisson model.

Therefore, $\mu|x$ will be considered to follow a distribution with mean $\mathrm{f}(x,\hat{\beta}) = G\hat{\beta}_2 x^2 + \hat{\beta}_1 x + \hat{\beta}_0$ and variance $\mathrm{v}(x,\hat{\beta}) = \nabla \cdot \hat{\Sigma}_{\hat{\beta}} \cdot \nabla^t$, where:

$$\nabla = \left( \frac{\partial \mathrm{f}}{\partial \beta_0}, \frac{\partial \mathrm{f}}{\partial \beta_1}, \frac{\partial \mathrm{f}}{\partial \beta_2} \right) = \left(1, x, Gx^2\right).$$

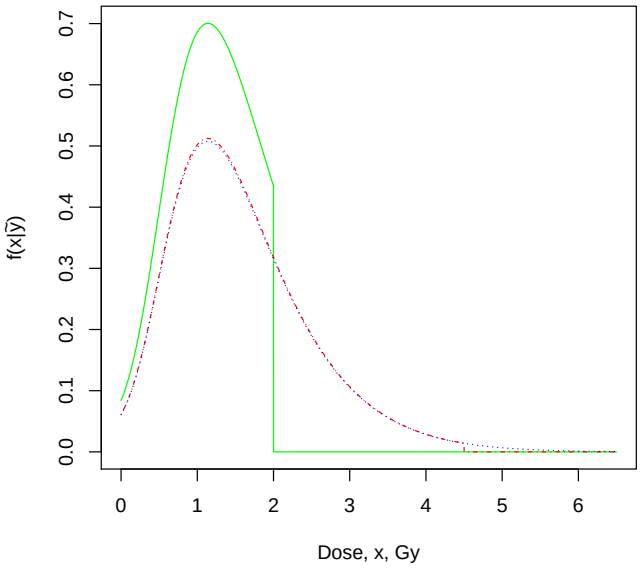


Figure 5.3: Calibrative densities of [86] Patient 1 test data calculated from a Gamma mean prior density, with a $\mathcal{U}(0,2)$ (green/solid line), a $\mathcal{U}(0,4.5)$ (red/dash-dot line) prior dose distribution and a improper $\mathcal{U}(0,+\infty)$ (blue/dotted line) prior dose distribution.

The condition $\mathrm{f}(x,\hat{\beta}) - m\mathrm{v}(x,\hat{\beta}) \geq 0$ is held when $x \leq 0.945$ Gy. This range of doses is very small for our purposes and consequently a gamma mean prior is preferred instead of a normal.

According to (5.8), for a gamma mean prior, the predictive posterior distribution $\mathrm{q}_{13}(x)$ represents the probability of a negative binomial random variable taking a value of 13 counts, with mean $0.185x^2 + 6.8x + 5$ and variance $0.006x^4 - 0.399x^3 + 7.987x^2 + 6.8x + 9.95$.

Three calibrative densities have been calculated applying two different proper uniform prior dose distributions, both using information given in [86]. An administered radioiodine activity that produces a blood dose less than 2 Gy is considered safe, so we could take a uniform dose prior distribution from 0 to 2, assuming that doctors use prudent doses. On the other hand, the calibration curve was calculated up to a dose of 4.5 Gy, so another uniform dose prior distribution could be from 0 to 4.5. An improper uniform prior dose distribution from 0 to $+\infty$ is also applied.

Figure 5.3 shows the plot of the three densities of the estimated dose for the data test. Their statistics are indicated in Table 5.4. These results agree with those displayed in [86], where the dose estimate for Patient 1 was 1.14 Gy.

Table 5.4: Statistics summary of the calibrative densities for two proper and one improper uniform dose priors.

| Prior dose distribution | Mode | Expected | SD | 95% CI |
|---|---|---|---|---|
| $\mathcal{U}(0, 2)$ | 1.140 | 1.141 | 0.481 | $(0.203, 1.945)$ |
| $\mathcal{U}(0, 4.5)$ | 1.140 | 1.561 | 0.858 | $(0.203, 3.615)$ |
| $\mathcal{U}(0, +\infty)$ | 1.140 | 1.593 | 0.921 | $(0.253, 3.829)$ |

## 5.4 The simplified compound Poisson calibration model

We now consider a data set that follows a compound Poisson distribution. The likelihood function of the test data has been previously described in (5.4), and the calculation of the calibrative density (5.5) requires to use numerical integration or Monte Carlo methods. However the model can be simplified by replacing $\delta$ in $\mathrm{L}(\tilde{y}|\mu, \delta)$ with the MLE $\hat{\delta}$ obtained from the calibration data. The performance of this simplification is analysed and compared in the example 5.4.1. Then the likelihood function $\mathrm{L}(\tilde{y}|\mu, \hat{\delta})$, which we prefer to denote as $\mathrm{L}(\tilde{y}, \hat{\delta}|\mu)$, is equivalent to the probability function of the sum of the observations, that is the probability function of a compound Poisson observation,

$$\mathrm{L}(\tilde{y}, \hat{\delta}|\mu) \propto \mathrm{p}(s, \hat{\delta}|m\mu),$$

where $s = \sum_{i=1}^{m} \tilde{y}_i$. Then, the calibrative density is as described in (5.6) with,

$$\mathrm{q}_s(x) = \int_{-\infty}^{\infty} \mathrm{p}(s, \hat{\delta}|m\mu)\phi(\mu|x)\mathrm{d}\mu, \tag{5.9}$$

if the mean prior is a normal density or,

$$\mathrm{q}_s(x) = \int_{0}^{\infty} \mathrm{p}(s, \hat{\delta}|m\mu)\Gamma(\mu|x)\mathrm{d}\mu, \tag{5.10}$$

when the mean prior is gamma distributed. Expressions (5.9) and (5.10) correspond to the probability function of mixed compound Poisson random variables, where the mixing density is respectively normal or gamma, evaluated at $s$. The operations of compounding and mixing are interchangeable for these models ([44], [35]), e.g., mixing a Neyman A with a normal results in the following,

$$\begin{aligned}
&\mathrm{NA}\left(m\mu, \hat{\delta}\right) \bigwedge_{\mu} \mathrm{N}\left(\mathrm{f}(x, \hat{\beta}), \mathrm{v}(x, \hat{\beta})\right) = \\
&\mathrm{Pois}\left(\frac{m\mu}{\hat{\delta} - 1}\right) \bigvee \mathrm{Pois}\left(\hat{\delta} - 1\right) \bigwedge_{\mu} \mathrm{N}\left(\mathrm{f}(x, \hat{\beta}), \mathrm{v}(x, \hat{\beta})\right) = \\
&\mathrm{Pois}\left(\frac{m\mu}{\hat{\delta} - 1}\right) \bigwedge_{\mu} \mathrm{N}\left(\mathrm{f}(x, \hat{\beta}), \mathrm{v}(x, \hat{\beta})\right) \bigvee \mathrm{Pois}\left(\hat{\delta} - 1\right) = \\
&\mathrm{Herm}\left(\frac{m\mathrm{f}(x, \hat{\beta})}{\hat{\delta} - 1}, 1 + \frac{m\mathrm{v}(x, \hat{\beta})}{(\hat{\delta} - 1)\mathrm{f}(x, \hat{\beta})}\right) \bigvee \mathrm{Pois}\left(\hat{\delta} - 1\right).
\end{aligned} \tag{5.11}$$

This is providing that (5.9) and (5.10) are respectively the probability functions of compound Hermite and compound negative binomial random variables.

Therefore, according to the different choices of the compound Poisson distribution we obtain the following compound distributions for $q_s(x)$:

$$\begin{aligned}
\text{NA}: &\quad \mathcal{F}\left(\frac{m\mathrm{f}(x,\hat{\beta})}{\hat{\delta}-1}, 1+\frac{m\mathrm{v}(x,\hat{\beta})}{(\hat{\delta}-1)\mathrm{f}(x,\hat{\beta})}\right)\bigvee \mathrm{Pois}\left(\hat{\delta}-1\right)\\
\text{NB}: &\quad \mathcal{F}\left(\frac{m\mathrm{f}(x,\hat{\beta})\log(\hat{\delta})}{\hat{\delta}-1}, 1+\frac{m\mathrm{v}(x,\hat{\beta})\log(\hat{\delta})}{(\hat{\delta}-1)\mathrm{f}(x,\hat{\beta})}\right)\bigvee \mathrm{Log}\left(\frac{\hat{\delta}-1}{\hat{\delta}}\right)\\
\text{Hermite}: &\quad \mathcal{F}\left(\frac{m\mathrm{f}(x,\hat{\beta})}{2(\hat{\delta}-1)}, 1+\frac{m\mathrm{v}(x,\hat{\beta})}{2(\hat{\delta}-1)\mathrm{f}(x,\hat{\beta})}\right)\bigvee \mathrm{Bin}\left(2,\hat{\delta}-1\right)
\end{aligned} \tag{5.12}$$

Here $\mathcal{F}(\mu_{\mathcal{F}},\delta_{\mathcal{F}})$ indicates a Hermite or a negative binomial distribution, according to (5.9) or (5.10), parameterized by its population mean and dispersion index. When $\mathcal{F}$ is the Hermite distribution, these representations make sense only when $\mathrm{f}(x,\hat{\beta})(\hat{\delta}-1)\geq m\mathrm{v}(x,\hat{\beta})$ for the NA, $\mathrm{f}(x,\hat{\beta})(\hat{\delta}-1)\geq m\mathrm{v}(x,\hat{\beta})\log(\hat{\delta})$ for the NB and $2\mathrm{f}(x,\hat{\beta})(\hat{\delta}-1)\geq m\mathrm{v}(x,\hat{\beta})$ for the Hermite.

Compound negative binomial distributions have been studied and applied in several publications. Properties, characterizations and references can be found in [44]. Compound Hermite distributions are less common, so far there is one recent publication [41] that studies the continuous compound Hermite gamma distribution.

When $\mathcal{F}(\mu_{\mathcal{F}},\delta_{\mathcal{F}})$ is negative binomial, the probabilities of the associated compound distributions can be calculated using the Panjer recursion formula [72]. This formula is based on the fact that the probabilities $p_n=\mathrm{P}(X=n)$ of a random variable $X$ distributed as a $\mathrm{NB}(\mu_{\mathcal{F}},\delta_{\mathcal{F}})$ satisfy a first order recurrence relation $p_n=p_{n-1}(a+b/n)$, where $a=(\delta_{\mathcal{F}}-1)/\delta_{\mathcal{F}}$ and $b=(\mu_{\mathcal{F}}-\delta_{\mathcal{F}}+1)/\delta_{\mathcal{F}}$. Then, if the probabilities of the generalizing distribution are denoted as $f_k$, the probabilities $q_i$ of the corresponding negative binomial compound distribution satisfy the recursion [72],

$$q_0=\frac{p_0}{(1-f_0a)^{1+b/a}},\ q_i=\sum_{j=1}^{i}\left(a+\frac{bj}{i}\right)f_jq_{i-j},\ i\geq 1. \tag{5.13}$$

Expression (5.13) can be efficiently used to calculate (5.10). The values of $a$ and $b$ will be taken according to the chosen distribution of the observations, using the corresponding expression of $\mu_{\mathcal{F}}$ and $\delta_{\mathcal{F}}$ of the negative binomial ($\mathcal{F}$) indicated in (5.12). In the next section we will give an example of application.

When $\mathcal{F}$ is Hermite, the probabilities of a Hermite compound distribution cannot be calculated using the Panjer recursion formula because the probabilities of the Hermite do not follow a linear recursion. To calculate the probabilities in this case we state and prove (in Appendix 5.6) the following proposition:

**Proposition 5.4.1** *Let $q_n$, $n=0,1,2\ldots$ be the probabilities of a compound Hermite distribution of the form Herm$(\mu_h,\delta_h)\bigvee\mathcal{P}$, where $\mathcal{P}$ is a count distribution with probabilities $f_k$, $k=0,1,2\ldots$. We define $r_j=\sum_{i=0}^{j}f_if_{j-i}$, $j=0,1,2\ldots$, then*

$$q_n=\frac{\mu_h}{n}\sum_{i=0}^{n-1}(n-i)q_i\Big\{(2-\delta_h)f_{n-i}+\frac{(\delta_h-1)}{2}r_{n-i}\Big\}, \tag{5.14}$$

*and $q_0=\exp(\mu_h((2-\delta_h)(f_0-1)+(\delta_h-1)(f_0^2-1)/2))$.*

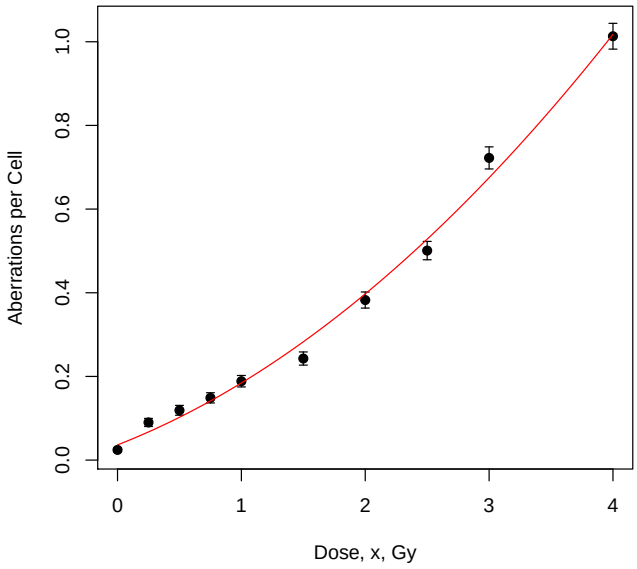


Figure 5.4: Observed means (dots), plus/minus twice their standard errors (error bars), and predicted means (solid line) of the number of micronuclei for NB fitting, based on the data in Table 5.5, omitting the 0.1 Gy test data.

It is important to remark that, to calculate $q_s(x)$ in (5.9) and (5.10), a computationally intensive direct numerical integration can be done instead to use the Panjer recursion or Proposition 5.4.1. To this end, it would be enough to obtain numerically the probabilities which are available for a more wide range of models than those studied in this chapter.

The use of (5.14) will be illustrated with a real data analysis in the next section.

### 5.4.1 Example: High LET exposure

Puig and Valero [77] studied the fitting of an experiment of eleven samples of peripheral blood exposed to different doses of $\gamma$–rays (Table 5.5), where the dose rate was 0.93 cGy/min. For each sample, approximately 5000 binucleated cells were inspected, and the numbers of micronuclei were counted.

The $u$ values shown in Table 5.5 confirm the overdispersion, thus Poisson regression is not adequate.

Similarly to the example analyzed in 5.3.1 the 0.1 Gy data will be removed to be used as test data. This distribution has a total of 250 micronuclei in a total of 5000 cells so $s = 250$ and $m = 5000$.

The appropriate dose–response curve, i.e. the regression model, is again a linear–quadratic model with intercept, $\mathrm{f}(x, \beta) = \beta_2 x^2 + \beta_1 x + \beta_0$ (Figure 5.4). Table 5.6 shows the BIC values for the four different models studied in this work. Note how these values support the use of the NB model.

Table 5.5: Frequency distributions of the number of micronuclei after exposure to 11 doses of gamma–rays, and the sample means, dispersion coefficients and $u$ values for each distribution. Test data in emphasis.

| | Number of micronuclei | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dose (Gy) | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | $\bar{y}$ | $d$ | $u$ |
| 0.00 | 4887 | 106 | 5 | 2 | | | | | 0.024 | 1.156 | 7.839 |
| *0.10* | *4773* | *206* | *19* | *2* | | | | | *0.050* | *1.150* | *7.526* |
| 0.25 | 4261 | 324 | 41 | 12 | 2 | | | | 0.090 | 1.306 | 15.306 |
| 0.50 | 4536 | 364 | 76 | 17 | 7 | | | | 0.119 | 1.449 | 22.484 |
| 0.75 | 4383 | 512 | 85 | 18 | 2 | | | | 0.149 | 1.257 | 12.876 |
| 1.00 | 4225 | 636 | 115 | 19 | 5 | | | | 0.189 | 1.240 | 12.009 |
| 1.50 | 4018 | 805 | 139 | 26 | 9 | 1 | 2 | | 0.243 | 1.270 | 13.495 |
| 2.00 | 3499 | 1194 | 238 | 45 | 13 | 10 | 1 | | 0.383 | 1.209 | 10.471 |
| 2.50 | 3171 | 1313 | 393 | 94 | 24 | 3 | 2 | | 0.501 | 1.201 | 10.077 |
| 3.00 | 2582 | 1575 | 598 | 190 | 44 | 9 | 2 | 6 | 0.722 | 1.206 | 10.307 |
| 4.00 | 1974 | 1674 | 869 | 342 | 102 | 26 | 13 | 2 | 1.013 | 1.172 | 8.628 |

Table 5.6: BIC values using a second degree polynomial dose–response curve for the different models.

| Model | Poisson | Hermite | NA | NB |
|---|---|---|---|---|
| BIC | 67360.01 | 66537.46 | 66467.85 | 66437.93 |

Using the NB model, the maximum likelihood estimation provides the following results:

- Fitted coefficients:

$$\hat{\beta}_0 = 3.639 \cdot 10^{-2}, \quad \hat{\beta}_1 = 1.156 \cdot 10^{-1}, \hat{\beta}_2 = 3.241 \cdot 10^{-2}, \quad \hat{\delta} = 1.231.$$

- Estimated covariance matrix:

$$\hat{\Sigma}_{\hat{\Theta}} = \begin{pmatrix} 73.749 & -115.908 & 29.210 & 13.976 \\ -115.908 & 373.338 & -110.398 & 36.919 \\ 29.210 & -110.398 & 38.102 & -3.625 \\ 13.976 & 36.919 & -3.625 & 1133.825 \end{pmatrix} \cdot 10^{-7}.$$

Then, the prior densities are:

- **Complete Model:** According to (5.3), $(\mu, \delta)|x$ follows a bivariate normal distribution with mean $\left(\hat{\beta}_2 x^2 + \hat{\beta}_1 x + \hat{\beta}_0, \hat{\delta}\right)$ and variance–covariance $\nabla \cdot \hat{\Sigma}_{\hat{\Theta}} \cdot \nabla^t$, where:

$$\nabla = \begin{pmatrix} \dfrac{\partial \mathrm{f}}{\partial \beta_0} & \dfrac{\partial \mathrm{f}}{\partial \beta_1} & \dfrac{\partial \mathrm{f}}{\partial \beta_2} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & x & x^2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

so the variance–covariance is:

$$\begin{pmatrix} \hat{\Sigma}_{33}x^4 + 2\hat{\Sigma}_{32}x^3 + 2\hat{\Sigma}_{31}x^2 + \hat{\Sigma}_{22}x^2 + 2\hat{\Sigma}_{21}x + \hat{\Sigma}_{11} & \hat{\Sigma}_{43}x^2 + \hat{\Sigma}_{42}x + \hat{\Sigma}_{41} \\ \hat{\Sigma}_{43}x^2 + \hat{\Sigma}_{42}x + \hat{\Sigma}_{41} & \hat{\Sigma}_{44} \end{pmatrix}.$$

Table 5.7: Statistics summary of the calibrative densities for the complete model, and the simplified models using a gamma and a normal mean prior with a uniform prior dose distribution.

| Model | Complete | S. Norm. p. | S. Gam. p. |
|---|---|---|---|
| Mode | 0.125 | 0.115 | 0.115 |
| Expected | 0.124 | 0.114 | 0.114 |
| SD | 0.033 | 0.034 | 0.034 |
| 95% CILB | 0.059 | 0.047 | 0.047 |
| 95% CIUB | 0.190 | 0.182 | 0.181 |

For this example the calibrative density (5.5) is calculated via numerical integration in order to be compared with those calculated using the simplified models.

- **Simplified Models:** According to the arguments given in Section 5.4, $\mu|x$ follows a gamma or a normal distribution with mean $\mathrm{f}(x,\hat{\beta}) = \hat{\beta}_2 x^2 + \hat{\beta}_1 x + \hat{\beta}_0$ and variance $\mathrm{v}(x,\hat{\beta}) = \nabla \cdot \hat{\Sigma}_\beta \cdot \nabla^t$, where:

$$\nabla = \left( \frac{\partial \mathrm{f}}{\partial \beta_0}, \frac{\partial \mathrm{f}}{\partial \beta_1}, \frac{\partial \mathrm{f}}{\partial \beta_2} \right) = \left(1, x, x^2\right),$$

so the variance is $\hat{\Sigma}_{33}x^4 + 2\hat{\Sigma}_{32}x^3 + 2\hat{\Sigma}_{31}x^2 + \hat{\Sigma}_{22}x^2 + 2\hat{\Sigma}_{21}x + \hat{\Sigma}_{11}$. According to (5.12), for a normal or a gamma mean prior, the predictive posterior distribution $\mathrm{q}_{250}(x)$ represents respectively the probability of a compound Hermite- or compound negative binomial-Logarithmic random variable taking a value of 250 counts, both with same $\mathrm{f}(x,\hat{\beta}) = 0.032x^2 + 0.116x + 0.036$, $\mathrm{v}(x,\hat{\beta}) = 3.81 \cdot 10^{-6}x^4 + 1.525 \cdot 10^{-5}x^3 + 5.842 \cdot 10^{-6}x^2 - 2.318 \cdot 10^{-5}x + 7.375 \cdot 10^{-6}$, and $\hat{\delta} = 1.231$.

To use the normal mean prior (5.9) in this calibration set for NB responses, there is a condition to be satisfied: $\mathrm{f}(x,\hat{\beta})(\hat{\delta} - 1) - m\mathrm{v}(x,\hat{\beta})\log(\hat{\delta}) \geq 0$. It is satisfied when $x \leq 4.294$ Gy. In this example this is not a problem and it could be used as prior information about the dose, that is $\mathrm{p}(x) \sim \mathcal{U}(0, 4.294)$. For the range of the likely doses studied, the minimum value of the shape parameter of the mean prior gamma is 179.605, and consequently both gamma and normal mean priors are almost indistinguishable (red and blue curves in Figure 5.5).

Figure 5.5 shows the plot of the three densities (one from the complete model and two from the simplified ones) of the estimated dose for the data test. Note that both calibrative densities from the simplified models are practically the same. The statistics of these densities are shown in Table 5.7. These results incorporate the real dose (0.1 Gy) and also show their similarities, chiefly between the simplified models.

## 5.5 Conclusion

In this chapter we have presented several Bayesian–type methods for count data inverse regression, showing its application in the field of cytogenetic dosimetry. First, in Section 5.2 we defined our methodology for inverse regression, where

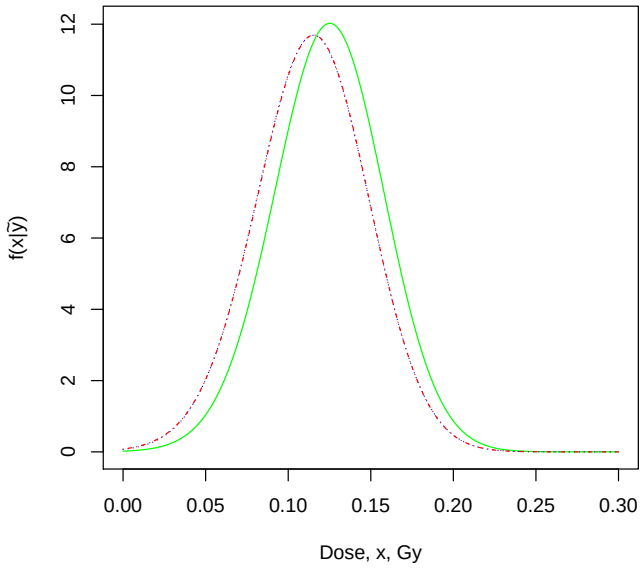


Figure 5.5: Calibrative densities of the 0.1 Gy test data using the complete model (5.5) (green/solid line), and the simplified ones with a normal (blue/dotted line) and a gamma (red/dash-dot line) mean prior density; all with a uniform prior dose distribution. Blue and red curves are indistinguishable.

responses are either Poisson or two–parameter compound Poisson. We have assumed that the dispersion index is constant along the different doses. This methodology leads to a bivariate normal prior density when the responses follow a two–parameter compound Poisson distribution, and an univariate normal or gamma mean prior density when the responses follow a Poisson distribution. To use our methodology, only the estimates of the parameters and covariance matrix of the dose–response curve are required. This information is available from the standard frequentist analysis suggested by the IAEA manual, with many examples published by other researchers or laboratories. Therefore, our method is not a full Bayesian approach because the dose–response curve is estimated using frequentist analysis. MCMC methods could be used if the models were more complex or the prior densities more complicated. They might also be used for model averaging, since one might aim to avoid choosing one of the presented four models, preferring to use a weighted amalgam of them.

The Poisson model is developed in Section 5.3, leading to a closed form of the calibrative density. Two examples of dose estimation based on the dicentric assay are reported.

In Section 5.4 we treated two–parameter compound Poisson models, simplifying them to get the calibrative densities into a closed form. For this purpose, we have presented a method which involves calculating the probabilities of compound negative binomial distributions, using Panjer's recursion [72], and com-

pound Hermite distributions, using a recursion relation described in Proposition 5.4.1. Another example of dose estimation is shown, based on data obtained with the micronucleus assay. We have assumed a constant dispersion coefficient, but our methods could be also extended to dose dependent dispersion models of the form $\delta_{ij} = \mathrm{g}(x_i, \gamma)$, $\gamma \in \mathbb{R}^q$.

The illustrative examples show applications using the most frequent calibrative curves, that are second order polynomials (the linear–quadratic model). However, other response functions can be directly analyzed using the same methodology. It should be noted that the approaches presented here may also prove useful in areas other than biological dosimetry.

## 5.6 Appendix A. Proof of Proposition 5.4.1

First of all, let us recall some topics related to the probability-generating function (pgf). Given a count random variable $X$, its pgf $\Phi_X(s)$ is defined as

$$\Phi_X(s) = \sum_{k=0}^{\infty} p_k s^k$$

where the coefficients of this power series are the probabilities $p_k = \mathrm{P}(X = k)$ and consequently the derivatives at $s = 0$ divided by $k!$ provide the probability mass function of $X$. The pgf of a compound probability distribution described in (5.1) is

$$\Phi_X(s) = \Phi_N(\Phi_\xi(s)), \tag{5.15}$$

where $\Phi_N(s)$ is the pgf of $N$ and $\Phi_\xi(s)$ is the common pgf of the $\xi_i$ [44].

One property of pgf's is that the sum of independent random variables is a random variable whose pgf is the product of the pgf's of the summed variables; e.g., given $X$ and $Y$ independent random variables with pgf's $\Phi_X(s)$ and $\Phi_Y(s)$ respectively, the pgf of $X + Y$ results

$$\Phi_{X+Y}(s) = \Phi_X(s)\Phi_Y(s). \tag{5.16}$$

According to [48] the pgf of a random variable $X$ Hermite distributed with mean $\mu_h$ and dispersion coefficient $\delta_h$ is

$$\Phi_X(s) = \mathrm{e}^{\mu_h\{(2-\delta_h)(s-1)+(\delta_h-1)(s^2-1)/2\}}, \tag{5.17}$$

therefore, according to (5.15), the pgf of a $\mathrm{Herm}(\mu_h, \delta_h) \bigvee \mathcal{P}$ distribution, being $\psi(s)$ the pgf of $\mathcal{P}$, is

$$\phi(s) = \mathrm{e}^{\mu_h\{(2-\delta_h)(\psi(s)-1)+(\delta_h-1)(\psi^2(s)-1)/2\}}, \tag{5.18}$$

thus the probability in 0 is

$$q_0 = \phi(0) = \mathrm{e}^{\mu_h\{(2-\delta_h)(f_0-1)+(\delta_h-1)(f_0^2-1)/2\}}.$$

Note that $\psi^2(s)$ is the pgf of a sum of two independent identically distributed random variables having both a pgf equal to $\psi$, so

$$\varphi(s) = \psi^2(s) = \sum_{n=0}^{\infty} r_n s^n, \quad r_n = \sum_{i=0}^{n} f_i f_{n-i}.$$

The derivative of $\phi$ is

$$\phi'(s) = [\mu_h\{(2-\delta_h)(\psi'(s)-1) + (\delta_h - 1)(\varphi'(s)-1)/2\}]\phi(s),$$

therefore,

$$\begin{aligned}\sum_{n=1}^{\infty} nq_n s^{n-1} &= \mu_h\Big\{(2-\delta_h)\sum_{n=1}^{\infty} nf_n s^{n-1} + \frac{(\delta_h-1)}{2}\sum_{n=1}^{\infty} nr_n s^{n-1}\Big\}\sum_{n=0}^{\infty} q_n s^n \\ &= \mu_h \sum_{n=1}^{\infty} n\Big\{(2-\delta_h)f_n + \frac{(\delta_h-1)}{2}r_n\Big\}s^{n-1}\sum_{n=0}^{\infty} q_n s^n,\end{aligned}$$

matching the coefficients with same degree in $s$ in both sides leads to,

$$q_n = \frac{\mu_h}{n}\sum_{i=0}^{n-1}(n-i)q_i\Big\{(2-\delta_h)f_{n-i} + \frac{(\delta_h-1)}{2}r_{n-i}\Big\}, \quad n \geq 1,$$

and this finishes the proof. □

# Chapter 6

# `radir` package: An `R` implementation for cytogenetic biodosimetry dose estimation

This chapter corresponds to the contents of [65].

**Abstract:** The Bayesian framework has now been shown to be very useful in cytogenetic dose estimation. This approach allows description of the probability of an event in terms of previous knowledge, e.g. its expectation and/or its uncertainty. A new `R` package entitled `radir` (`rad`iation `i`nverse `r`egression) has been implemented with the aim of reproducing a recent Bayesian–type dose estimation methodology. `radir` takes the method of dose–estimation under the Poisson assumption of the responses (the chromosomal aberrations counts) for the required dose–response curve (typically linear or quadratic). The individual commands are described in detail and relevant examples of the use of the methods and the corresponding `radir` software tools are given. The suitability of this methodology is highlighted and its application encouraged by providing a user–friendly command type software interface inside the `R` statistical software (version 3.1.1 or higher), which provides a complete manual.

**Keywords:** Bayesian, biological dosimetry, `R` software, calibrative density, Poisson distribution.

## 6.1 Introduction

The classical methods for dose estimation in radiation cytogenetics are well established and described in detail in the manual of the International Atomic Energy Agency (IAEA) [42]. First, calibration data (generally yields of chromosome aberrations in blood lymphocytes) are collected and fitted to a linear or quadratic model, the coefficients of which are then used to calculate doses. The Poisson model is used to describe the uncertainty on the yield of aberrations, and this is combined with uncertainty on the fitted calibration coefficient(s) using standard methodology in order to give the total uncertainty associated with

the estimated dose.

In the classical or 'frequentist' framework the coefficients of the calibrated dose–response curve are considered 'fixed' thus providing an estimate of radiation dose and associated confidence limits using standard likelihood methods. Therefore, assignment of a probability to an event is based solely on the observed frequency of occurrence of the event.

Alternatively, the Bayesian approach considers the parameters, for instance the dose–response curve coefficients in this case, random variables for which previous information could exist. This information could come from previous analysis in the field literature or even from the experts' opinion/knowledge. The newly collected data is combined with this *prior* information to produce a *posterior* model.

The Bayesian inference uses distributions for all the parameters, leading to an important advantage: the uncertainty of the system is an intrinsic part of the analysis. A review of these methods can be found in [4]. Bayesian methods in cytogenetic biodosimetry give the estimation of the absorbed dose by an individual in a form of random variable distribution, called the calibrative dose density [15].

Some specific software has been developed to fit dose–response curves and to estimate the dose absorbed by an individual; e.g. CABAS [18], DoseEstimate [1] and BioDoser [98]. There is also one recent program, CytoBayesJ [3], which provides some basic software tools for Bayesian analysis of cytogenetic radiation dosimetry data.

In this chapter we present a new `R` statistical software package which implements the Poisson models developed in [39]. These models are Bayesian–type inverse regression. They use dose–effect calibration curves estimated by the frequentist approach.

This methodology collects the prior information of the yield of chromosomes per cell (the prior population mean) from the dose–effect calibration curve in an univariate parameter by means of the delta method. This prior distribution is assumed normal (under some constraints) or gamma distributed and the calibrative dose density results in terms of the probability functions of the Hermite or negative binomial distributions.

## 6.2 The `radir` `R` software package

The software introduced in this work has been written in the `R` programming language [81], which is becoming more popular in the cytogenetic biodosimetry context in recent years because of its availability; it can be freely downloaded from `http://cran.r-project.org/` and used on the most common operating systems. In fact, an `R` script for fitting dose–response curves written by H. Braselmann was included in [42]. In addition, the `hermite` `R` package version 1.0.1 [66] has been utilised for the management of the Hermite distribution. The general workflow of `radir` package is summarised in Section 6.2.2. A video tutorial has been prepared with the aim of helping `radir` users in the installation and general usage of the `R` statistical software and, in particular, the `radir` package. It can be found in the next open access link, `http://polimedia.uab.cat/#v_592`.

### 6.2.1 Features of `radir` package

In version 1.0, available from `http://cran.r-project.org/src/contrib/radir_1.0.tar.gz`, the following tools are included:

- Calculation of the calibrative dose density for a given:
  - expression of the dose–response curve;
  - hyperparameters set;
  - estimate of the parameter set;
  - variance–covariance matrix of the estimation;
  - total number of cells examined;
  - number of chromosomal aberrations;
  - prior distribution of the chromosomal aberration mean: normal or gamma;
  - prior distribution of the absorbed dose: uniform or gamma;
  - parameters of the distribution of the dose prior.
- Summary statistics of the calibrative dose density: best estimate, expected value, standard deviation and the 95% highest posterior density (HPD) interval, defined as the shortest range that contains the 95% (or the required percentage) region of the probability density.
- Calculation of the HPD interval for a given credible region.
- Calculation of the probability between two given doses.
- Plots of the
  - calibrative dose density;
  - HPD interval for a given credible region;
  - probability between two given doses;
  - cumulative dose distribution function.

### 6.2.2 `radir` package workflow

The calibrative density is computed explicitly for the Poisson model in [39], which is the most common situation and is also the case covered by the `radir` package. The software takes as inputs laboratory information such as the dose–response curve, the maximum likelihood estimates of its parameters, the variance–covariance matrix and the mean prior distribution, which can be normal or gamma, together with patient information such as the number of cells examined, chromosomal aberrations counts, and the prior dose distribution, which can be uniform or gamma. From this input data, the calibration dose density is calculated and summary statistics, probability between two given doses, HPDs and several plots can be obtained. This workflow is summarised in Figure 6.1.

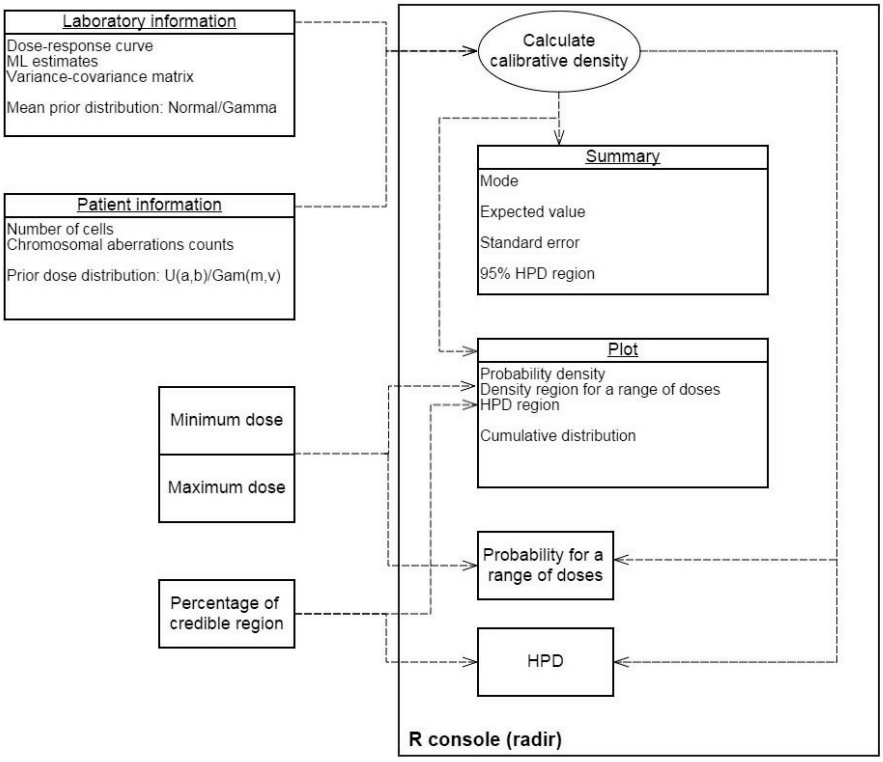


Figure 6.1: `radir` package workflow

### 6.2.3 Calibrative dose density calculation

The calculation of the calibrative density is based in the fact that the likelihood function of the data coming from the patient is proportional to a Poisson probability function evaluated at the total number of chromosomal aberrations. Then, the calibrative density remains proportional to the prior dose density multiplied by a probability function of a mixed–Poisson distribution evaluated at the total number of chromosomal aberrations. The nature of the mixed–Poisson distribution depends on the mean prior distribution that acts as the "mixing" distribution. When the mean prior is gamma (the default case for the `radir` package) it leads to a negative binomial, and for the normal density option it leads to the Hermite distribution. The details of all the process and methodology are widely described in [39].

The first step is to compute the calibrative dose density, by means of the function `dose.distr()`.

A call to this function might be

```
dose.distr(f, pars, beta, cov, cells, dics, m.prior, d.prior,
 prior.param, stdf)
```

The description of these arguments can be summarised as follows:

- `f`: Dose–response function, as an expression. Must be differentiable in the

domain of parameters.

- `pars`: String vector containing the parameters in `f`.
- `beta`: Estimates of the parameters.
- `cov`: Variance–covariance matrix of estimates `beta`.
- `cells`: Patient information, number of cells examined.
- `dics`: Patient information, observed number of aberrations.
- `m.prior`: String containing the prior distribution of the mean. In the current version of the package, it can be gamma (the default value) or normal.
- `d.prior`: String containing the prior distribution of the dose. In the current version of the package, it can be gamma or uniform (the default value).
- `prior.param`: Vector of length 2 containing the parameters of the distribution of the dose prior. Its default value is the non–informative prior. If `d.prior` is a gamma distribution, the mean and standard deviation should be given, otherwise the function will return an error.
- `stdf`: Approximated standard deviation factor. This input is useful to control the ends of the calibrative density; i.e. in case the tails of the calibrative dose density are very long this value could be reduced, or *vice versa*. Its default value is 6.

The gamma and normal distributions are the alternatives for the mean prior distribution. In principle, the normal distribution would be the most natural choice because, according to the maximum likelihood theory, the mean is asymptotically normal distributed with expectation and variance depending on the dose and the dose–response function. Then, the calibrative density is proportional to the prior dose density multiplied by a probability function of a Hermite distribution, that is, a Poisson–mixed normal distribution. However, to mix a Poisson with a normal distribution only makes sense when the population mean of the normal distribution is greater than its variance. For this reason, when using the normal mean prior option, the range of doses of the calibrative density could be truncated to the right, and potentially not being enough sensible for the analysed sample. Conversely, using the gamma mean prior distribution (the default), mixing a Poisson with a gamma distribution (negative binomial distribution) does not create such problems and the range of doses is not truncated. Moreover, a gamma distribution with a large shape parameter is a good approximation to the normal distribution. Therefore, it is recommendable in general to use the gamma mean prior option.

The gamma and uniform distributions are the alternatives for dose prior distributions. The gamma distribution has been used for instance in [33], and is parameterized here in terms of its mean and standard deviation. The uniform distribution is parameterised by its minimum and maximum, and to the knowledge of the authors it has not been previously used for cytogenetic dosimetry, even though is a sensible choice in general dosimetry. The dose prior choice

depends on expert opinion and/or the information collected from the irradiation event. If there is no prior information, the most appropriate option is to choose an improper uniform prior defined between zero and infinity. This non–informative option is the default. If limited knowledge about the dose is available, for instance its maximum range, then we could use as a prior a proper uniform distribution defined over zero and the maximum range. More information about the dose (for instance mean and standard deviation), like in the example 3.1, can lead to the usage of a gamma prior.

The function output collects the sequence of doses and their respective probability density.

### 6.2.4 Statistics summary, credible region, and probability between doses

A summary containing the most relevant information about the estimated doses can be obtained via `summary()`.

This function, when applied to the output of `dose.distr`, gives the most interesting statistics in this context including mode, expected value, standard deviation and the 95% HPD credibility interval.

The HPD credible interval for an object of class `dose.radir` can be obtained numerically by means of the function `ci.dose.radir`, with parameters

- `object`: An object of class `dose.radir` containing the estimated doses.
- `cr`: Credible region size. Its default value is 0.95.

The probability between two doses can be obtained numerically by means of the function `pr.dose.radir`, with parameters

- `object`: An object of class `dose.radir` containing the estimated doses.
- `lod`: Lower dose value. Its default value is 0.
- `upd`: Upper dose value. Its default value is the maximum dose in `object`.

### 6.2.5 Plots

Graphics can be obtained in the standard way by means of `R plot()` or `lines()` functions.

The `plot` function can also be used to present credible intervals through the argument `ci=TRUE`. The desired credible region size can be fixed using the argument `cr`, that is 0.95 by default. The color of the shaded credible region is grey by default, but it may be changed by using the argument `col.ci`. For instance, to see the credible region shaded in red, the user should write `col.ci=''red''`.

The probability between two doses can be graphically represented by means of the argument `prob=c(d1,d2)`, where `d1` and `d2` are respectively the lower and upper doses considered. The color of the shaded region is grey by default, but it may be changed by using the argument `col.pr` in the same way as for the parameter `col.ci`.

The distribution function can be plotted as well, using the argument `distr= TRUE` in the `plot` function.

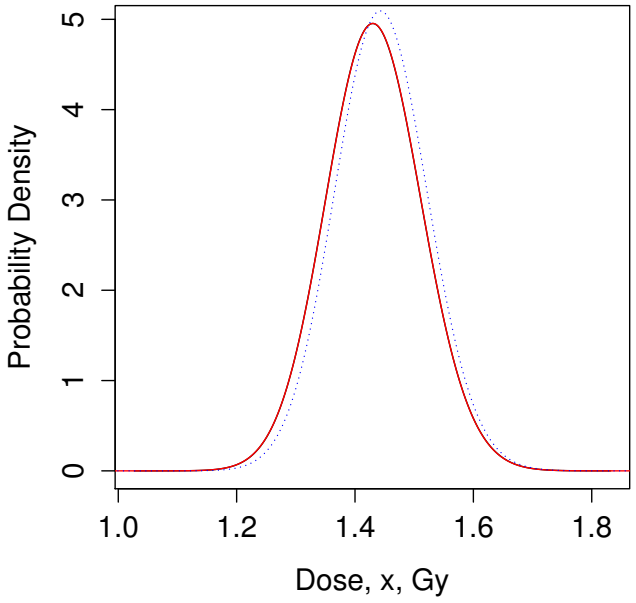


Figure 6.2: Calibrative densities of the 1.5 Gy test data for a normal mean prior and a $\mathcal{U}(0,\infty)$ dose prior (black, *ex1.a*), for gamma mean priors and a $\mathcal{U}(0,\infty)$ dose prior (red, *ex1.b*) and a gamma dose prior (blue dotted line, *ex1.c*). Note that the black and red lines are indistinguishable.

## 6.3 Examples

Several examples of use of the `radir` package (some of them introduced in [39]) are described in detail in this section.

### 6.3.1 Cobalt–60 gamma-ray irradiation

In [39] the authors consider an example from an *in vitro* Cobalt–60 gamma–ray exposure. From the calibration data (Table 1 in [39]) the 1.5 Gy row is removed to be inferred later. The model consists of a linear–quadratic dose response without a constant term curve, $\beta_2 x^2 + \beta_1 x$, where $x$ represents the absorbed dose, assuming that the counts of the chromosomal aberrations are Poisson distributed. The specific data used in this example (and in the others) were obtained from real experiments, and are reasonable for a Cobalt–60 gamma–ray exposure.

Its maximum likelihood estimation provides

$$\hat{\beta}_1 = 3.126 \cdot 10^{-3}, \quad \hat{\beta}_2 = 2.537 \cdot 10^{-2}, \quad \hat{\Sigma}_{\hat{\beta}} = \begin{pmatrix} 7.205 & -3.438 \\ -3.438 & 2.718 \end{pmatrix} \cdot 10^{-6}.$$

The 1.5 Gy sample consisted of 102 observed dicentrics in a total of 1811

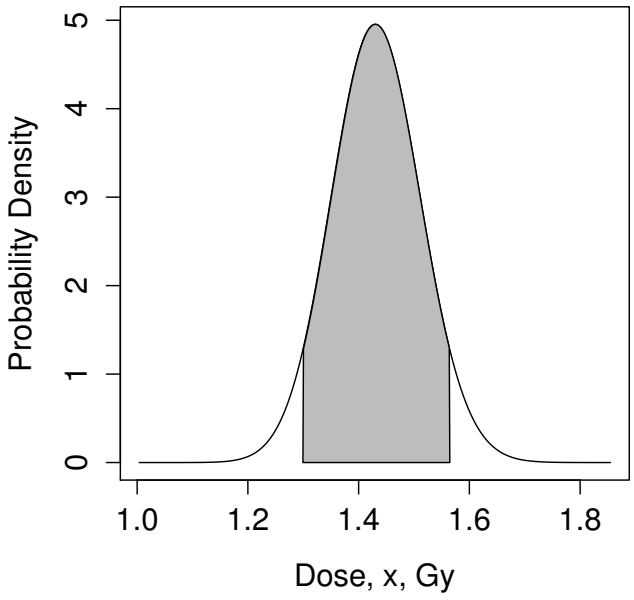


Figure 6.3: 90% HPD interval of the calibrative density of the 1.5 Gy test data for a normal mean prior and a $\mathcal{U}(0, \infty)$ dose prior.

blood cells. Therefore, the calibrative dose density for this observed data (assuming an improper uniform dose prior) can be calculated with the `radir` package by means of,

```
library(radir)
f <- expression(b1*x+b2*x^2)
pars <- c("b1","b2")
beta <- c(3.126e-3, 2.537e-2)
cov  <- matrix(c(7.205e-06,-3.438e-06,-3.438e-06,2.718e-06),nrow=2)
ex1.a <- dose.distr(f, pars, beta, cov, cells=1811, dics=102, m.prior="normal")
```

The default situation in `dose.distr()` is a gamma mean prior, as

```
ex1.b <- dose.distr(f, pars, beta, cov, cells=1811, dics=102)
```

In [39] the authors consider that assuming that the real dose is unknown a reasonable prior dose distribution is gamma with mean 1.75 and standard error 0.375. This can be implemented in the `radir` package by means of

```
ex1.c <- dose.distr(f, pars, beta, cov, cells=1811, dics=102, d.prior="gamma",
 prior.param=c(1.75,0.375))
```

Figure 6.2 shows the plot of the three densities of the estimated dose for the test data. It has been obtained using the `plot()` function on the outputs of function `dose.distr` (`ex1.a`, `ex1.b` and `ex1.c`), generated from the following code:

```
plot(ex1.a)
lines(ex1.b, col="red")
lines(ex1.c, col="blue", lty=3)
```

It should be observed that these results incorporate the real dose (1.5 Gy) and show the equivalence of both mean priors. Note that the gamma mean prior is moderately more conservative. A summary table of the statistics of the three calibrative densities calculated in this example can be obtained via the `summary()` function. For instance, for the first case

```
summary(ex1.a)
```

```
    Mode
    ----------------------------
     1.43


    Expected value
    ----------------------------
     1.432


    Standard Dev.
    ----------------------------
     0.081


    95% CI
    ----------------------------
     (1.275; 1.591)
```

A figure showing the density and the 90% HPD interval (Figure 6.3) can be obtained by means of

```
plot(ex1.a, ci=T, cr=0.90)
```

### 6.3.2 Analysis of doses in thyroid cancer patients

Serna et al. ([86]) studied chromosomal damage in lymphocytes of thyroid cancer patients after radioiodine treatment. The authors carried out the micronucleus assay in binucleated cells of blood samples from 25 patients 3 days after Iodine–131 (3.7 GBq) exposure. The *in vitro* dose–response curve was fitted by a linear–quadratic model, $f(x, \beta) = G\beta_2 x^2 + \beta_1 x + \beta_0$ according to Poisson's law, and the estimate of $\beta_0$ was not taken into account, because the authors in [86] argued that the intercept could change for each patient. Constant $G$ is the Lea–Catcheside generalised dose–protraction factor, which modifies the quadratic term according to the temporal pattern of exposure, being $G = 1$ for the *in vitro* assay. The authors calculated the following parameter estimates $(\hat{\beta}_i \pm SE(\hat{\beta}_i))$,

$$\hat{\beta}_1 = (13.6 \pm 5.5) \cdot 10^{-3}, \quad \hat{\beta}_2 = (3.7 \pm 1.6) \cdot 10^{-2}, \quad \rho = -0.89,$$

where $\rho$ is the correlation coefficient. Taking into account the characteristics of the Iodine–131 treatment, the authors in [86] found the factor $G$ to be close to 0.1. Then $\beta_0$, the background of each patient, was estimated by counting the micronuclei of the patient from a blood sample taken before the treatment, information provided in [86]. This leads to the fitted regression model $f(x, \hat{\beta}) =$

$G\hat{\beta}_2 x^2 + \hat{\beta}_1 x + \hat{\beta}_0$ with a covariance matrix that incorporates the variance of $\hat{\beta}_0$ without correlation with $\hat{\beta}_1$ and $\hat{\beta}_2$. For instance, Patient 1 presented 487 normal cells and 13 cells with just one micronucleus each for a total of 500 cells scored. Before the treatment 5 micronuclei were found in 500 blood cells, thus $\hat{\beta}_0 = (10 \pm 4.472) \cdot 10^{-3}$.

A gamma mean prior is preferred instead of a normal in this case, because the range of doses supported by the normal mean prior is very small, due to mathematical constraints. In [39] the authors show that in that case, the predictive posterior distribution represents the probability of a negative binomial random variable taking a value of 13 counts, with mean $4.810 \cdot 10^{-3}x^2 + 0.177x + 0.130$ and variance $4.326 \cdot 10^{-6}x^4 - 2.647 \cdot 10^{-4}x^3 + 1.008 \cdot 10^{-2}x^2 + 0.177x + 0.133$, for Patient 1. It is possible to define all needed input values for the `radir` package to analyse the patient data via

```
f <- expression(b0+b1*x+0.1*b2*x^2)
pars <- c("b0","b1","b2")
beta <- c(0.01, .0136, .0037)
cov <- matrix(c(1.98e-05,0,0,0,.3121*10^(-4),-.0798*10^(-4),0,-.0798*10^(-4),
 .0256*10^(-4)),nrow=3)
```

Three calibrative densities have been calculated applying two different proper uniform prior dose distributions, both using information given in [86]. An administered radioiodine activity that produces a blood dose less than 2 Gy is considered safe in the context of medical uses of radiation ([46]), so one could take a uniform dose prior distribution from 0 to 2 (`ex2.u1`). On the other hand, the calibration curve was calculated up to a dose of 4.5 Gy, so another uniform dose prior distribution could be from 0 to 4.5 (`ex2.u2`). An improper uniform prior dose distribution from 0 to $\infty$ is also applied (`ex2.u3`). This can be done with the `radir` package by means of

```
ex2.u1 <- dose.distr(f, pars, beta, cov, cells=500, dics=13, prior.param=c(0,
 2))
ex2.u2 <- dose.distr(f, pars, beta, cov, cells=500, dics=13, prior.param=c(0,
 4.5))
ex2.u3 <- dose.distr(f, pars, beta, cov, cells=500, dics=13)
```

Table 6.1 shows the summary results for the 25 patients described in [86]; these results were obtained using a loop that runs the function `dose.distr` for each patient taking the pre–radiotherapy and the post–radiotherapy information, for each of the two uniform prior dose distributions indicated ($U(0, 2)$ and an improper uniform distribution).

### 6.3.3 New model for low and high doses

In [79] the authors present a new model for biological dosimetry under a weighted Poisson assumption, where the mean of the underlying Poisson is a Gompertz function of the dose, and the underdispersion level is a linear function of the dose. This leads to a model where the mean of dicentrics is

$$f(x, \beta) = \beta_0 e^{-\beta_1 e^{-\beta_2 x}} \left(1 + \frac{\beta_3 x(2\beta_0 e^{-\beta_1 e^{-\beta_2 x}} + 1)}{1 + \beta_3 x(\beta_0^2 (e^{-\beta_1 e^{-\beta_2 x}})^2 + \beta_0 e^{-\beta_1 e^{-\beta_2 x}})}\right). \quad (6.1)$$

This model is especially useful for high dose exposures.

| Prior dose | $\mathcal{U}(0,2)$ | | | | $\mathcal{U}(0,\infty)$ | | |
|---|---|---|---|---|---|---|---|
| Patient | Expected | SD | 95% HPD | Mode | Expected | SD | 95% HPD |
| 1 | 1.141 | 0.481 | (0.319, 2.000) | 1.140 | 1.593 | 0.919 | (0.027, 3.362) |
| 2 | 1.155 | 0.471 | (0.355, 2.000) | 1.141 | 1.588 | 0.894 | (0.079, 3.365) |
| 3 | 0.759 | 0.477 | (0.000, 1.679) | 0.475 | 0.867 | 0.637 | (0.000, 2.121) |
| 4 | 1.237 | 0.444 | (0.470, 2.000) | 1.279 | 1.739 | 0.908 | (0.231, 3.599) |
| 5 | 1.099 | 0.470 | (0.326, 2.000) | 1.007 | 1.434 | 0.828 | (0.057, 3.068) |
| 6 | 0.847 | 0.487 | (0.000, 1.748) | 0.605 | 1.001 | 0.693 | (0.000, 2.358) |
| 7 | 1.172 | 0.459 | (0.398, 2.000) | 1.143 | 1.587 | 0.870 | (0.144, 3.343) |
| 8 | 1.011 | 0.488 | (0.206, 1.977) | 0.871 | 1.290 | 0.806 | (0.000, 2.841) |
| 9 | 0.842 | 0.478 | (0.000, 1.735) | 0.602 | 0.982 | 0.671 | (0.000, 2.288) |
| 10 | 1.116 | 0.457 | (0.374, 2.000) | 1.008 | 1.431 | 0.797 | (0.133, 3.041) |
| 11 | 0.791 | 0.499 | (0.000, 1.729) | 0.482 | 0.949 | 0.715 | (0.000, 2.356) |
| 12 | 0.452 | 0.389 | (0.000, 1.266) | 0.000 | 0.471 | 0.430 | (0.000, 1.344) |
| 13 | 0.851 | 0.495 | (0.000, 1.760) | 0.608 | 1.025 | 0.722 | (0.000, 2.433) |
| 14 | 1.020 | 0.479 | (0.236, 1.976) | 0.871 | 1.280 | 0.777 | (0.000, 2.773) |
| 15 | 0.580 | 0.435 | (0.000, 1.470) | 0.222 | 0.624 | 0.515 | (0.000, 1.655) |
| 16 | 0.542 | 0.437 | (0.000, 1.450) | 0.000 | 0.589 | 0.524 | (0.000, 1.653) |
| 17 | 0.746 | 0.467 | (0.000, 1.655) | 0.471 | 0.840 | 0.610 | (0.000, 2.038) |
| 18 | 0.607 | 0.449 | (0.000, 1.521) | 0.225 | 0.663 | 0.544 | (0.000, 1.757) |
| 19 | 0.940 | 0.470 | (0.138, 1.880) | 0.734 | 1.117 | 0.700 | (0.000, 2.465) |
| 20 | 0.771 | 0.486 | (0.000, 1.699) | 0.478 | 0.899 | 0.673 | (0.000, 2.215) |
| 21 | 0.771 | 0.486 | (0.000, 1.699) | 0.478 | 0.895 | 0.663 | (0.000, 2.202) |
| 22 | 1.141 | 0.481 | (0.319, 2.000) | 1.140 | 1.590 | 0.913 | (0.027, 3.362) |
| 23 | 1.075 | 0.490 | (0.261, 2.000) | 1.005 | 1.445 | 0.874 | (0.000, 3.119) |
| 24 | 0.934 | 0.482 | (0.100, 1.872) | 0.736 | 1.128 | 0.724 | (0.000, 2.529) |
| 25 | 0.931 | 0.491 | (0.070, 1.861) | 0.738 | 1.145 | 0.756 | (0.000, 2.609) |

Table 6.1: Statistics summary of the calibrative densities of the 25 Patients in [86] for $\mathcal{U}(0,2)$ and $\mathcal{U}(0,\infty)$ dose priors. Note that the mode is the same for both priors.

| Dose (Gy) | Dicentrics | Cells | Mode | Expected | SD | 95% HPD |
|---|---|---|---|---|---|---|
| 2 | 155 | 498 | 1.748 | 1.757 | 0.110 | (1.544, 1.975) |
| 6 | 425 | 150 | 5.850 | 5.859 | 0.228 | (5.415, 6.308) |
| 12 | 869 | 150 | 9.747 | 9.763 | 0.311 | (9.158, 10.378) |
| 17 | 914 | 100 | 16.143 | 16.814 | 1.778 | (14.148, 19.953) |

Table 6.2: Cells analyzed and total dicentrics counts for the simulated whole body irradiations for testing, and the statistics summary of their respective calibrative densities.

The *in vitro* irradiation experiment was performed using 10 different doses, from 0 to 25 Gy and numbers of dicentrics in blood lymphocytes were then counted. The models in `radir` are only for the Poisson assumption, so a Poisson model is defined with the dose–response curve defined by expression (6.1). The maximum likelihood estimation is,

$$\hat{\beta}_0 = 8.676, \quad \hat{\beta}_1 = 7.262, \quad \hat{\beta}_2 = 0.230 \quad \hat{\beta}_3 = 2.388,$$
$$\hat{\Sigma}_{\hat{\beta}} = \begin{pmatrix} 0.056 & 0.006 & -0.001 & 0.019 \\ 0.006 & 0.089 & 0.001 & 0.305 \\ -0.001 & 0.001 & 0.000 & 0.003 \\ 0.019 & 0.305 & 0.003 & 1.146 \end{pmatrix}.$$

To check the methodology and the `radir` performances, doses are inferred from test data shown in [79] (Table 6.2).

Therefore, the input parameters for the `radir` package should be

```
f <- expression(b0*exp(-b1*exp(-b2*x))*(1+b3*x*(2*b0*exp(-b1*exp(-b2*x))+1)/
 (1+b3*x*(b0^2*(exp(-b1*exp(-b2*x)))^2+b0*exp(-b1*exp(-b2*x))))))
pars <- c("b0","b1","b2","b3")
```

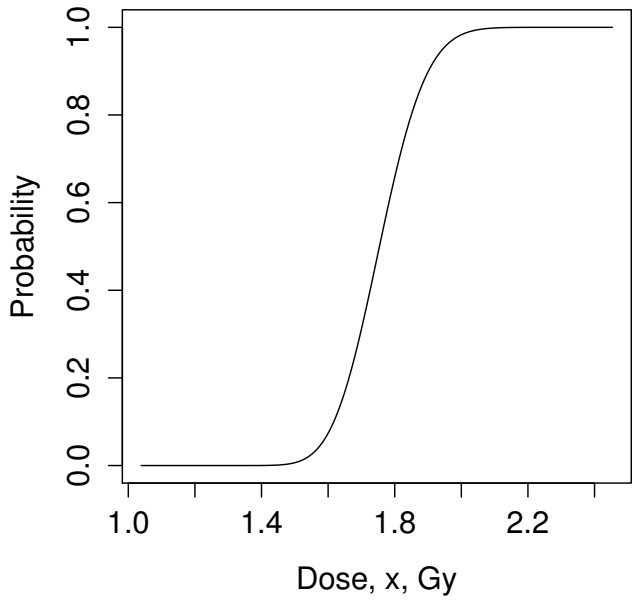


Figure 6.4: Cumulative distribution function of the 2 Gy test data.

```
beta <- c(8.6759674, 7.2624173, 0.2296528, 2.3875238)
cov <- matrix(c(0.0562628690,0.0056047214,-8.120599e-04,0.018587644,0.0056047214,
 0.0894182387,9.727568e-04,0.304724328,-0.0008120599,0.0009727568,3.792577e-05,
 0.002753902,0.0185876441,0.3047243281,2.753902e-03,1.145724697),nrow=4)
```

And then, the four situations proposed in Table 6.2 can be introduced in R by means of

```
ex3.a <- dose.distr(f, pars, beta, cov, cells=498, dics=155)
ex3.b <- dose.distr(f, pars, beta, cov, cells=150, dics=425)
ex3.c <- dose.distr(f, pars, beta, cov, cells=150, dics=869)
ex3.d <- dose.distr(f, pars, beta, cov, cells=100, dics=914)
```

Again, the `summary()` function can be used to check that the results are similar to the expected, for instance for the first experiment we have

```
summary(ex3.a)

    Mode
    ----------------------------
     1.748


    Expected value
    ----------------------------
     1.757


    Standard Dev.
```

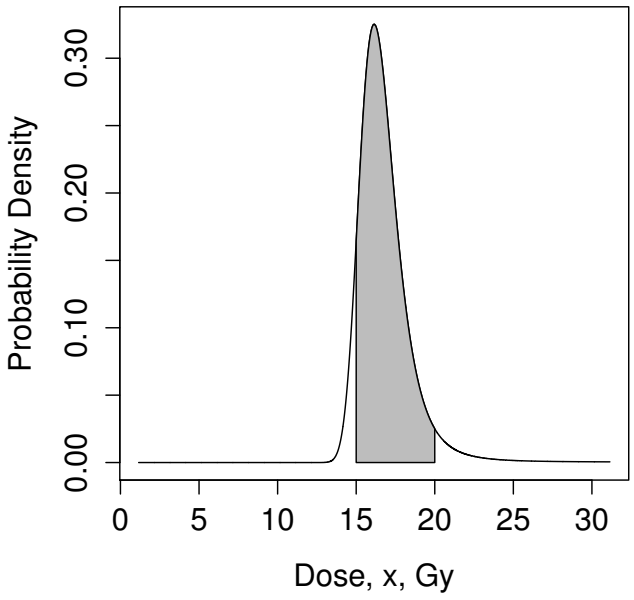


Figure 6.5: Calibrative density of the 17 Gy test data and the probability of the dose to be in (15, 20) Gy.

```
----------------------------
 0.11


95% CI
----------------------------
(1.544; 1.975)
```

The probability of a dose exposure between 1.544 and 1.977 is, as expected, approximately 0.95; it can be checked using

```
pr.dose.radir(ex3.a, as.numeric(substr(summary(ex3.a)[[4]],2,6)),
 as.numeric(substr(summary(ex3.a)[[4]],9,13)))

    [1] 0.9499784
```

The cumulative distribution for this particular example (Figure 6.4) can be plotted with

```
plot(ex3.a, distr=T)
```

The region under the calibrative curve for doses from 15 to 20 Gy can be plotted for the 17 Gy test data calibrative density (Figure 6.5) using

```
plot(ex3.d, prob=c(15,20))
```

## 6.4 Discussion

Biological dosimetry is necessary in many situations when dealing with radiation events, and a quick and accurate estimation of the radiation doses received by individuals undergoing medical radiation treatments or following a radiation accident is essential in many scenarios. This work presents a readily available package in the framework of the well known and widely distributed `R` software that represents a novel and useful tool to achieve this goal. It allows the user to estimate radiation doses received by an individual on the basis of a new inverse regression methodology and using the recently validated Bayesian framework which is able to compute true probability intervals.

The package uses as inputs the estimated parameters and variance–covariance matrix of the dose–response function, obtained using classical (frequentist) maximum likelihood methods. Therefore, the `radir` package can be seen as a complement of other existing packages (CABAS [18], DoseEstimate [1] or BioDoser [98]), which have to be used to obtain the required inputs.

The usability of the `radir` package in several common radiation related situations is demonstrated through the proposed examples, although the methodologies introduced in [39] and therefore the package itself could also be used in areas not related to biological dosimetry. Improvements planned for the package `radir` include the consideration of a normal prior dose distribution truncated at negative values ([15]), the ability to fit the dose–response curve given the calibration data, and the analysis of high-LET scenarios using the compound–Poisson models described in [39]. A Bayesian dose estimation of partial body irradiated blood samples following the new methodology of [40] will also be considered for future enhancement of the package.

# Chapter 7

# A new Bayesian model applied to cytogenetic partial body irradiation estimation

This chapter corresponds to the contents of [40].

**Abstract:** A new zero–inflated Poisson model is introduced for the estimation of partial body irradiation dose and fraction of body irradiated. Bayes factors are introduced as tools to help determine whether a data set of chromosomal aberrations obtained from a blood sample reflects partial or whole body irradiation. Two examples of simulated cytogenetic radiation exposure data are presented to demonstrate the usefulness of this methodology in cytogenetic biological dosimetry.

## 7.1 Introduction

The main goal of biological dosimetry is the estimation of the radiation dose received by an exposed individual, in scenarios such as radiation accidents or in radiotherapy settings. Radiation exposure produces breaks in the chromosomal DNA, and the resulting fragments can be repaired in different patterns from their original arrangement. Consequently, frequencies of chromosome aberrations including dicentrics and centric rings increase with the amount of absorbed radiation and are a reliable and very well established biomarker of radiation exposure. The estimation of the dose received by an individual requires dose-effect calibration curves, which are produced by exposing peripheral blood lymphocytes to a range of doses, simulating whole body irradiation. The manual of the International Atomic Energy Agency (IAEA) [42] describes the standards for these calibration experiments.

The construction of a calibration curve starts with the irradiation of blood samples from a healthy donor with different doses. Next, the counts of observed chromosomal aberrations are recorded. It is typically assumed that after exposure to X– or $\gamma$–rays the number of chromosomal aberrations per cell follows a

Poisson distribution with a population mean which is a linear-quadratic function of the dose. The set of parameters of this regression model is usually estimated by maximum likelihood (ML), recording the estimator, called maximum likelihood estimator (MLE) and its variance-covariance matrix, which measures the uncertainty of the estimated parameters. Thus, for an irradiated patient, a blood sample is taken and several tens to thousand lymphocytes are scored for chromosomal aberrations. The established approach for calculating the absorbed dose and its confidence limits is to use the classical inverse regression method described as a standard procedure in [42].

The mathematical distribution of chromosomal aberrations in partial body irradiation (PBI) scenarios is different from Poisson. The non irradiated cells contribute an extra amount of zeros in comparison to the distribution of chromosomal aberrations following whole body irradiation. This proportion of extra zeros can be described by the so–called zero inflated models. The zero counts of aberrations in this zero inflated process result from a mixture of cells with zero aberrations from the irradiated population and extra zeros which represent the non irradiated cells. Zero-inflated count models provide one method to account for the excess zeros in the data by modelling the data as a mixture of two distributions: a distribution taking a single value at zero and a count distribution, in this case Poisson.

The probability mass function (pmf) of a zero-inflated Poisson (ZIP) random variable $\mathcal{Y}$ taking $k$ counts is defined as

$$\mathrm{P}(\mathcal{Y}=k;\mu,\omega)=\begin{cases}\omega+(1-\omega)\mathrm{e}^{-\mu} & \text{if } k=0\\ (1-\omega)\dfrac{\mu^k\mathrm{e}^{-\mu}}{k!} & \text{otherwise,}\end{cases}$$

where $0\leq\omega\leq 1$ is the proportion of extra zeros. Note that this pmf is expressed in terms of the Poisson pmf with population mean $\mu$.

The standard procedure to detect PBI is to calculate the sample dispersion index, the ratio of the sample variance to the sample mean, of the counts obtained from the blood sample and then to use the $u$–test to reject or not the Poisson assumption [80], because of the overdispersion caused by the excess of zeros [42]. If the Poisson hypothesis is not rejected the recommendation is to perform a whole body cytogenetic dose estimation. Otherwise, if the Poisson hypothesis is rejected, the yield of chromosomal aberrations in the irradiated fraction, and the fraction of cells irradiated can both be estimated by the MLE of the parameters of a ZIP distribution. Then, the absorbed dose is estimated by performing the whole body cytogenetic dose estimation method for the estimated yield of chromosomal aberrations in the irradiated fraction. Finally, the fraction of the body irradiated is estimated by a formula described in [42].

The alternative to the classic methodology described above is Bayesian analysis, which lends itself to retrospective dosimetry, because it presents the probability of an event in terms of prior knowledge about its expectation and uncertainty. There is a recent publication [39] with Bayesian models that allow the reconsideration of most of the published examples of radiation exposures that were analysed using the classical methods for dose estimation following homogeneous exposure. A review of Bayesian methods in biodosimetry can be found in [4], containing one appendix dedicated to the description and derivation of a Bayesian model for cytogenetic dose estimation. There are also two recent programs, CytoBayesJ [3], which provides some basic software tools for Bayesian

Table 7.1: Calibration data.

| Dose (Gy) | Cells | Dic+CR |
|---|---|---|
| 0.00 | 5389 | 5 |
| 1.96 | 924 | 279 |
| 3.96 | 475 | 504 |
| 5.88 | 331 | 675 |
| 7.84 | 300 | 1028 |
| 9.59 | 157 | 735 |
| 16.36 | 34 | 295 |
| 20.30 | 60 | 927 |

analysis of cytogenetic radiation dosimetry data, and radir [65], which is an R [81] package for applying the models in [39].

However, in the field of cytogenetic biodosimetry there are currently no Bayesian methodologies in use for the detection of PBI. In this chapter a new cytogenetic method is presented for the estimation of the absorbed dose and the FBI in PBI scenarios, based on prior determination of whether a partial or whole body exposure had been received.

## 7.2 Calibration and test data

Table 7.1 displays *in vitro* dose response data for dicentrics plus centric rings (Dic+CR) obtained in a recent calibration experiment within a high dose range [95] that was performed in accordance with the IAEA requirements [42]. The detailed protocols and conditions of blood irradiation, lymphocyte culturing, cell fixing, metaphase staining, aberration scoring and QA/QC procedures were fully presented in [95]. The dose rate was 0.5 Gy $\text{min}^{-1}$ for 2, 4 and 6 Gy and 1.2 Gy $\text{min}^{-1}$ for 8, 10, 16 and 20 Gy.

The test datasets representing the Dic+CR scored after the simulated PBI (Table 7.2) were generated within a separate experiment, which was carried out in the framework of the IAEA Coordinated Research Project E.3.50.08. They have not yet been published elsewhere. Peripheral blood was taken from one healthy male volunteer, corresponding to Donor I in calibration experiment [95], to minimise the potential impact of intrinsic individual variations in our study. The donor participation was with written informed consent and in accordance with the institutional ethics protocol.

Blood sampling and irradiation conditions, cell culturing details and aberration scoring criteria in the PBI simulation experiment were identical to those used in the calibration experiment [95]. Briefly, the donor's blood samples were collected into Vacutainer$^{\text{TM}}$ tubes with lithium heparin anticoagulant and exposed to 1.98 and 11.88 Gy $^{60}$Co acute $\gamma$-rays in accordance with the IAEA requirements [42]. The dose rate was 0.5 Gy $\text{min}^{-1}$ for 2 Gy and 1.2 Gy $\text{min}^{-1}$ for 12 Gy. The irradiated samples were accompanied by a sham treated zero dose control sample.

To simulate PBI, the unirradiated and irradiated blood at each dose was mixed, so that the exposed blood fraction comprised 90% for 12 Gy and 10% for 2 Gy, thus the equivalent whole body doses are 10.8 and 0.2 Gy respectively.

Table 7.2: Frequency distributions of the number of Dic+CR of the test data samples; sample 1, 2 Gy and 10% irradiated fraction, and sample 2, 12 Gy and 90% irradiated fraction. [a]4 cells with 3 Dic+CR, 11 with 4, 10 with 5, 9 with 6, 7 with 7, 3 with 8, 2 with 9 and 1 with 11.

| | Number of Dic+CR | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Sample | 0 | 1 | 2 | ≥3 | $\bar{y}$ | $d$ | $u$ | $2\log BF$ |
| 1 | 1043 | 16 | 3 | | 0.021 | 1.253 | 5.967 | 9.410 |
| 2 | 148 | | 4 | [a]47 | 1.357 | 4.667 | 36.551 | 505.480 |

Lymphocyte cultures were set up according to the standard technique [42] with cell cycle control using BrdU. After 52 h of culturing cells were fixed and conventional chromosomal analysis was carried out on metaphase preparations stained with the fluorescence–plus–Giemsa method. The microscopy of all preparations was performed on coded slides by the same cytogeneticist, to avoid an observer bias and possible inter-operator variability. The scorer has over 20 years of experience in microscopy of aberrations and participated actively in generating the calibration data, and scored essential number of cells and aberrations at each dose response point, with all QA/QC procedures described in [95]. Chromosomal damage was scored only in the 1$^{\text{st}}$ *in vitro* division, diploid metaphases and registered using the stringent aberration scoring criteria [42]. All unstable chromosome type aberrations were recorded; polycentrics were converted into the equivalent number of dicentrics. For the present work the yields of dicentrics and centric rings accompanied by a fragment were selected.

## 7.3 The zero–inflated Poisson model

Assuming the test (patient) data $y = \{y_1, y_2, \ldots, y_n\}$, formed by $n$ count data observations, representing the number of chromosomal aberrations in $n$ blood cells, to be ZIP$(\mu, \omega)$ ($\mu$ the population mean and $\omega$ the proportion of extra zeros) distributed, the likelihood of this sample, following [27], is proportional to

$$\mathrm{L}(y|\mu,\omega) = \prod_{i=1}^{n} \mathrm{P}(y_i) \propto \mu^s \sum_{j=1}^{n_0} \binom{n_0}{j} \frac{\omega^j (1-\omega)^{n-j}}{\mathrm{e}^{(n-j)\mu}}, \tag{7.1}$$

where $n_0$ and $s$ are the sample frequency of zeros and the sum of the total number of chromosomal aberrations, respectively. Multiplying and dividing by $(n-j)^s$ inside the summatory of (7.1), the likelihood leads to

$$\mathrm{L}(y|\mu,\omega) \propto \sum_{j=1}^{n_0} \binom{n_0}{j} \frac{\omega^j (1-\omega)^{n-j}}{(n-j)^s} \frac{[(n-j)\mu]^s}{\mathrm{e}^{(n-j)\mu}}. \tag{7.2}$$

Note that the last fraction of (7.2) is proportional to the probability of a Poisson distribution with population mean $(n-j)\mu$ taking $s$ counts.

## 7.4 Bayes factor

The Bayes factor is the main Bayesian model comparison tool. Given a dataset $y$, the probabilities of two different models, in this case ZIP and Poisson, on $y$

are compared. Following [11] this Bayes factor remains

$$BF = \frac{n_0!}{(n+1)!}\sum_{j=0}^{n_0}\frac{(n-j)!}{(n_0-j)!}(1-j/n)^{-(s+1/2)}, \tag{7.3}$$

and in accordance with [45] $2\log BF > 0, 2, 6, 10$ gives 'weak positive', 'positive', 'strong' and 'very strong' evidence in support of the ZIP model, respectively.

## 7.5 Dose and body fraction estimation

Here, the probability of extra zeros $\omega$ represents the proportion of non–irradiated cells, and $D$ the absorbed dose in the blood sample $y$. Following [42] the fraction of the body irradiated, $F$, is calculated as

$$F = \frac{(1-\omega)\mathrm{e}^{D/d_0}}{\omega + (1-\omega)\mathrm{e}^{D/d_0}}, \tag{7.4}$$

where $d_0$ is the 37% cell survival dose, with experimental evidence to be between 2.7 and 3.5 Gy [42]. Following Equation (7.4), the proportion of non–irradiated cells $\omega$ as a function of $x$, $F$ and $d_0$ results,

$$\omega = \frac{1-F}{F\mathrm{e}^{-D/d_0} - F + 1}, \tag{7.5}$$

thus substituting $\omega$ by Equation (7.5), the likelihood in (7.2) results,

$$\begin{aligned}\mathscr{L}(y|\mu, F, d_0) = &\ \mathrm{L}\left(y\,\middle|\mu, \frac{1-F}{F\mathrm{e}^{-D/d_0} - F + 1}\right)\\ \propto &\ (F\mathrm{e}^{-D/d_0} - F + 1)^{-n}\sum_{j=1}^{n_0}\binom{n_0}{j}\frac{F^{n-j}(1-F)^j}{(n-j)^s}\frac{[(n-j)\mu]^s}{\mathrm{e}^{(n-j)\mu}},\end{aligned} \tag{7.6}$$

Let $\mu = \mathrm{f}(D, \beta)$ be the calibration curve, and $\mathrm{v}(D, \hat{\beta}) = \nabla \cdot \hat{\Sigma}_{\hat{\beta}} \cdot \nabla^T$, where $\nabla$ denotes the gradient of $\mathrm{f}(D, \beta)$ respect $\beta$ and $\hat{\Sigma}_{\hat{\beta}}$ is the variance matrix of $\beta$, the *mean prior* is defined for the purposes of inverse regression as in [39], i.e.

$$\mu|D \sim \mathrm{Gamma}\left(\frac{\mathrm{f}(D,\hat{\beta})^2}{\mathrm{v}(D,\hat{\beta})}, \frac{\mathrm{f}(D,\hat{\beta})}{\mathrm{v}(D,\hat{\beta})}\right). \tag{7.7}$$

Analogously to [39], because the knowledge of $\mu$ implies the knowledge of $D$, then $\mathscr{L}(y|\mu, F, d_0) = \mathscr{L}(y|D, \mu, F, d_0)$, (7.6). Therefore, an application of Bayes' theorem shows the expression of the likelihood of the absorbed dose $D$ and the FBI $F$ for the given the test data $y$,

$$\begin{aligned}\mathcal{L}(y|D, F, d_0) \propto &\ \int_0^{+\infty}\mathscr{L}(y|\mu, F, d_0)\mathrm{P}(\mu|D)\mathrm{d}\mu\\ = &\ (F\mathrm{e}^{-D/d_0} - F + 1)^{-n}\sum_{j=1}^{n_0}\binom{n_0}{j}\frac{F^{n-j}(1-F)^j}{(n-j)^s}\mathrm{P}(\mathcal{X}_j = s|D),\end{aligned} \tag{7.8}$$

where $\mathcal{X}_j$ is a random variable negative binomial distribution with mean $(n-j)\mathrm{f}(x,\hat{\beta})$ and variance $(n-j)\mathrm{f}(D,\hat{\beta})+(n-j)^2\mathrm{v}(D,\hat{\beta})$. Different fonts to indicate the likelihood as function of different variables have been used in Equations (7.1), (7.6) and (7.8).

Considering $D$, $F$ and $d_0$ as independent random variables, consequently their joint probability density remains $\mathrm{P}(D,F,d_0)=\mathrm{P}(D)\mathrm{P}(F)\mathrm{P}(d_0)$. For the FBI, F, a beta prior is applied. For the cell survival dose, $d_0$, a uniform prior $\mathcal{U}(c,d)$ is applied, where $c\geq 2.7$ and $d\leq 3.5$. These least informative choices are applied in case there is no information about $F$ and $d_0$, $F\sim\mathrm{Beta}(1,1)\equiv\mathcal{U}(0,1)$ and $d_0\sim\mathcal{U}(2.7,3.5)$.

A prior gamma distribution is applied for the absorbed dose $D$, as in [33]. This *prior information* can be defined through expert judgement or by an empirical Bayes method. The empirical Bayes method is applied using the MLE of $D$ and its standard error from (7.8), i.e. a gamma distribution with mean $\hat{D}$ and variance $\hat{\sigma}^2_{\hat{D}}$,

$$D\sim\mathrm{Gamma}\left(a=\frac{\hat{D}^2}{\hat{\sigma}^2_{\hat{D}}},b=\frac{\hat{D}}{\hat{\sigma}^2_{\hat{D}}}\right). \tag{7.9}$$

Therefore, the joint posterior density remains,

$$\mathrm{P}(D,F,d_0|y)=\frac{\mathcal{L}(y|D,F,d_0)\mathrm{P}(D,F,d_0)}{\displaystyle\int\mathcal{L}(y|D,F,d_0)\mathrm{P}(D,F,d_0)\mathrm{d}D\mathrm{d}F\mathrm{d}d_0}. \tag{7.10}$$

To calculate this joint posterior density (7.10), with a non–tractable form, and its marginal densities, the acceptance–rejection method is used to simulate the posterior distribution. This approach generates a random vector $(\mathcal{D},\mathcal{F})$ of size $m$ following $\mathrm{P}(D,F,d_0|y)$:

S1 Generate $u$ from $\mathcal{U}(0,1)$.

S2 Generate a random variate $D^*$ from $\mathrm{Gamma}(a,b)$, a random variate $F^*$ from $\mathrm{Beta}(\alpha,\beta)$, and a random variate $d_0^*$ from $\mathcal{U}(c,d)$, all them independent of $u$.

S3 If $u\leq\mathcal{L}(y|D^*,F^*,d_0^*)/\mathcal{L}(y|\hat{D},\hat{F},\hat{d}_0)$, then set $(\mathcal{D},\mathcal{F})\leftarrow(\mathcal{D},\mathcal{F})\bigcup(D^*,F^*)$. While the size of $(\mathcal{D},\mathcal{F})<m$, go to S1.

Here $\hat{D}$, $\hat{F}$ and $\hat{d}_0$ represent the MLE from (7.8) of the absorbed dose, the FBI and the cell survival dose, respectively. Thus $\mathcal{L}(y|\hat{D},\hat{F},\hat{d}_0)$ is the maximum of (7.8). The efficiency is $\mathrm{eff}=m/M$, where $M$ is the total number of times that this process is repeated until $m$ simulations are achieved. An approximation of $C$, the normalising constant of the joint posterior density, is $C\approx\mathrm{eff}\cdot\mathcal{L}(y|D^*,F^*,d_0^*)$.

## 7.6 Results

The values in the last column of Table 7.2 show that the Bayes factors, Equation (7.3), give 'strong' and 'very strong' evidence in support to the ZIP assumption respectively for samples 1 and 2.

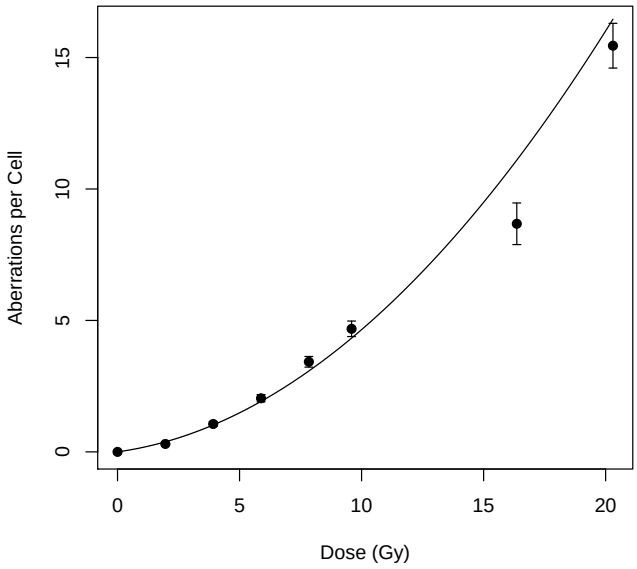


Figure 7.1: Observed means (dots), plus/minus twice their standard errors (error bars), and predicted means (solid line) of the number of Dic+CR.

Table 7.3: Statistics summary of the marginal posterior densities of the absorbed dose and the FBI ($D$ in Gy, $F$ in fraction).

| Sample | Variable | Modal | Expected | SD | 95% CI |
|---|---|---|---|---|---|
| 1 | $D\|y$ | 1.392 | 1.502 | 0.498 | (0.658, 2.577) |
| | $F\|y$ | 0.111 | 0.134 | 0.048 | (0.071, 0.250) |
| 2 | $D\|y$ | 10.724 | 10.736 | 0.281 | (10.182, 11.277) |
| | $F\|y$ | 0.910 | 0.910 | 0.025 | (0.856, 0.951) |

Following the notation in Expression (7.1), in sample 1 $n_0 = 1043$, $s = 22$ and $n = 1062$, i.e. 1043 cells free of Dic+CR and 22 scored Dic+CR in 1062 blood cells, and in sample 2 $n_0 = 148$, $s = 270$ and $n = 199$.

In this example, the appropriate dose-response curve is linear-quadratic, $\mathrm{f}(D, \beta) = \beta_2 D^2 + \beta_1 D + \beta_0$ (Figure 7.1). For high doses, Gompertz–type dose responses curves are also suitable, as shown in a recent publication [79]. Note that the saturation effect is often quoted as leading to incorrect dose estimates at doses $>\approx$5 Gy, however, for low LET irradiation, doses of up to 20 Gy have been demonstrated to induce approximately linear–quadratic responses in terms of chromosome aberrations (e.g. [83]).

Following [95] it is assumed that the distributions of Dic+CR is Poisson and thus the ML parameter estimates and their estimated covariance matrix are the following:

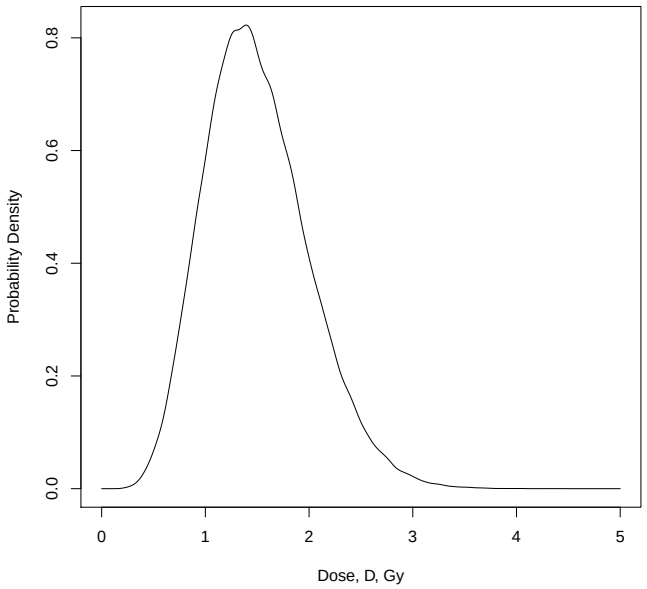


Figure 7.2: Marginal posterior dose density for sample 1, 2 Gy and 10% irradiated fraction.

- Fitted coefficients:

$$\hat{\beta}_0 = 9.073 \cdot 10^{-4}, \quad \hat{\beta}_1 = 1.291 \cdot 10^{-1}, \\ \hat{\beta}_2 = 3.357 \cdot 10^{-2}.$$

- Estimated covariance matrix:

$$\hat{\Sigma}_{\hat{\beta}} = \begin{pmatrix} 168.354 & -73.146 & 4.933 \\ -73.146 & 85611.557 & -9365.947 \\ 4.933 & -9365.947 & 1560.911 \end{pmatrix} \cdot 10^{-9}.$$

According to (7.7), $\mu|D$ will follow a Gamma distribution with mean $\mathrm{f}(D, \hat{\beta}) = \hat{\beta}_2 D^2 + \hat{\beta}_1 D + \hat{\beta}_0$ and variance $\mathrm{v}(x, \hat{\beta}) = \nabla \cdot \hat{\Sigma}_{\hat{\beta}} \cdot \nabla^t$, where:

$$\nabla = \left( \frac{\partial \mathrm{f}}{\partial \beta_0}, \frac{\partial \mathrm{f}}{\partial \beta_1}, \frac{\partial \mathrm{f}}{\partial \beta_2} \right) = \left(1, D, D^2\right),$$

and therefore $\mathrm{v}(D, \hat{\beta}) = \hat{\Sigma}_{33} D^4 + 2\hat{\Sigma}_{23} D^3 + 2\hat{\Sigma}_{13} D^2 + \hat{\Sigma}_{22} D^2 + 2\hat{\Sigma}_{12} D + \hat{\Sigma}_{11}$.

According to Expression (7.8), $\mathrm{P}(\mathcal{X}_j = s|D)$ represents the probability of a negative binomial random variable taking a value of 74 counts for sample 1, 270 for sample 2, with mean $(n-j)(3.357 \cdot 10^{-2} D^2 + 1.291 \cdot 10^{-1} D + 9.073 \cdot 10^{-4})$ and variance $(n-j)(3.357 \cdot 10^{-2} D^2 + 1.291 \cdot 10^{-1} D + 9.073 \cdot 10^{-4}) + (n-j)^2 (15.609 D^4 - 187.319 D^3 + 856.214 D^2 - 1.463 D + 1.683) \cdot 10^{-7}$.

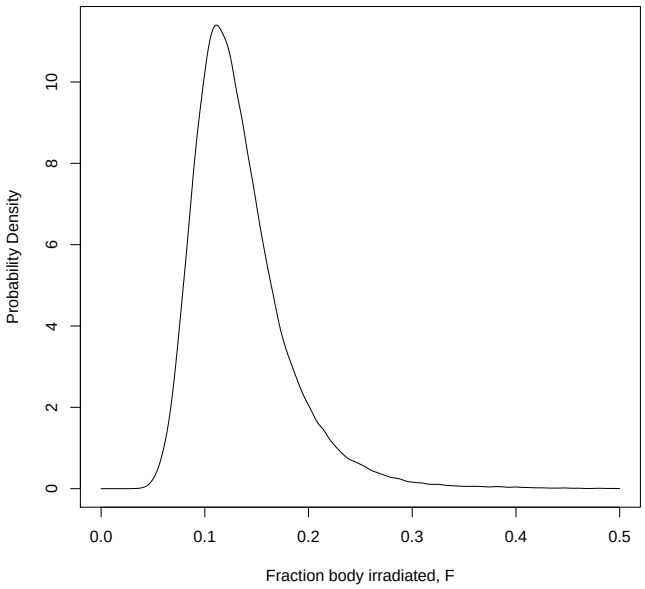


Figure 7.3: Marginal posterior FBI density for sample 1, 2 Gy and 10% irradiated fraction.

To define the prior of $D$, the empirical Bayes method is applied. The MLE of sample 1 for $D$ is $\hat{D} = 1.636$ and $\hat{\sigma}^2_{\hat{D}} = 0.508$, then according to (7.9) the *prior* of the absorbed dose is $D \sim \text{Gamma}(5.271, 3.221)$. For sample 2, $\hat{D} = 10.746$, $\hat{\sigma}^2_{\hat{D}} = 0.156$, and $x \sim \text{Gamma}(739.362, 68.800)$.

For both samples $F \sim \mathcal{U}(0, 1)$ and $d_0 \sim \mathcal{U}(2.7, 3.5)$, the less informative priors.

Table 7.3 and Figures 7.2, 7.3, 7.4, 7.3, 7.6, 7.7 show the results of the posterior densities. These results are sensible compared with the real doses and FBIs, only the real absorbed dose in sample 2 is not covered by the 95% credible interval (CI). It is remarkable that the marginal densities of the FBI in both samples have narrow shapes, localising the FBI most likely values in a small range.

Applying Dolphin's method in the classical framework ([42]) to sample 1 (2 Gy, 10% FBI) gives a dose estimate of 1.631 Gy with a confidence interval of (0, 2.994) Gy, and the estimation of the FBI is 0.113; for sample 2 (12 Gy, 90% FBI) the dose estimate is 10.658 Gy with a confidence interval of (9.634, 11.717) Gy and an FBI estimate of 0.922.

In trying to demonstrate the practical applicability of the suggested method, we selected two marginal scenarios of PBI, which can be of great interest for clinicians which are traditionally not easy for conventional biodosimetry. Very localized (10%) exposure to 2 Gy and sub–total (90%) exposure to 12 Gy could easily be misinterpreted as total body irradiation, and such a mistake may lead to serious clinical consequences. In the 2 Gy/10% scenario the calculations of

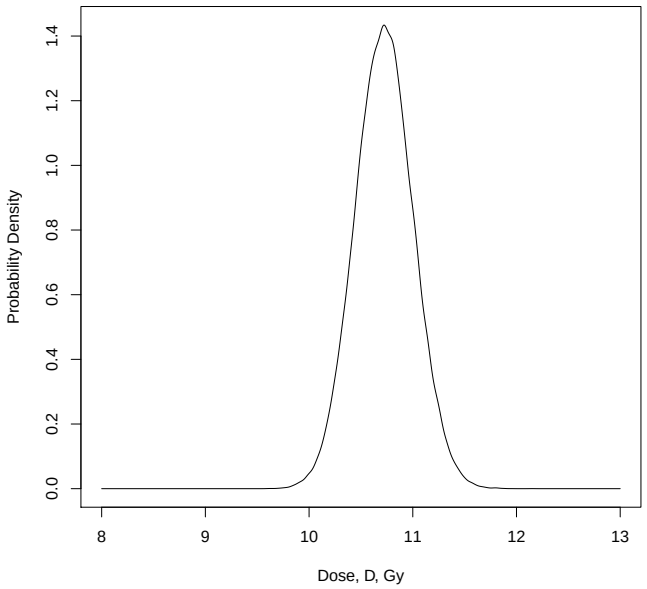


Figure 7.4: Marginal posterior dose density for sample 2, 12 Gy and 90% irradiated fraction.

cancer risk to organs would be based on the underestimated dose, and for 12 Gy/90% ignoring the partial nature of exposure might result in an incorrect decision regarding the necessity of bone marrow transplantation. However, the new technique presented here should help to avoid such mistakes. In Figures 7.2, 7.3, 7.4 and 7.5 the shapes of the curves clearly indicate the partial nature of exposure, and the maximum probable values of both $D$ and $F$ are in acceptable agreement with the true values. It is important to remark the histograms in Figures 7.6 and 7.7 because to the best of the authors knowledge, this is the first time that the correlation between the estimates of the dose and irradiated fraction in cytogenetic dosimetry has been highlighted and visualised. From this the end user of the information (e.g. a clinician dealing with an irradiated patient) can make a judgment by analysing the most probable dose/volume estimates to the clinical symptoms.

In contrast to the classical estimation method for PBI, the Bayesian methods return results in the form of probability densities which intrinsically contain both modal dose and uncertainty information, and more information such as the skewness can be observed. In addition, the cell survival dose is considered a random variable, whilst in the classical method it is a point value, fixed before the estimation. The posterior joint density for $(D, F)$ allows the correlation to be studied between the absorbed dose and the FBI. However, the prior distribution choices influence the final results. Potential future work includes to study the application of informative priors for $d_0$.

The number of simulations for both samples is $m = 100000$. This is a huge number of simulations and consequently the joint posterior density calculation

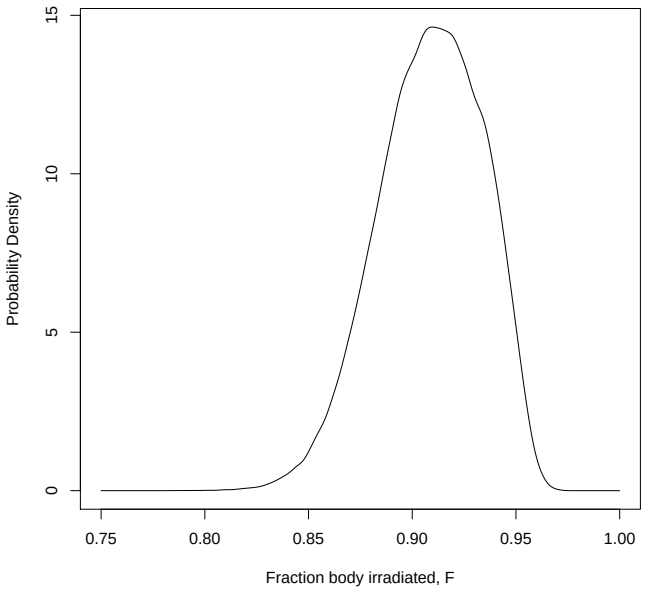


Figure 7.5: Marginal posterior FBI density for sample 1, 12 Gy and 90% irradiated fraction.

takes time (≈40 min). It is necessary to use such a large number of simulations to plot accurate bivariate histograms for $(D, F)$ of both samples, Figures 7.6 and 7.7. In case there is only interest in plotting the marginal posteriors, a lower number of simulations is recommended and this calculation will be quicker (e.g. for $m = 10000$ it takes ≈3 min).

## 7.7 Conclusions

This chapter presents a new ZIP Bayesian method for PBI estimation. The ZIP distribution is introduced and its application in cytogenetic biodosimetry is described.

The method to estimate the absorbed dose and the FBI is also derived. To use this methodology, only the estimates of the parameters and covariance matrix of the dose–response curve are required, which are available from the classical analysis [42] and many examples of which are published. Acceptance–rejection sampling is applied to simulate the joint posterior density.

Finally, the illustrative examples show the application of this methodology in biological dosimetry. This methodology has other potential uses, it could be applied for other types of chromosomal aberration assays, biomarkers and agents (e.g. chemicals).

Further evolution of the methodology presented here will focus on the development of zero–inflated compound Poisson models for estimating dose and fraction of the body irradiated in partial body exposure scenarios where the

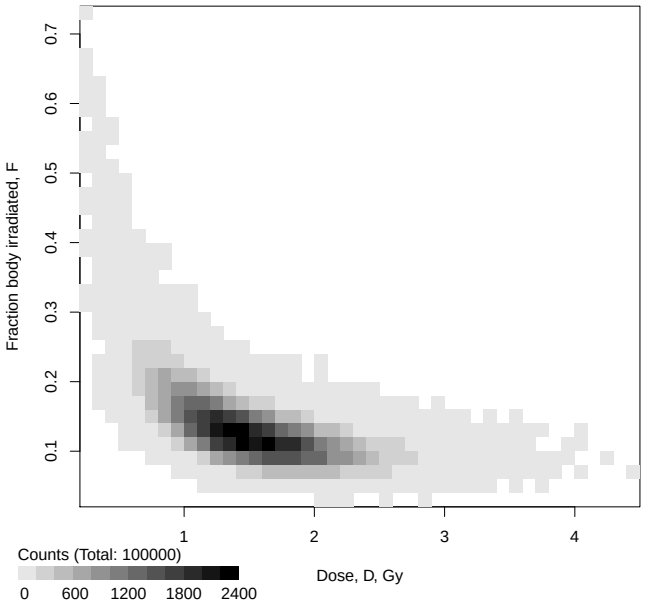


Figure 7.6: Histogram of the joint posterior density of $(D, F)$ for sample 1, 2 Gy and 10% of irradiated fraction.

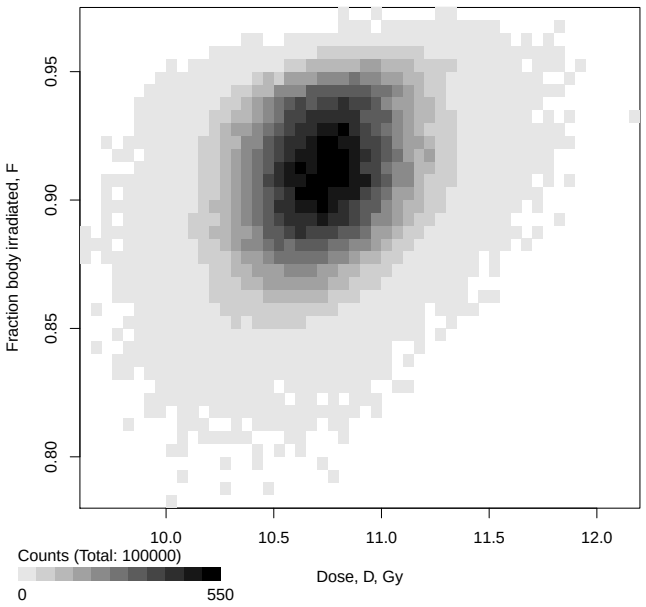


Figure 7.7: Histogram of the joint posterior density of $(D, F)$ for sample 2, 12 Gy and 90% of irradiated fraction.

chromosomal aberration yield in the irradiated fraction is overdispersed, e.g. high LET sources like $\alpha$–particles.

# Chapter 8

# Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study

Submitted to *Biometrical Journal*.
**Abstract:** Within the field of cytogenetic biodosimetry, Poisson regression is the classical approach for modelling the number of chromosome aberrations as a function of radiation dose. However, it is common to find data that exhibit overdispersion. In practice, the assumption of equidispersion may be violated due to unobserved heterogeneity in the cell population, which will render the variance of observed aberration counts larger than their mean, and/or the frequency of zero counts greater than expected for the Poisson distribution. This phenomenon is observable for both full and partial body exposure, but more pronounced for the latter. In this work, different methodologies for analysing cytogenetic chromosomal aberrations datasets are compared, with special focus on zero–inflated Poisson and zero–inflated negative binomial models. A score test for testing for zero–inflation in Poisson regression models under the identity link is also developed.
**Keywords:** biological dosimetry, chromosome aberrations, count data, overdispersion, zero–inflation, score tests.

## 8.1 Introduction

Data from biological systems regarding the effects of environmental or manmade mutagens frequently consist of count variables. This is the case in biological dosimetry, where the measurement of chromosome aberration frequencies in human lymphocytes is used for assessing absorbed doses of ionising radiation to individuals. For that purpose, dose–effect calibration curves are required

which are produced by irradiating peripheral blood lymphocytes to a range of doses and quantifying the amount of damage induced by radiation at a cellular level, for instance by counting dicentrics or micronuclei (IAEA, 2011). That is, $d$ blood samples from a healthy donor are irradiated with several doses $x_i$, $i = 1 \ldots, d$. Then for each irradiated sample, $n_i$ cells are examined and the number of observed chromosomal aberrations $y_{ij}$, $j = 1, \ldots, n_i$ is recorded. The aberrations most commonly analyzed are the dicentrics, centric rings, and micronuclei.

These chromosomal aberrations appear because when cells are exposed to radiation, breaks are induced in the chromosomal DNA, and the broken fragments may rejoin incorrectly. Therefore, the frequency of chromosome aberrations increases with the amount of radiation and is a reliable and very well established biological indicator of radiation absorbed dose. Dicentrics are the interchange between the fragments of two separate chromosomes resulting in unstable, aberrant chromosomes with two centromeres. A ring chromosome, or centric ring, is an exchange between two breaks on separate arms of the same chromosome and is also accompanied by an acentric fragment (chromosome without centromere). Micronuclei are lagging chromosomal fragments or whole chromosomes at anaphase that are not included in the nuclei of daughter cell.

For such count data, the Poisson distribution is the most widely recognized and commonly used distribution and constitutes the standard framework for explaining the relationship between the outcome variable and the dose (Lloyd and Edwards, 1983; IAEA, 2011). However, in practice, the assumption of equidispersion implicit in the Poisson distribution is often violated, which is a well–known effect under high LET (Linear Energy Transfer) radiation, also known as densely ionising radiation (IAEA, 2011). Moreover, the distributions of micronuclei are in general overdispersed for both high and low LET radiation exposure.

The focus of the research presented in this manuscript is the identification of adequate response distributions for the modelling of cytogenetic dose–response curves. The cytogenetic dose estimation is a subsequent inverse regression problem that depends on this previous curve fitting. If the initial response distribution is incorrectly specified, this will impact on the accuracy of the model parameter estimates of the fitted curve and, more strongly, of their standard errors. In addition, the inverse regression step is sensitive to the initial model specification, and may behave unreliably if that specification is incorrect. Summarizing, an incorrectly specified response distribution may or may not lead to reasonable dose estimates, but it will certainly lead to an incorrect assessment of the uncertainty associated to these dose estimates. This subsequent inverse regression step is not the subject of this manuscript, see Higueras et al. (2015a, 2015b) for recent advances in this respect.

Due to the mentioned violations of the Poisson distribution, other distributions have been considered in the literature for dealing with overdispersed data in biodosimetry. These alternatives include the negative binomial distribution, which has been shown to accurately characterize aberration data in cases of overdispersion (Brame and Groer, 2002); the Neyman type A distribution, which has been shown to be useful for characterisation of aberration induced by high LET radiation (Gudowska–Nowak *et al.*, 2007) and the univariate $r$th-order Hermite distributions (Puig and Barquinero, 2011). These distributions have recently been tested for suitability to a selection of chromosome aberra-

tion data collected in different exposure scenarios (Ainsbury *et al.* 2013a [2]) and used for cytogenetic dose estimation through a Bayesian-like inverse regression technique (Higueras *et al.*, 2015a). Further, Poisson–inverse Gaussian and Pólya–Aeppli distributions have been considered in Puig and Valero (2006).

Also, a commonly observed characteristic of count data is the number of zeros in the sample exceeding the expected number of zeros generated by a Poisson distribution having the same mean. This phenomenon, known as zero–inflation, is frequently related to overdispersion. Distributions which account for overdispersion will also – to some extent – allow for zero-inflation. For instance, the families of Compound Poisson and Mixed Poisson distributions (which include the distributions mentioned in the previous paragraph as special cases) are overdispersed and zero-inflated.

However, the extra zeros (relative to the Poisson model) generated by these models may still be insufficient to account for the total observed number of zeros in the data. Count datasets with an excessive number of zero outcomes are abundant in many disciplines such as manufacturing applications (Lambert, 1992), medicine (Böhning *et al.*, 1999), econometrics (Gurmu *et al.*, 1999) and agriculture (Hall, 2000). In most of these works, a special kind of zero–inflated models are considered, using a mixture of a distribution degenerate at zero and a count distribution such a Poisson or a negative binomial. These models can be especially useful in partial body irradiation scenarios which feature a mixture of populations of non-irradiated and irradiated cells.

In this manuscript we will introduce and advocate the use of zero–inflated models for cytogenetic count data. We will compare zero–inflated models to other models previously proposed in the field of radiation biodosimetry, and we will devote particular attention to the question of whether overdispersion needs to be taken into account on top of the zero–inflation. The manuscript is organized as follows: In Section 8.2, zero–inflated Poisson and zero–inflated negative binomial models are reviewed. The models are applied to several datasets with different radiation exposure patterns in Section 8.3. In Section 8.4 we provide a small simulation study in a radiation induced chromosome aberration context to study the identifiability of zero–inflated and overdispersed regression models. The paper is concluded in Section 8.5.

The supplementary material in [70] contains the data sets, along with a description of the code used for the data analysis, as well as the mathematical forms of count distributions used in Section 8.3. It further contains the derivations for a score test for zero–inflation under the identity link which is employed in Sections 8.3 and 8.4.

## 8.2 Zero–inflated regression models applied to biodosimetry

In this section, zero–inflated regression models are reviewed in a general framework in Section 8.2.1 and details on how these models are applied for modelling the number of chromosome aberrations as a function of radiation doses are given in Section 8.2.2.

### 8.2.1 Zero–inflated count regression overview

Zero–inflated count models provide one method to account for the excess zeros in data by modelling the data as a mixture of two distributions: a distribution taking a single value at zero and a count distribution such as Poisson or negative binomial distributions.

The zero–inflated Poisson (ZIP) regression model was first introduced by Lambert (1992) who applied the model to the data collected from a quality control study. Since then, the ZIP regression model has been applied in many and different fields, such as, dental epidemiology (Böhning *et al.* 1999 [13]), occupational health (Lee *et al.*, 2001), and children's growth and development (Cheung 2002).

Let $Y_{ij}$, $i = 1, \ldots, d$, $j = 1, \ldots n_i$ be the response variable which in our context represents numbers of chromosomal aberrations at dose level $i$ for cell $j$. A ZIP regression model is defined as

$$\mathrm{P}(Y_{ij} = y_{ij}) = \begin{cases} p_i + (1 - p_i)\exp(-\lambda_i), & y_i = 0, \\ (1 - p_i)\exp(-\lambda_i)\lambda_i^{y_{ij}}/y_{ij}!, & y_{ij} > 0, \end{cases}$$

where $0 \leq p_i \leq 1$ and $\lambda_i > 0$. For the ZIP, $\mathrm{E}(Y_{ij}) = (1 - p_i)\lambda_i = \mu_i$ and $\mathrm{Var}(Y_{ij}) = (1 - p_i)\lambda_i(1 + p_i\lambda_i)$. Both the mean $\lambda_i$ of the underlying Poisson distribution and the mixture parameter $p_i$ (also referred to as 'zero–inflation parameter') can depend on vectors of covariates.

Since $\mathrm{Var}(Y_{ij}) = \mu_i(1 + p_i\lambda_i) \geq \mu_i$ it is clear that zero-inflation can be considered as a special form of overdispersion. When overdispersion is attributed to the large number of zeros with respect to the Poisson model, a ZIP model may provide a good fit. A ZIP model assumes that the zero observations have two different origins: some of them are zeros produced at random by the Poisson distribution, while some others (with proportion $p_i$) are "structural". The structural zeros have to be justified by the nature of data (in our case, by non–irradiated lymphocytes; for instance after partial body exposure). In addition, there may exist another source of overdispersion that cannot be attributed to the excess zeros. That is, even after accounting for zero–inflation, the non–zero part of the count distribution may be overdispersed (in our context, this will be mainly observed for densely ionising radiation). For dealing with this situation, Greene (1994) introduced an extended version of the negative binomial model for excess zero count data, the zero–inflated negative binomial (ZINB). In that case, when the overdispersion is both due to the heterogeneity of data and the excess of zeros, the ZINB regression model often is more appropriate than the ZIP.

For the ZINB regression model, the probability mass function of the response variable $Y_{ij}$ $(i = 1, \ldots, d,\ j = 1, \ldots, n_i)$ is given by

$$\mathrm{P}(Y_{ij} = y_{ij}) = \begin{cases} p_i + (1 - p_i)(1 + \alpha\lambda_i^c)^{-\lambda_i^{1-c}/\alpha}, & y_{ij} = 0, \\ (1 - p_i)\frac{\Gamma(y_{ij} + \lambda_i^{1-c}/\alpha)}{y_{ij}!\Gamma(\lambda_i^{1-c}/\alpha)}(1 + \alpha\lambda_i^c)^{-\lambda_i^{1-c}/\alpha}(1 + \lambda_i^{-c}/\alpha)^{-y_{ij}}, & y_{ij} > 0, \end{cases}$$

where $\alpha > 0$ is an overdispersion parameter, and the index $c \in \{0, 1\}$ identifies the form of the underlying negative binomial distribution. These models will be denoted by ZINB1 and ZINB2, respectively. The mean and variance of the ZINB distribution are $\mathrm{E}(Y_{ij}) = (1-p_i)\lambda_i = \mu_i$ and $\mathrm{Var}(Y_{ij}) = (1-p_i)\lambda_i(1+p_i\lambda_i+\alpha\lambda_i^c)$, respectively. The ZINB model reduces to the ZIP model as $\alpha \to 0$, in analogy to the relationship between the negative binomial and the Poisson distribution.

### 8.2.2 Application to biological dosimetry

Count regression models such as Poisson and negative binomial and their zero–inflated versions have been widely applied in many and different fields. However their application to biological dosimetry deserves special attention.

In biodosimetry, it is assumed that the mean of the number of aberrations is a linear or a quadratic function of the dose (IAEA, 2011). For sparsely ionising radiation there is very strong evidence that the mean yield of chromosome aberrations, $\mu_i$, is related to dose $x_i$ by the quadratic equation:

$$\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2,\ i = 1, \ldots, d, \tag{8.1}$$

whereas for densely ionising radiation, the larger relative amount of energy deposited (and the increase in the density of the ionisations which lead to the damage measured) results in an increase in the linear term and the quadratic term becomes biologically less relevant and so, the dose response may be approximated by a linear equation.

The linear quadratic model is used for low linear-energy-transfer (LET) radiations (i.e. gamma and X-rays) based on the justification that dicentric chromosome aberrations and micronuclei result from interactions between two independently damaged chromosomes (Hall *et al.*, 2012) and that the number of 'tracks' along which damage take place is linearly proportional to dose, so that the number of track (and thus damage) pairs is approximately proportional to dose squared (Hlatky *et al.*, 2002). For higher LET radiations, induction of chromosome aberrations becomes a linear function of dose because the more densely ionising nature of the radiation leads to a corresponding 'one track' distribution of damage. The same is true of fractionated or protracted doses, where there is time for repair of damage along one or more tracks between exposures.

Consequently, the link function used in (8.1) is simply the identity link function, as opposed to the log–link which is used for count data modelling in many other fields. The identity link is the accepted standard in biodosimetry since there is no evidence that the increase of aberration counts with dose is of exponential shape, and it avoids the undesired effect that dose–response curves start decreasing from about the maximum dose considered (IAEA, 2011). While we do not have strong arguments to change this standard, we point out that the log–link does have a few conceptual advantages, such as easier access to inferential tools for model testing, and the avoidance of problems with negative values of the linear predictor. In addition, using the log–link, that is,

$$\log(\mu_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2,\ i = 1, \ldots, d, \tag{8.2}$$

a simple second order approximation of $\mu_i$ can be directly obtained applying Taylor's formula at $x_i = 0$,

$$\mu_i = \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2) \approx a + b x_i + c x_i^2\ i = 1, \ldots, d,$$

with $a = \exp(\beta_0)$, $b = \exp(\beta_0)\beta_1$ and $c = \exp(\beta_0)(\beta_2 + \beta_1^2/2)$. Therefore, for low doses the results obtained using the identity–link or the log–link have to be very similar. Indeed, we will find in our detailed study in the next section that the results obtained for the two link functions are largely interchangeable.

A consequence of using the identity–link is that the maximum likelihood estimate of the parameter $\beta_0$ obtained by maximizing the log–likelihood function

of the corresponding model may be negative, i.e., may lead to a fitted negative control level, which makes no sense biologically. Therefore, in order to avoid negative values for the intercept, constraints in the domain of the parameters must be included when the model is fitted. Note that, though in some papers (e.g, Puig and Barquinero, 2011) the intercept is ignored in specific situations, it is well known that even when blood samples are not irradiated, the background level of aberrations could be positive (IAEA, 2011). In the absence of an intercept, the likelihood function at dose 0 (and, hence, the full data likelihood) would take the value zero, for which reason we would advocate the general use of an intercept in any model. Furthermore, since radiation protection practises are generally very good these days, most 'real life' cytogenetic dose estimates are likely to be in the region of zero.

A decision is required on which mean function is to be modelled: the mean of the zero-inflated distribution, $\mu_i$, or the mean of the underlying Poisson or negative binomial distribution, $\lambda_i$, which are related via $\lambda_i = \mu_i/(1-p_i)$. For compliance with formulation (1) and with practice in this particular field, we decided that it is adequate to model the mean of the corresponding zero–inflated distribution, $\mu_i$, via the linear predictor in (1). If no covariates are assumed for $p_i$, then this is equivalent to modelling $\lambda_i$ through a quadratic form.

The mixture parameter $p_i$ will be modelled as usual with logistic regression, where three different scenarios will be investigated: Firstly, it is assumed that the proportion of the mixture is constant:

$$\text{logit}(p_i) = \gamma_0,\ i = 1, \ldots, d, \tag{8.3}$$

secondly, $p_i$ is also modelled as a linear function of the dose:

$$\text{logit}(p_i) = \gamma_1 x_i,\ i = 1, \ldots, d, \tag{8.4}$$

and finally, $p_i$ is also modelled as a linear function of the dose but an intercept is included:

$$\text{logit}(p_i) = \gamma_0 + \gamma_1 x_i,\ i = 1, \ldots, d. \tag{8.5}$$

These different approaches will be applied on several data sets in Section 8.3.3, and further discussed in Section 8.3.4.

It should be noted that the zero–inflated Poisson distribution has been previously applied to estimate the mean yield of aberrations of the irradiated fraction of cells and the dose received by this fraction by a patient who has been exposed to an inhomogeneous irradiation. This methodology, proposed by Dolphin (1969), is known in biodosimetry as Dolphin's method or contaminated Poisson method (IAEA, 2011). However, this methodology does not constitute 'zero–inflated regression' from the viewpoint of modern statistical modelling, as outlined in this section. So, while the concept of zero–inflation is not completely new in this context, at the best of our knowledge, zero–inflated *regression* models have not been employed for the construction of dose–response curves, neither for partial nor whole body exposure scenarios.

## 8.3 Comparative study

In order to study the performance of zero–inflated models to describe the number of chromosome aberrations in biological dosimetry a substantial analysis

has been performed where these models are compared with those models already considered in the literature: Poisson, negative binomial, Neyman type A, Hermite, Pólya–Aeppli and Poisson–inverse Gaussian. The mathematical forms of these distributions are given as supplementary material in [70].

These models have been fitted following the "standards" given in Section 8.2.2 by using self–programmed code, which has been developed in the free software environment R (R Development Core Team, 2014). Function `maxLik` from package `maxLik` has been used in order to maximize the corresponding log–likelihood function. With the goal of facilitating the use of these techniques by practitioners, the function used for fitting the different models is available as supplementary material in [70] jointly with a detailed description of its usage and the datasets used in the study.

### 8.3.1 Scenarios: description of the data

The models have been applied to several real datasets obtained under four different scenarios: whole and partial body exposure with sparsely and densely ionising radiation. A brief description of them is given below.

(A) Whole body exposure – sparsely ionising radiation:

- (A1) These data consist of the frequency of dicentrics chromosomes after acute whole body *in vitro* exposure to eight uniform doses between 0 and 4.5 Gy of Cobalt–60 gamma rays (dose rate: 0.27 Gy/min). Blood was taken from fourteen healthy donors (six for the 0 Gy controls, and eight for the irradiated samples). Data were collected within the MULTIBIODOSE project and can be found in Table 6 of Romm *et al.* (2013).
- (A2) This dataset consists of scores of micronuclei obtained after irradiating eleven samples of peripheral blood with different doses (between 0 and 4 Gy) of gamma irradiation, where the dose rate was 0.93 cGy/min. In this case, for each sample, approximately 5000 binucleated cells were inspected and the numbers of micronuclei were counted. Data can be found in Table 6 of Puig and Valero (2006).
- (A3) Frequencies of dicentrics + centric rings aberrations are analysed in a total of 51600 metaphases from two volunteers after whole body exposure with 200 kV X-rays. Data considered here were obtained by scoring in metaphases reaching the first mitosis after a culture time of 56 h. Data can be found in Table 2 of Heimers *et al.* (2006).

(B) Whole body exposure – densely ionising radiation:

- (B1) This dataset corresponds to the number of dicentrics after exposure of peripheral blood samples to 10 different doses (between 0 and 1.6 Gy) of 1480 MeV oxygen ions. Data can be found in Table 2 of Di Giorgio *et al.* (2004) and was studied by Puig and Barquinero (2011).
- (B2) The second dataset considered in this scenario was obtained after irradiating blood samples with five different doses between 0.1 and 1 Gy of 2.1 MeV neutrons. In this case, frequencies of dicentrics + centric rings are analysed. Data are from Table 3 from Heimers *et al.* (2006) and corresponds to a culture time of 72 h.

Table 8.1: Doses, frequency distributions of the number of dicentrics, sample size and sum, and $u$-test values, for data set (B1).

| | $y_{ij}$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | $n_i$ | $y_i$ | $u_i$ |
| 0.000 | 1999 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2000 | 1 | 0 |
| 0.092 | 737 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 753 | 16 | -0.399 |
| 0.120 | 1438 | 55 | 5 | 2 | 0 | 0 | 0 | 0 | 1500 | 71 | 7.261 |
| 0.205 | 1300 | 104 | 14 | 2 | 0 | 0 | 0 | 0 | 1420 | 138 | 5.173 |
| 0.300 | 471 | 73 | 15 | 1 | 0 | 0 | 0 | 0 | 560 | 106 | 2.560 |
| 0.405 | 437 | 66 | 15 | 1 | 1 | 0 | 0 | 0 | 520 | 103 | 4.377 |
| 0.600 | 473 | 119 | 34 | 3 | 2 | 0 | 0 | 0 | 631 | 204 | 3.876 |
| 0.820 | 253 | 99 | 38 | 17 | 5 | 0 | 0 | 1 | 413 | 253 | 7.158 |
| 1.200 | 92 | 55 | 27 | 11 | 4 | 1 | 0 | 0 | 190 | 163 | 2.948 |
| 1.600 | 80 | 49 | 26 | 13 | 5 | 0 | 0 | 0 | 173 | 160 | 2.512 |

(C) Partial body exposure – sparsely ionising radiation:

(C1–C3) Three datasets were considered here. The scenario is the same as for dataset (A3) but, they correspond to partial body exposure simulation, with unirradiated blood mixed with irradiated blood from the same donors. The proportion of irradiated blood is 25%, 50% and 75%, respectively.

(D) Partial body exposure – densely ionising radiation:

(D1–D3) Finally, three datasets are considered in this scenario. Data were obtained by irradiating blood samples with 2.1 MeV neutrons (as in B2) and the same culture time is considered. The proportion of irradiated blood is 25%, 50% and 75%, respectively.

Quadratic dose models of type (8.1) and (8.2) will be used under sparsely ionising radiation, that is for data sets (A1) to (A3) and (C1) to (C3), and, following Puig and Barquinero (2011), also for data set (B1). Following the reasoning outlined in Section 8.2.2, the quadratic term will be removed for data sets (B2) and (D1) to (D3).

To illustrate the nature of the data, the full data set (B1) is displayed in Table 8.1 and visualized in image Figure 8.1. (Analogous tables and graphs for the remaining datasets are available in the supplementary material in [70].) Recall that we denote $y_i = \sum_{j=1}^{n_i} y_{ij}$ the total number of counts observed for dose $x_i$, that is, $y_i$ is the sufficient statistics to estimate the mean of the Poisson distribution under dose $x_i$. In the graphical representation, the circles have location $(x_i, y_i/n_i)$ and size $n_i$. The solid curve is the dose–response curve that would be fitted according to the Poisson model with identity link. However, consider the $u_i$ figures shown in Table 8.1 which are the values of the $u$-test statistic of Rao and Chakravarti (1956) to measure the overdispersion, suggested by IAEA 2011. Most of these $u$ values are $> 1.96$ (except for the control and the 0.092 Gy samples), rejecting in general the equidispersion assumption, thus the classical Poisson model is not appropriate for fitting this dataset.

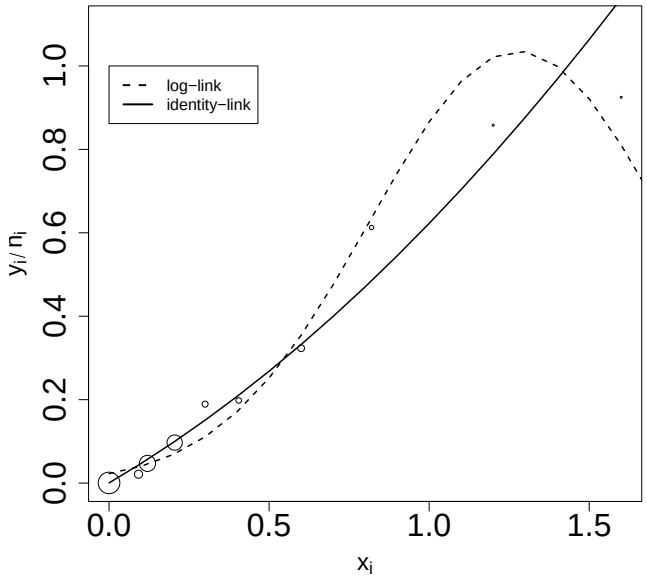


Figure 8.1: Dataset (B1): Proportions $y_i/n_i$ (symbolized by circles of radius $\propto n_i$) and dose-response curves fitted with Poisson model and two link functions.

### 8.3.2 Score tests

Before we provide our detailed overview of fitted models, we will give some more evidence for the presence of zero–inflation, and overdispersion on top of the zero–inflation, in the datasets introduced in Section 8.3.1. Score (or Rao) tests are a convenient tool for this purpose. A score test for testing a Poisson against a ZIP regression model was developed by van den Broek (1995). Similar score tests do exist for testing a Poisson against a negative binomial (NB) regression model (Dean and Lawless, 1989), as well as ZIP against a ZINB regression model (Ridout *et al.*, 2001). All these tests assume that constant probabilities (8.3) are employed. Furthermore, all these tests require that the mean is modelled through a log–link function. For the Poisson/ZIP case, we developed a variant of van den Broek's score test which also works under the identity link; see the appendix for details. As one can see from Table 8.2, the values of the test statistic are quite similar for the two link functions, and in any case lead to the same conclusions.

Similar adaptions of the score test for the identity link could be developed for the Poisson/NB and the ZIP/ZINB comparisons though this is beyond the scope of this paper. Hence, for these two latter situations, we constrain ourselves to log–link models when applying the score–test (our considerations in Section 8.2.2, as well as the results of the Poisson/ZIP test, suggest that this is not a serious restriction).

The values of the score test statistics for all considered datasets are given in Table 8.2. The values given in this table need to be compared with quantiles of

Table 8.2: Values of the score test statistic which, under the null hypothesis, has a $\chi^2$ distribution with one degree of freedom. The form of the (zero–inflated) negative binomial considered in each case is the one that provided the best fit according to the log–likelihood value in Tables 8.3 to 8.6. For tests involving zero–inflated models, the mixture parameter has been modelled according to (8.3).

| link | test | (A1) | (A2) | (A3) | (B1) | (B2) | (C1) | (C2) | (C3) | (D1) | (D2) | (D3) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| id | P/ZIP | 18.17 | 383.58 | 0.92 | 87.72 | 61.32 | 2007.39 | 1418.28 | 776.55 | 416.20 | 387.91 | 168.13 |
| | P/ZIP | 16.89 | 378.69 | 1.00 | 87.16 | 47.20 | 1996.30 | 1417.96 | 745.84 | 421.48 | 398.38 | 168.74 |
| log | P/NB | 20.79 | 1699.91 | 0.90 | 159.26 | 136.89 | 6009.35 | 3281.00 | 1210.34 | 770.62 | 693.80 | 285.61 |
| | ZIP/ZINB | 1.54 | 1043.94 | | 47.20 | 64.96 | 0.22 | 1.74 | 0.01 | 11.49 | 35.94 | 36.24 |

the chi–squared distribution with one degree of freedom; for instance at the 5% levels of significance this quantile takes the value 3.84. The higher the provided value of the test statistics, the stronger the evidence against the smaller model. This leads to the following conclusions:

- only for dataset (A3) — sparsely ionising whole body exposure — the assumption of a Poisson distribution cannot be rejected;
- for dataset (A1) — again, sparsely ionising whole body exposure — the Poisson assumption is rejected;
- for all datasets involving densely ionising radiation, that is (B) and (D), as well as for the micronuclei (A2), the Poisson model is rejected in favour of the ZIP and NB models, and furthermore the ZIP model is rejected in favour of the ZINB model.
- for all data sets involving partial body exposure, that is (C) — sparsely ionising — and (D) — densely ionising —, the Poisson assumption is rejected in favour of the ZIP and NB models.

It is worth noting that the current IAEA recommendation is for the Poisson distribution to be applied to all sparsely ionising data, testing for overdispersion then applying Dolphin's (1969) contaminated Poisson method or similar when Poisson assumptions are violated — which is expected in partial body exposure scenarios (IAEA, 2011). This recommendation is not entirely at odds with the result of our initial score tests, but is clearly too vague to be actually useful, so that cytogenists, in the absence of further guidance, tend to use the — apparently incorrect — Poisson assumption in most of the cases.

From the score test results one can further observe that, while for some datasets it will be sufficient to model either overdispersion or zero–inflation, for other datasets overdispersion appears to be separately present on top of the zero–inflation. We continue with a comprehensive analysis, fitting these and a variety of other related models, which will confirm these results.

### 8.3.3 Results and discussion

In order to compare the performance of the different models, classical likelihood measures of goodness of fit are used: The Akaike Information Criterion (AIC) and the Bayesian (Schwarz) Information Criterion (BIC). The AIC (Akaike 1974 [6]) penalizes a model with a larger number of parameters, and is defined

Table 8.3: Results of fitting various models to datasets (A1), (A2) and (A3), obtained under whole body exposure with sparsely ionising radiation. For each model, results obtained with identity–link (first row) and log–link (second row, *italic*) are shown.

| | | (A1) | | | (A2) | | | (A3) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Models | k | loglik | AIC | BIC | loglik | AIC | BIC | loglik | AIC | BIC |
| Poisson | 3 | -3748.59 | 7503.17 | 7526.14 | -34679.48 | 69364.95 | 69391.70 | -3806.89 | **7619.77** | **7638.94** |
| | 3 | -3749.36 | 7504.73 | 7527.70 | -34721.03 | 69448.07 | 69474.81 | -3808.27 | ***7622.55*** | ***7641.72*** |
| NB1 | 4 | -3742.82 | 7493.65 | 7524.28 | -34199.14 | 68406.28 | 68441.94 | -3806.89 | 7621.77 | 7647.33 |
| | 4 | -3743.69 | 7495.39 | 7526.02 | -34231.24 | 68470.47 | 68506.13 | -3808.28 | 7624.55 | 7650.11 |
| NB2 | 4 | -3739.23 | **7486.46** | **7517.09** | -34398.42 | 68804.84 | 68840.50 | -3806.98 | 7621.97 | 7647.53 |
| | 4 | -3740.55 | ***7489.10*** | ***7519.73*** | -34440.88 | 68889.76 | 68925.42 | -3808.28 | 7624.56 | 7650.12 |
| Neyman A | 4 | -3742.96 | 7493.91 | 7524.54 | -34214.13 | 68436.25 | 68471.91 | -3806.93 | 7621.86 | 7647.42 |
| | 4 | -3743.78 | 7495.57 | 7526.20 | -34246.64 | 68501.27 | 68536.93 | -3808.30 | 7624.60 | 7650.16 |
| Hermite (r=2) | 4 | -3743.09 | 7494.19 | 7524.82 | -34249.44 | 68506.88 | 68542.54 | -3806.89 | 7621.77 | 7647.33 |
| | 4 | -3743.89 | 7495.78 | 7526.41 | -34282.67 | 68573.35 | 68609.01 | -3808.30 | 7624.60 | 7650.16 |
| Polya-Aeppli | 4 | -3742.90 | 7493.79 | 7524.42 | -34204.54 | 68417.09 | 68452.75 | -3806.93 | 7621.86 | 7647.42 |
| | 4 | -3743.74 | 7495.47 | 7526.10 | -34236.76 | 68481.51 | 68517.17 | -3808.31 | 7624.63 | 7650.18 |
| PIG | 4 | -3742.75 | 7493.50 | 7524.13 | -34196.84 | 68401.69 | **68437.35** | -3806.91 | 7621.82 | 7647.38 |
| | 4 | -3743.62 | 7495.25 | 7525.88 | -34228.97 | 68465.94 | ***68501.60*** | -3808.28 | 7624.56 | 7650.12 |
| ZIP (3) | 4 | -3739.79 | 7487.58 | 7518.21 | -34490.47 | 68988.94 | 69024.60 | -3806.44 | 7620.87 | 7646.43 |
| | 4 | -3741.18 | 7490.36 | 7520.99 | -34534.06 | 69076.12 | 69111.78 | -3807.78 | 7623.57 | 7649.12 |
| ZIP (4) | 4 | -3741.26 | 7490.52 | 7521.15 | -34352.76 | 68713.53 | 68749.19 | -3806.89 | 7621.78 | 7647.34 |
| | 4 | -3742.85 | 7493.69 | 7524.32 | -34395.01 | 68798.02 | 68833.68 | -3808.28 | 7624.55 | 7650.11 |
| ZIP (5) | 5 | -3739.18 | 7488.36 | 7526.65 | -34266.33 | 68542.66 | 68587.23 | **-3806.21** | 7622.41 | 7654.36 |
| | 5 | -3740.19 | 7490.38 | 7528.67 | -34299.43 | 68608.87 | 68653.44 | ***-3807.55*** | 7625.11 | 7657.06 |
| ZINB1 (3) | 5 | -3739.69 | 7489.38 | 7527.67 | -34199.16 | 68408.31 | 68452.89 | -3806.45 | 7622.91 | 7654.85 |
| | 5 | -3740.72 | 7491.44 | 7529.73 | -34231.63 | 68473.26 | 68517.84 | -3808.19 | 7626.39 | 7658.33 |
| ZINB1 (4) | 5 | -3741.27 | 7492.53 | 7530.82 | **-34195.50** | **68400.99** | 68445.57 | -3807.24 | 7624.49 | 7656.43 |
| | 5 | -3742.81 | 7495.62 | 7533.90 | -34226.81 | 68463.62 | 68508.19 | -3808.38 | 7626.76 | 7658.71 |
| ZINB1 (5) | 6 | -3742.82 | 7497.65 | 7543.59 | -34195.73 | 68403.46 | 68456.96 | -3807.03 | 7626.06 | 7664.40 |
| | 6 | -3739.38 | 7490.75 | 7536.70 | ***-34224.79*** | ***68461.58*** | 68515.07 | -3808.31 | 7628.62 | 7666.95 |
| ZINB2 (3) | 5 | -3739.14 | 7488.27 | 7526.56 | -34398.60 | 68807.20 | 68851.78 | -3806.44 | 7622.87 | 7654.82 |
| | 5 | -3740.49 | 7490.98 | 7529.26 | -34440.92 | 68891.84 | 68936.41 | -3807.78 | 7625.57 | 7657.51 |
| ZINB2 (4) | 5 | -3739.08 | 7488.16 | 7526.45 | -34281.79 | 68573.58 | 68618.16 | -3806.89 | 7623.78 | 7655.73 |
| | 5 | -3740.37 | 7490.74 | 7529.03 | -34322.27 | 68654.54 | 68699.12 | -3815.15 | 7640.30 | 7672.25 |
| ZINB2 (5) | 6 | **-3738.15** | 7488.30 | 7534.25 | -34210.50 | 68433.00 | 68486.49 | -3806.21 | 7642.42 | 7662.75 |
| | 6 | ***-3739.25*** | 7490.50 | 7536.45 | -34242.98 | 68497.96 | 68551.45 | -3807.55 | 7627.11 | 7665.44 |

as AIC $= -2\log L + 2k$ , where $\log L$ denotes the fitted log–likelihood and $k$ the number of parameters. The BIC (Schwarz, 1978), defined as BIC $= -2\log L + k\log n$, works similarly to AIC but increases the penalty with increasing sample size $n$ (with our notation $n = \sum_{i=1}^{d} n_i$). According to these criteria, models with smaller values of AIC and BIC are considered preferable. It is standard practise to include both criteria in model fitting. Tables 8.3–8.6 show the results for each dataset considered, for both the identity link and the log–link (first and second row, respectively, for each given model). The value $k$ used for AIC and BIC is given explicitly in each table, and is computed as the sum of regression and model parameters. The best model in each column and for each link function is provided in bold face. Note that Hermite ($r$=1) is just Poisson, and that the results for higher–order Hermite models ($r$=3,4) are relegated to the supplementary material in [70].

**(A) Whole body exposure – sparsely ionising radiation**

Firstly, we observe from Table 8.3 that, as expected from the result of the score test, for dataset (A3) the Poisson model comes up as the preferred model under both the AIC and the BIC criterion. This corresponds to accepted practice for dicentrics under whole body exposure and sparsely ionising radiation.

However, for dataset (A1), the values of the maximized log–likelihood as well as the information criteria indicate that NB2 and zero–inflated models fit the data better than other models. Although the results are not shown here, a similar behavior has been observed for other datasets (e.g., for data correspond-

Table 8.4: Results of fitting various models to datasets (B1) and (B2), obtained under whole body exposure with densely ionising radiation. For each model, results obtained with identity–link (first row) and log–link (second row, *italic*) are shown. Separate columns for $k$ are provided for dataset (B1), which employs a quadratic model, and dataset (B2), which uses a linear predictor without quadratic term.

| | | (B1) | | | (B2) | | | |
|---|---|---|---|---|---|---|---|---|
| Models | k | loglik | AIC | BIC | loglik | AIC | BIC | k |
| Poisson | 3 | -2855.85 | 5717.70 | 5738.73 | -3004.72 | 6013.45 | 6026.57 | 2 |
| | 3 | -2904.50 | 5815.00 | 5836.02 | -3028.27 | 6060.54 | 6073.66 | 2 |
| NB1 | 4 | -2800.29 | 5608.58 | 5636.60 | -2960.17 | 5926.33 | 5946.02 | 3 |
| | 4 | -2846.15 | 5700.30 | 5728.33 | -2977.92 | 5961.83 | 5981.52 | 3 |
| NB2 | 4 | -2807.48 | 5622.96 | 5650.99 | -2976.16 | 5958.32 | 5978.00 | 3 |
| | 4 | -2856.61 | 5721.22 | 5749.25 | -2996.11 | 5998.22 | 6017.91 | 3 |
| Neyman A | 4 | -2799.74 | 5607.47 | **5635.50** | -2958.86 | **5923.72** | **5943.41** | 3 |
| | 4 | -2845.21 | 5698.41 | 5726.44 | -2976.94 | ***5959.88*** | ***5979.57*** | 3 |
| Hermite (r=2) | 4 | -2802.15 | 5612.30 | 5640.32 | -2959.02 | 5924.03 | 5943.72 | 3 |
| | 4 | -2847.60 | 5703.19 | 5731.22 | -2978.22 | 5962.44 | 5982.13 | 3 |
| Polya-Aeppli | 4 | -2799.81 | 5607.61 | 5635.64 | -2959.48 | 5924.96 | 5944.65 | 3 |
| | 4 | -2845.48 | 5698.97 | 5727.00 | -2977.25 | 5960.50 | 5980.19 | 3 |
| PIG | 4 | -2801.91 | 5611.81 | 5639.84 | -2961.98 | 5929.97 | 5949.66 | 3 |
| | 4 | -2848.04 | 5704.08 | 5732.11 | -2979.74 | 5965.48 | 5985.17 | 3 |
| ZIP (3) | 4 | -2814.53 | 5637.07 | 5665.09 | -2979.06 | 5964.13 | 5983.82 | 3 |
| | 4 | -2861.85 | 5731.69 | 5759.72 | -3005.82 | 6017.64 | 6037.33 | 3 |
| ZIP (4) | 4 | -2805.36 | 5618.71 | 5646.74 | -2967.53 | 5941.05 | 5960.74 | 3 |
| | 4 | -2854.06 | 5716.12 | 5744.15 | -2990.43 | 5986.87 | 6006.56 | 3 |
| ZIP (5) | 5 | -2800.58 | 5611.17 | 5646.20 | -2958.35 | 5924.71 | 5950.96 | 4 |
| | 5 | -2847.77 | 5705.53 | 5740.57 | -2977.43 | 5962.86 | 5989.12 | 4 |
| ZINB1 (3) | 5 | -2797.41 | 5604.82 | 5639.85 | -2960.81 | 5929.62 | 5955.87 | 4 |
| | 5 | -2842.31 | 5694.63 | 5729.66 | -2977.92 | 5963.84 | 5990.09 | 4 |
| ZINB1 (4) | 5 | **-2797.30** | **5604.61** | 5639.64 | -2958.76 | 5925.52 | 5951.77 | 4 |
| | 5 | -2842.34 | 5694.68 | 5729.72 | -2976.85 | 5961.70 | 5987.95 | 4 |
| ZINB1 (5) | 6 | -2797.33 | 5606.67 | 5648.71 | -2957.40 | 5924.79 | 5957.61 | 5 |
| | 6 | -2842.04 | 5696.07 | 5738.11 | **-2975.95** | 5961.90 | 5994.71 | 5 |
| ZINB2 (3) | 5 | -2807.47 | 5624.93 | 5659.97 | -2976.38 | 5960.76 | 5987.01 | 4 |
| | 5 | -2856.41 | 5722.82 | 5757.86 | -2996.13 | 6000.26 | 6026.51 | 4 |
| ZINB2 (4) | 5 | -2800.06 | 5610.13 | 5645.16 | -2964.11 | 5936.22 | 5962.47 | 4 |
| | 5 | -2847.84 | 5705.68 | 5740.71 | -2984.50 | 5976.99 | 6003.24 | 4 |
| ZINB2 (5) | 6 | -2798.59 | 5609.17 | 5651.22 | **-2957.30** | 5924.60 | 5957.41 | 5 |
| | 6 | ***-2809.61*** | ***5631.21*** | ***5673.25*** | -2976.29 | 5962.58 | 5995.40 | 5 |

ing to lab 3 shown in Table 3 in Romm, 2003) obtained using an automatic scoring procedure. In this case, one could speculate that the automatic scoring procedure used for (A1) may skew the data away from Poisson. However, more datasets would be needed to demonstrate such an effect reliably.

For data (A2), the Poisson distribution does not provide a good fit (see Table 8.3). In this case, it should be pointed out that micronuclei counts differ from dicentrics in that i) the quadratic component of the dose dependence is frequently weaker (for sparsely irradiation), ii) baseline counts of unirradiated samples are much higher than for dicentrics and iii) even after uniform total body irradiation micronucleus distributions tend to be overdispersed.

Therefore, although for whole–body exposure and sparsely ionising radiation, it is usually assumed that data follow a Poisson model, data under this scenario may depart from the Poisson model due to other circumstances (e.g., the scoring procedure).

Table 8.5: Results of fitting various models to datasets (C1), (C2) and (C3), obtained under partial body exposure with sparsely ionising radiation. For each model, results obtained with identity–link (first row) and log–link (second row, *italic*) are shown.

| | | (C1) | | | (C2) | | | (C3) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Models | k | loglik | AIC | BIC | loglik | AIC | BIC | loglik | AIC | BIC |
| Poisson | 3 | -2674.93 | 5355.86 | 5376.50 | -3526.90 | 7059.81 | 7079.70 | -3472.24 | 6950.47 | 6969.64 |
| | 3 | -2676.09 | 5358.18 | 5378.83 | -3528.70 | 7063.39 | 7083.28 | -3468.15 | 6942.30 | 6961.46 |
| NB1 | 4 | -2090.11 | 4188.21 | 4215.74 | -3011.85 | 6031.70 | 6058.23 | -3229.20 | 6466.40 | 6491.95 |
| | 4 | -2091.83 | 4191.65 | 4219.18 | -3011.69 | 6031.38 | 6057.90 | -3224.49 | 6456.98 | 6482.54 |
| NB2 | 4 | -2088.53 | 4185.07 | 4212.59 | -2939.48 | 5886.97 | 5913.49 | -3155.36 | 6318.71 | 6344.27 |
| | 4 | -2052.98 | 4113.96 | 4141.48 | -2940.52 | 5889.05 | 5915.57 | -3153.54 | 6315.08 | 6340.64 |
| Neyman A | 4 | -2103.10 | 4214.20 | 4241.73 | -3021.07 | 6050.13 | 6076.66 | -3232.00 | 6471.99 | 6497.55 |
| | 4 | -2104.75 | 4217.50 | 4245.03 | -3022.38 | 6052.76 | 6079.28 | -3229.16 | 6466.31 | 6491.87 |
| Hermite (r=2) | 4 | -2248.89 | 4505.77 | 4533.30 | -3122.52 | 6253.03 | 6279.56 | -3265.83 | 6539.65 | 6565.21 |
| | 4 | -2249.82 | 4507.64 | 4535.17 | -3123.51 | 6255.01 | 6281.53 | -3263.19 | 6534.39 | 6559.95 |
| Polya-Aeppli | 4 | -2087.21 | 4182.42 | 4209.94 | -3007.02 | 6022.04 | 6048.56 | -3227.37 | 6462.75 | 6488.31 |
| | 4 | -2088.91 | 4185.82 | 4213.35 | -3007.94 | 6023.89 | 6050.41 | -3223.57 | 6455.14 | 6480.69 |
| PIG | 4 | -2109.59 | 4227.19 | 4254.72 | -3035.89 | 6079.79 | 6106.31 | -3241.24 | 6490.47 | 6516.03 |
| | 4 | -2111.35 | 4230.69 | 4258.22 | -3035.98 | 6079.96 | 6106.48 | -3235.40 | 6478.81 | 6504.37 |
| ZIP (3) | 4 | -2010.84 | 4029.68 | ***4057.21*** | -2852.63 | 5713.26 | 5739.79 | -3092.40 | 6192.79 | 6218.35 |
| | 4 | -2010.76 | 4029.53 | ***4057.05*** | -2852.29 | 5712.59 | 5739.11 | -3092.73 | 6193.46 | 6219.02 |
| ZIP (4) | 4 | -2034.75 | 4077.51 | 4105.03 | -2844.89 | 5697.77 | **5724.29** | -3087.70 | 6183.41 | ***6208.97*** |
| | 4 | -2026.52 | 4061.05 | 4088.58 | -2845.72 | 5699.44 | ***5725.96*** | -3086.79 | 6181.58 | 6207.14 |
| ZIP (5) | 5 | -2007.01 | ***4024.02*** | 4058.43 | **-2842.39** | **5694.79** | 5727.94 | -3085.56 | ***6181.13*** | 6213.08 |
| | 5 | -2006.57 | ***4023.13*** | 4057.54 | ***-2843.70*** | ***5697.40*** | 5730.55 | -3081.33 | ***6172.66*** | ***6204.61*** |
| ZINB1 (3) | 5 | -2010.85 | 4031.70 | 4066.10 | -2852.65 | 5715.31 | 5748.46 | -3092.45 | 6194.91 | 6226.85 |
| | 5 | -2010.78 | 4031.55 | 4065.96 | -2852.31 | 5714.61 | 5747.77 | -3092.75 | 6195.50 | 6227.44 |
| ZINB1 (4) | 5 | -2017.66 | 4045.31 | 4079.72 | -2844.21 | 5698.43 | 5731.58 | -3087.84 | 6185.67 | 6217.62 |
| | 5 | -2015.63 | 4041.25 | 4075.66 | -2845.06 | 5700.13 | 5733.28 | -3086.80 | 6183.59 | 6215.54 |
| ZINB1 (5) | 6 | -2006.97 | 4025.94 | 4067.23 | -2842.41 | 5696.81 | 5736.60 | -3085.55 | 6183.10 | 6221.44 |
| | 6 | -2006.50 | 4024.99 | 4066.28 | -2843.71 | 5699.42 | 5739.20 | -3080.99 | 6173.97 | 6212.31 |
| ZINB2 (3) | 5 | -2010.70 | 4031.40 | 4065.81 | -2852.63 | 5715.26 | 5748.42 | -3092.37 | 6194.74 | 6226.68 |
| | 5 | -2010.66 | 4031.31 | 4065.72 | -2852.29 | 5714.59 | 5747.74 | -3092.73 | 6195.46 | 6227.40 |
| ZINB2 (4) | 5 | -2022.00 | 4054.00 | 4088.41 | -2844.88 | 5699.75 | 5732.90 | -3087.68 | 6185.37 | 6217.32 |
| | 5 | -2021.05 | 4052.10 | 4086.50 | -2845.72 | 5701.44 | 5734.59 | -3086.79 | 6183.58 | 6215.53 |
| ZINB2 (5) | 6 | ***-2006.47*** | 4024.93 | 4066.22 | -2842.42 | 5696.83 | 5736.62 | ***-3084.93*** | 6181.87 | 6220.20 |
| | 6 | ***-2006.20*** | 4024.41 | 4065.70 | -2843.70 | 5699.40 | 5739.18 | ***-3080.94*** | 6173.87 | 6212.21 |

**(B) Whole body exposure – densely ionising radiation**

For the two datasets in this scenario values for the Poisson regression model are clearly worse than for the other models, confirming the overdispersion reported for several authors concerning high LET radiation exposures. According to the results shown in Table 8.4, there are several models which are very competitive. In this case, it seems that overdispersion can be modelled through different models, including the Neyman A and zero–inflated negative binomial models.

**(C) Partial body exposure – sparsely ionising radiation**

For datasets considered in this scenario (C1–C3), zero–inflated models are notably better than the other models as shown in Table 8.5. This result is in line with the philosophy of Dolphin's method (Dolphin, 1969). The zero–inflated Poisson models perform consistently well for all three datasets, and the information criteria give little support for (possibly zero–inflated) negative binomial models. Hence, for this type of datasets, it seems clear that overdispersion is due to the excess of zeros.

**(D) Partial body exposure – densely ionising radiation**

For datasets in this scenario (D1–D3), the Poisson model is clearly rejected. From Table 8.6, it can be observed that, in general, the ZINB models provide the best fits which indicates that overdispersion is due to both the excess of zeroes (caused by the partial body exposure) and the heterogeneity (caused by the densely ionising radiation). However, there is quite a wide range of models

Table 8.6: Results of fitting various models to datasets (D1), (D2) and (D3), obtained under partial body exposure with densely ionising radiation. For each model, results obtained with identity–link (first row) and log–link (second row, *italic*) are shown.

| | | (D1) | | | (D2) | | | (D3) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Models | k | loglik | AIC | BIC | loglik | AIC | BIC | loglik | AIC | BIC |
| Poisson | 2 | -1477.95 | 2959.89 | 2973.54 | -2302.09 | 4608.18 | 4621.51 | -2394.99 | 4793.98 | 4806.93 |
| | 2 | -1482.50 | 2969.00 | 2982.65 | -2323.30 | 4650.60 | 4663.94 | -2415.08 | 4834.17 | 4847.12 |
| NB1 | 3 | -1370.07 | 2746.13 | 2766.61 | -2148.66 | 4303.31 | 4323.31 | -2310.45 | 4626.89 | 4646.32 |
| | 3 | -1373.15 | 2752.30 | 2772.77 | -2163.63 | 4333.26 | 4353.26 | -2326.58 | 4659.17 | 4678.59 |
| NB2 | 3 | -1366.28 | **2738.57** | **2759.04** | -2151.63 | 4309.26 | 4329.26 | -2322.30 | 4650.59 | 4670.02 |
| | 3 | -1370.12 | ***2746.25*** | ***2766.72*** | -2167.41 | 4340.81 | 4360.81 | -2337.16 | 4680.31 | 4699.74 |
| Neyman A | 3 | -1372.33 | 2750.65 | 2771.13 | -2146.95 | 4299.91 | 4319.91 | -2306.20 | **4618.39** | **4637.82** |
| | 3 | -1375.41 | 2756.81 | 2777.29 | -2161.79 | 4329.58 | 4349.58 | -2322.23 | ***4650.47*** | ***4669.89*** |
| Hermite (r=2) | 3 | -1382.18 | 2770.37 | 2790.84 | -2164.80 | 4335.59 | 4355.59 | -2308.57 | 4623.14 | 4642.56 |
| | 3 | -1385.58 | 2777.16 | 2797.63 | -2180.55 | 4367.10 | 4387.10 | -2325.06 | 4656.12 | 4675.55 |
| Polya-Aeppli | 3 | -1370.34 | 2746.67 | 2767.15 | -2146.66 | 4299.32 | **4319.32** | -2308.06 | 4622.11 | 4641.54 |
| | 3 | -1373.41 | 2752.83 | 2773.30 | -2161.53 | 4329.06 | ***4349.06*** | -2324.10 | 4654.20 | 4673.63 |
| PIG | 3 | -1371.72 | 2749.43 | 2769.90 | -2155.04 | 4316.09 | 4336.08 | -2315.55 | 4637.11 | 4656.54 |
| | 3 | -1374.81 | 2755.63 | 2776.10 | -2170.28 | 4346.57 | 4366.57 | -2331.96 | 4669.93 | 4689.36 |
| ZIP (3) | 3 | -1369.48 | 2744.96 | 2765.43 | -2155.15 | 4316.30 | 4336.30 | -2322.37 | 4650.73 | 4670.16 |
| | 3 | -1373.58 | 2753.17 | 2773.64 | -2173.27 | 4352.53 | 4372.53 | -2341.29 | 4688.58 | 4708.01 |
| ZIP (4) | 3 | -1386.57 | 2779.15 | 2799.62 | -2172.81 | 4351.62 | 4371.61 | -2323.33 | 4652.66 | 4672.09 |
| | 3 | -1391.84 | 2789.68 | 2810.16 | -2193.54 | 4393.07 | 4413.07 | -2341.86 | 4689.72 | 4709.15 |
| ZIP (5) | 4 | -1368.96 | 2745.91 | 2773.21 | -2147.03 | 4302.07 | 4328.73 | -2308.05 | 4624.11 | 4650.01 |
| | 4 | -1372.62 | 2753.25 | 2780.55 | -2160.67 | 4329.34 | 4356.00 | -2321.58 | 4651.15 | 4677.06 |
| ZINB1 (3) | 4 | -1366.48 | 2740.96 | 2768.26 | **-2143.46** | **4294.93** | 4321.59 | -2308.53 | 4625.06 | 4650.97 |
| | 4 | -1369.76 | 2747.53 | 2774.83 | -2158.76 | ***4325.53*** | 4352.19 | -2324.87 | 4657.73 | 4683.64 |
| ZINB1 (4) | 4 | -1366.16 | 2740.32 | 2767.62 | -2143.59 | 4295.19 | 4321.85 | -2307.72 | 4623.44 | 4649.35 |
| | 4 | -1373.16 | 2754.32 | 2781.61 | -2158.79 | 4325.59 | 4352.25 | -2323.98 | 4655.97 | 4681.87 |
| ZINB1 (5) | 5 | -1366.13 | 2742.27 | 2776.39 | -2143.40 | 4296.80 | 4330.13 | -2306.96 | 4623.93 | 4656.31 |
| | 5 | ***-1369.22*** | 2748.45 | 2782.57 | ***-2158.67*** | 4327.34 | 4360.66 | -2321.48 | 4652.96 | 4685.35 |
| ZINB2 (3) | 4 | -1366.05 | 2740.09 | 2767.39 | -2150.67 | 4309.35 | 4336.01 | -2320.61 | 4649.22 | 4675.13 |
| | 4 | -1369.93 | 2747.85 | 2775.15 | -2166.94 | 4341.88 | 4368.54 | -2336.76 | 4681.52 | 4707.42 |
| ZINB2 (4) | 4 | -1366.44 | 2740.87 | 2768.17 | -2147.31 | 4302.62 | 4329.28 | -2313.89 | 4635.79 | 4661.69 |
| | 4 | -1369.97 | 2747.94 | 2775.23 | -2162.26 | 4332.53 | 4359.19 | -2328.61 | 4665.22 | 4691.12 |
| ZINB2 (5) | 5 | **-1365.88** | 2741.77 | 2775.89 | -2144.92 | 4299.85 | 4333.18 | **-2307.69** | 4625.38 | 4657.76 |
| | 5 | -1369.66 | 2749.32 | 2783.45 | -2158.93 | 4327.85 | 4361.18 | ***-2321.34*** | 4652.68 | 4685.07 |

which provided competitive results for some data sets under this scenario, among them NB2, Polya-Aeppli, and the Neyman type A model. The latter has been shown to perform well for densely ionising radiation by Virsik and Harder (1981).

In our analysis, the Poisson model provides the (by far) worst fit for almost all datasets, including the sparsely ionising scenarios. Thus, a Poisson model should be used only in cases where there is strong evidence that it is the correct specification. In any case, it is clear that the Poisson model will be inadequate under partial body exposure and/or for densely ionising radiation. In general, as compared to the Poisson model, the proposed zero–inflated regression models perform well in terms of log–likelihood and the model selection criteria employed, for both full and partial body exposure.

The wide range of model classes considered so far does not make the claim to be exhaustive, and there are further modelling strategies which deserve consideration. We investigated random effect models which effectively add a random intercept term to the linear predictor (8.1) or (8.2). Considering the responses as repeated measures $y_{ij}$ equipped with a two–level structure, the random effect is added to the upper (aggregated) data level, $i$, effectively imposing correlation within blood samples. While correlation between cells from the same blood sample is a reasonable assumption, we note that each blood sample got exposed to a different dose, which is included as covariate into the model. We certainly would expect the dose effect to be much larger than any possible within–sample correlation. Hence, we do not consider this approach as a truly hierarchical ('variance component') model, but rather as a simple overdispersion model. We

fitted random effect models using a Gaussian random effect with Poisson and negative binomial response distribution, and, for the former, also considered an unspecified random effect distribution. It was found that, using the log–link, these models indeed work well for some specific data sets such as (A1) and (D2) (see the supplementary material in [70] for detailed results). The identity–link version, for which we found a workable implementation only for the Poisson model with Gaussian random effect, is more difficult to use since the random effect can render the linear predictor negative, which is incompatible with its interpretation as a Poisson mean.

Of course, random effect models will show their actual power only in truly hierarchical setups, where they can be used to model inter–individual correlation rather than just overdispersion. To our knowledge, the first work in this direction has been produced by Mano and Suto (2014), using a Bayesian framework. None of the datasets that we have investigated does provide such hierarchical information, so we did not investigate this avenue further.

A second model class which should be mentioned here are two–part models, which, rather than allowing zeros to be generated via two different routes as in the ZIP model, define a separate model for zero– and non–zero response, where the latter part could be described by e.g. a truncated Poisson distribution (Alfò and Maruotti 2013 [7]). Such 'Hurdle' models have the appealing property of being based on a clear hierarchical structure: first, a decision is made on whether a zero is chosen or not, and secondly, the non–zero part of the model is invoked if chosen. These models, which are beyond the scope of the present manuscript, appear promising in the context of radiation biodosimetry and so deserve further investigation.

### 8.3.4 Discussion on how to model the zero–inflation parameter

Based on the results shown in Tables 8.3–8.6, the three considered forms of modelling the zero–inflation parameter $p_i$ provide similar results in terms of the log–likelihood. However, looking at the fitted values of this parameter, it can be observed that they can be very different depending on the specified model.

Figure 8.2 shows the fitted values of the parameter $p_i$ after fitting a ZIP regression model to data (C1–C3) and (A3) (left panel) and a ZINB1 regression model to data (D1–D3) and (B2) (right panel). The solid dots represent the fitted $p_i$ when these do not depend on covariates, and the dashed and solid lines give the fitted values when $p_i$ is modelled through a logit link as a linear function of the dose with and without intercept, respectively.

Both plots show that the mixture parameter takes similar values at the highest doses observed in each case, independently of how it is modelled. Moreover, the value of $p_i$ is influenced by the percentage of unirradiated blood, as expected (IAEA, 2011). However, for the lowest doses, it takes very different values. If $p_i$ is modelled as $\text{logit}(p_i) = \gamma_1 x_i$ $(i = 1, \ldots, d)$, then the mixture parameter is equal to 0.5 at zero dose. That is, the model (8.4) imposes the probability 0.5 of extra zeros for non–irradiation. However, this may be a very restrictive assumption. In order to allow for more flexibility, an intercept is included in model (8.5). If $\text{logit}(p_i) = \gamma_0 + \gamma_1 x_i$, different situations can occur. For example, the dashed lines in the right panel in Figure 8.2 show that for non–irradiated blood samples, the probability of extra zeros is quite similar for the four datasets (as

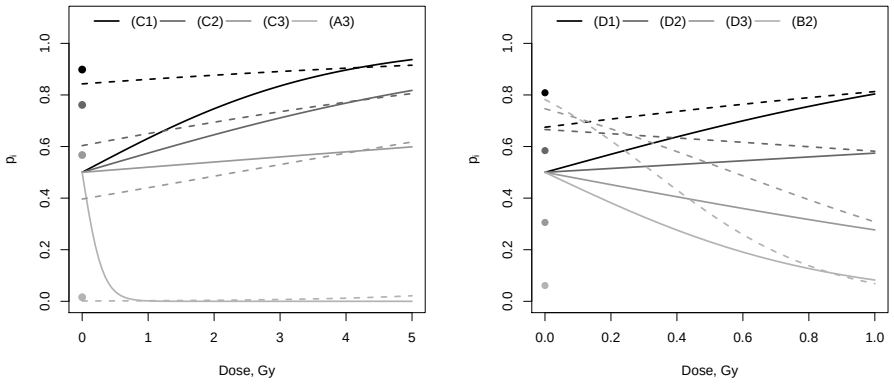


Figure 8.2: Fitted zero–inflation (mixture) parameters $p_i$ as a function of dose, $x_i$. Solid lines correspond to modelling the mixture parameter as $\text{logit}(p_i) = \gamma_1 x_i$ and dashed lines correspond to modelling it as $\text{logit}(p_i) = \gamma_0 + \gamma_1 x_i$. Solid dots indicate the fitted probabilities when $p_i$ is modelled as a constant, $\text{logit}(p_i) = \gamma_0$. Left panel: Results obtained from fitting a ZIP regression model to data (A3) and (C1–C3). Right panel: Results obtained from fitting a ZINB1 regression model to data (B2) and (D1–D3).

it would be expected). But, the dashed lines in the left panel show that the probability takes very different values at dose 0. This different behaviour may be explained by the first dose observed in each case. For data in the right panel, the smallest dose used was 0.1 Gy so, it is expected that the four datasets perform similarly (at dose 0, the four datasets should be practically equal). In the other hand, for data in the left panel, the smallest dose was 1 Gy and so, the value of $p_i$ is already influenced by the percentage of irradiated blood.

Model (8.5) is especially meaningful for fitting (C1–C3) and (D1–D3). In a partial body exposure simulation experiment, where a fixed proportion of blood $f$ is irradiated (for instance, 25%, 50% and 75%) to a dose $x$, this proportion $f$ is not the same as the proportion $(1-p)$ of irradiated cells in the zero–inflated model. Moreover, the magnitude of the difference also depends on the dose. The reason is that not all the irradiated cells transform and survive to metaphase, and those which do not survive can not be scored. According to Lloyd et al. (1973), the survival rate of the irradiated cells $s(x)$ follows a decreasing exponential function of the dose $x$ of the form, $s(x) = \exp(-\gamma_1 x)$. Note that for a dose of $x = 0$ the survival rate is 100%.

Suppose that in a partial body exposure we have $N$ irradiated cells and $N_0$ non irradiated cells. It is clear that the proportion of irradiated blood is,

$$f = \frac{N}{N_0 + N}\,, \tag{8.6}$$

but the proportion of scored irradiated cells (those which survive) is,

$$(1-p) = \frac{Ns(x)}{N_0 + Ns(x)}\,. \tag{8.7}$$

Replacing $N_0$ in (8.7) by that isolated from (8.6), we obtain,

$$(1-p) = \frac{Ns(x)}{N(1-f)/f + Ns(x)} = \frac{\exp(-\gamma_1 x)}{(1-f)/f + \exp(-\gamma_1 x)} = \frac{1}{1+\exp(\gamma_0+\gamma_1 x)}\,, \tag{8.8}$$

where $\text{logit}(f) = -\gamma_0$. This implies the relationship $\text{logit}(p) = \gamma_0 + \gamma_1 x$, justifying model (8.5).

The value of $\gamma_1$ depends on the kind or radiation and its capacity to damage the cells, and $\gamma_0$ is related to the fraction of irradiated blood.

Our application studies have demonstrated little difference in terms of log–likelihood between the three methods of modelling the mixture parameter. However, for partial body irradiation scenarios, we have shown that model (8.5) is conceptually preferable. Dataset (C2) constitutes an example where this conceptual advantage led to a superior practical performance.

## 8.4 Simulation study

In this section, we will give some more objective evidence for our claim that overdispersion and zero–inflation are in general separately identifiable. If that is true, we would expect

- the ZINB model to be favorable if both of these features are present;
- the NB model to be favorable if only overdispersion is present;
- the ZIP model to be favorable if only zero–inflation is present;
- the Poisson model to be favorable if none of these features are present.

Therefore, we generated 100 data sets from each of Poisson, ZIP, NB2 and ZINB2 models, then we fitted the data using all four models, and counted the proportion of times that each model gives the 'winning result' in terms of AIC and BIC. We also computed the score tests introduced earlier (where applicable) and give the proportions of rejection of the respective null hypothesis. For the data generation, we used the Poisson model fitted to (A3) as base model (we know from our previous analysis that this is a 'correct' model), with five doses $x_1 = 1, \ldots, x_5 = 5$. Then we instilled successively zero–inflation and overdispersion into this data-generating process, and observed the outcome. For the zero–inflation parameter $p_i$, we assumed scenario (8.3); that is, we did not assume dependence of this parameter on dose.

For the simulation of the ZIP models, we distinguished between (a) mild zero–inflation [$p = 0.1$], (b) moderate zero–inflation [$p = 0.2$] and (c) strong zero–inflation [$p = 0.5$]. For the NB models, we tried to match the degree of 'non-Poissonness" according to the following reasoning. Note that, for ZIP models, one has

$$\text{Var}(Y_i) = \mu_i(1+p\lambda_i) = \mu_i\left(1+\frac{p}{1-p}\mu_i\right).$$

For the negative binomial model (NB2), we know that

$$\text{Var}(Y_i) = \mu_i(1+\alpha\mu_i);$$

Table 8.7: Proportion of correctly identified models using AIC, for models using the identity–link (top) and log–link (bottom). The 'correct' model choice is provided in bold letters. Columns add up to 100%.

| | True model | P | ZIP | | | NB | | | ZINB | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| link | Fitted model | | mild | mod. | strong | mild | mod. | strong | mild | mod. | strong |
| | P | **91** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| id | ZIP | 6 | **88** | **96** | **96** | 5 | 0 | 0 | 8 | 0 | 0 |
| | NB | 3 | 0 | 0 | 0 | **92** | **91** | **90** | 3 | 0 | 2 |
| | ZINB | 0 | 12 | 4 | 4 | 3 | 9 | 10 | **89** | **100** | **98** |
| | P | **95** | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| log | ZIP | 3 | **50** | **67** | **96** | 42 | 41 | 7 | 9 | 0 | 0 |
| | NB | 2 | 0 | 0 | 0 | **51** | **54** | **85** | 7 | 1 | 1 |
| | ZINB | 0 | 50 | 33 | 4 | 5 | 5 | 8 | **84** | **99** | **99** |

Table 8.8: Proportion of correctly identified models using BIC, for models using the identity–link. The 'correct' model choice is provided in bold letters. Columns add up to 100%.

| | True model | P | ZIP | | | NB | | | ZINB | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| link | Fitted model | | mild | mod. | strong | mild | mod. | strong | mild | mod. | strong |
| | P | **100** | 0 | 0 | 0 | 13 | 0 | 0 | 0 | 0 | 0 |
| id | ZIP | 0 | **95** | **100** | **100** | 6 | 0 | 0 | 45 | 1 | 0 |
| | NB | 0 | 2 | 0 | 0 | **81** | **100** | **100** | 48 | 25 | 17 |
| | ZINB | 0 | 3 | 0 | 0 | 0 | 0 | 0 | **7** | **74** | **83** |
| | P | **100** | 1 | 0 | 0 | 23 | 0 | 0 | 0 | 0 | 0 |
| log | ZIP | 0 | **52** | **68** | **100** | 22 | 41 | 7 | 51 | 1 | 0 |
| | NB | 0 | 11 | 0 | 0 | **55** | **59** | **93** | 39 | 14 | 15 |
| | ZINB | 0 | 36 | 32 | 0 | 0 | 0 | 0 | **10** | **85** | **85** |

hence, for an equal degree of non-Poissonness we can equate $\alpha = p/(1-p)$. Following this reasoning, we considered in our simulation study data generated from a NB2 distribution with parameters (a) $\alpha = 1/9$ (mild overdispersion), (b) $\alpha = 1/4$ (moderate overdispersion) and (c) $\alpha = 1$ (strong overdispersion). For the ZINB models, we considered the pairings (a) mild/mild, (b) moderate/moderate and (c) strong/strong.

Tables 8.7 and 8.8 indicate clearly that, in the vast majority of cases, the underlying models were correctly identified. For the log–link, we observed a tendency of mildly zero–inflated Poisson models to be classified as ZINB models, and mildly overdispersed (NB) models to be classified as ZIP models. The stronger the overdispersion or zero–inflation, the better are the associated models separately identifiable. The proportion of correctly identified models was generally larger for the identity– than for the log–link, and was generally larger when using BIC rather than AIC. The only exception to this are 'mild/mild' zero–inflated negative binomial models, which tend to be classified as ZIP or NB models under BIC. The score tests in Table 8.9 speak a very clear language: The proportion of rejection of the Poisson and ZIP model is close to 0, when these models are true, and close or equal to 1, when these are false. Overall these simulations confirm impressively the separate dependability of zero–inflated and overdispersed models, as well as the need for models which are both overdispersed and zero–inflated.

Table 8.9: Proportion of rejection of the smaller model using score tests (at the 5% level of significance), for models using the identity link (top row) and the log–link (bottom three rows). Only values in bold are fully meaningful as in this case the true model corresponds to one of the two models tested against.

| | True model | P | ZIP | | | NB | | | ZINB | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| link | Fitted model | | mild | mod. | strong | mild | mod. | strong | mild | mod. | strong |
| id | P/ZIP | **0.05** | **1** | **1** | **1** | 0.86 | 1 | 1 | 1 | 1 | 1 |
| log | P/ZIP | **0.02** | **1** | **1** | **1** | 0.90 | 1 | 1 | 1 | 1 | 1 |
| | P/NB | **0.03** | 0.99 | 1 | 1 | **0.98** | **1** | **1** | 1 | 1 | 1 |
| | ZIP/ZINB | 0.00 | **0.03** | **0.01** | **0.05** | 0.84 | 1 | 1 | **0.77** | **1** | **1** |

## 8.5 Concluding remarks

Zero–inflated models have been proposed for modelling the number of aberrations per cell as a function of the dose. They have been compared with other models showing that they behave well in several scenarios, especially for partial body exposure. Moreover, results obtained by modelling the mean yield of aberrations through a log–link were compared with these ones obtained by using the identity link showing that both link functions give very similar results. Score tests justified the use of zero–inflated models for fitting several datasets. For the problem of testing a Poisson versus a ZIP model, we have presented in this manuscript a variant of van den Broek's score test which allows for the use of the identity link.

A relevant finding of this paper is that overdispersion needs to be taken into account irrespective of whether the data stem from full or partial body exposure. In the case of full body exposure, for densely ionising radiation or when micronuclei are analysed, the overdispersion will be relatively high and can often be addressed through a (possibly zero–inflated) negative binomial model or the Neyman A model, whereas for sparsely ionising radiation overdispersion will be relatively mild (but not always ignorable) and can often be addressed exchangeably through a negative binomial or a zero–inflated Poisson model (or even other models). Partial body exposure will in general require explicit modelling of the zero-inflation. While for sparsely ionising radiation zero–inflated Poisson models turned out to be sufficient in our analysis, for densely ionising radiation it was generally necessary to model the overdispersion *on top of* the zero–inflation, through a zero–inflated negative binomial model. A small simulation study has confirmed that the concept of considering overdispersion and zero–inflation as separately identifiable model properties is sensible.

Table 8.10 summarizes our recommended settings for different exposure scenarios. The important message from this table is that models which allow for overdispersion will be needed in the bottom row (due to the densely ionising radiation), and that zero-inflated models will need to be used in the right column (where the body exposure is only partial). The table should not be considered in an 'exclusive' sense – there will often be many other models which will fit well too. We have chosen the named models based on conceptual plausibility, and practical performance in our analysis in Tables 8.3 to 8.6. If one is in doubt about the exposure scenario, ZIP models (especially those which model the zero–inflation parameter linearly) will generally lead to good results.

Zero–inflated models are also directly biologically relevant, as partial body exposures always lead to a mixture of non-irradiated and irradiated blood lym-

Table 8.10: Summary of recommended settings under different exposure scenarios, when counting dicentrics and centric rings (low LET and high LET correspond to sparsely and densely ionising radiation, respectively). When analyzing micronuclei, we would advocate the use of ZINB models irrespective of the exposure pattern.

| exposure | | whole body | partial |
|---|---|---|---|
| LET | low | Poisson/NB | ZIP |
| | high | NB/Neyman A | ZINB |

phocytes within the body at the time of irradiation, and blood sampling for biological dosimetry takes place >24 hours after exposure, the timescale for full circulation of lymphocytes within the human body, so the exposed and un-exposed fractions can reasonably be expected to be fully mixed within the sample taken.

One issue that we have not discussed in this paper is how, given a fitted model, the dose can be estimated from the fitted model for a given aberration count. This is an inverse regression problem; two Bayesian–like solutions to which have been recently provided by Higueras et al. (2015a, 2015b) in whole and partial body exposure scenarios, respectively. Higueras et al. (2015a) can effectively be used for Poisson, NB1, Neyman A and Hermite (r = 2), and can be extended for all two parameter compound Poisson count distributions, this includes Poisson–inverse Gaussian and Pólya–Aeppli models. The approach by Higueras et al. (2015b) can be used for the ZIP(3) distribution. A well–fitting model is, however, absolutely crucial for the success of these techniques. We hope that our manuscript could contribute to addressing this question.

The models used with the cytogenetic example data presented in his work would certainly be applicable to other fields, specifically including nuclear radiation and technology research where the Poisson distribution is frequently applied but also potentially for chemical and other mutagens. Indeed the results of this work have shown that it is useful to formally assess the most appropriate models in a dynamic way wherever count data appear, or models are used to formally assess effects on biological systems.

# Appendix A

# Generalized Hermite distribution modelling with the R package hermite

Submitted to *The R Journal*.
**Abstract:** The Generalized Hermite distribution (and the Hermite distribution as a particular case) is often used for fitting count data in the presence of overdispersion or multimodality. Despite this, to our knowledge, no standard software packages have implemented specific functions to compute basic probabilities and make simple statistical inference based on these distributions. We present here a set of computational tools that allows the user to face these difficulties by modelling with the Generalized Hermite distribution using the `R` package `hermite`. The package can also be used to generate random deviates from a Generalized Hermite distribution and to use basic functions to compute probabilities (density, cumulative density and quantile functions are available), to estimate parameters using the maximum likelihood method and to perform the likelihood ratio test for Poisson assumption against a Generalized Hermite alternative. In order to improve the density and quantile functions performance when the parameters are large, Edgeworth and Cornish–Fisher expansions have been used. Hermite regression is also a useful tool for modeling inflated count data, so its inclusion to a commonly used software like `R` will make this tool available to a wide range of potential users. Some examples of usage in several fields of application are also given.
**Keywords:** discrete distributions, count data, Hermite distributions, Poisson, overdispersion, multi–modality.

## A.1 Introduction

The Poisson distribution is without a doubt the most common when dealing with count data. There are several reasons for that, including the fact that the maximum likelihood estimate of the population mean is the sample mean and the property that this distribution is closed under convolutions (see [44]). However, it is very common in practice that data presents overdispersion or

zero inflation, cases where the Poisson assumption does not hold. In these situations it is reasonable to consider discrete distributions with more than one parameter. The class of all two-parameter discrete distributions closed under convolutions and satisfying that the sample mean is the maximum likelihood estimator of the population mean are characterised in [76]. One of these families is just the Generalized Hermite distribution. Several generalizations of Poisson distribution have been considered in literature (see, for instance, [35, 47, 48, 52]), that are *compound–Poisson* or *contagious* distributions. They are families with probability generating function (PGF) defined by

$$P(s) = exp(\lambda(f(s)-1)) = \exp(a_1(s-1) + a_2(s^2-1) + \ldots + a_m(s^m-1) + \ldots), \tag{A.1}$$

where $f(s)$ is also a PGF and $\sum_{i=1}^{m} a_i = \lambda$. Many well known discrete distributions are included in these families, like the negative binomial, Polya–Aeppli or the Neyman A distributions. The Generalized Hermite distribution was first introduced in [34] as the situation where $a_m$ is significant compared to $a_1$ in (A.1), while all the other terms $a_i$ are negligible, resulting in the PGF

$$P(s) = \exp(a_1(s-1) + a_m(s^m-1)). \tag{A.2}$$

After fixing the value of the positive integer $m \geq 2$, the *order* or *degree* of the distribution, the domain of the parameters is $a_1 > 0$ and $a_m > 0$. Note that when $a_m$ tends to zero, the distribution tends to a Poisson. Otherwise, when $a_1$ tends to zero it tends to $m$ times a Poisson distribution. It is immediate to see that the PGF in (A.1) is the same than the PGF of $X_1 + mX_2$, where $X_1$ and $X_2$ are independent Poisson distributed random variables with population mean $a_1$ and $a_m$ respectively. From here, it is straightforward to calculate the population mean, variance, skewness and excess kurtosis of the Generalized Hermite distribution:

$$\begin{aligned} \mu &= a_1 + ma_m, \\ \sigma^2 &= a_1 + m^2 a_m, \\ \gamma_1 &= \frac{a_1 + m^3 a_m}{(a_1 + m^2 a_m)^{3/2}}, \\ \gamma_2 &= \frac{a_1 + m^4 a_m}{(a_1 + m^2 a_m)^2}. \end{aligned} \tag{A.3}$$

A useful expression for the probability mass function of the Generalized Hermite distribution in terms of the population mean $\mu$ and the population index of dispersion $d = \sigma^2/\mu$ is provided in [76].

$$P(Y=k) = P(Y=0)\frac{\mu^k(m-d)^k}{(m-1)^k} \sum_{j=0}^{[k/m]} \frac{(d-1)^j(m-1)^{(m-1)j}}{m^j \mu^{(m-1)j}(m-d)^{mj}(k-mj)!j!}, \tag{A.4}$$

where $k = 0, 1, \ldots$, $P(Y=0) = \exp(\mu(-1 + \frac{d-1}{m}))$ and $[k/m]$ is the integer part of $\frac{k}{m}$. Note that $m$ can be expressed as $m = \frac{d-1}{1+\log(p_0)/\mu}$. Because the denominator is a measure of zero inflation, $m$ can be understood as an index of the relationship between the overdispersion and the zero inflation.

The probabilities can be also written in terms of the parameters $a_1$, $a_m$ using the identities given in (A.3).

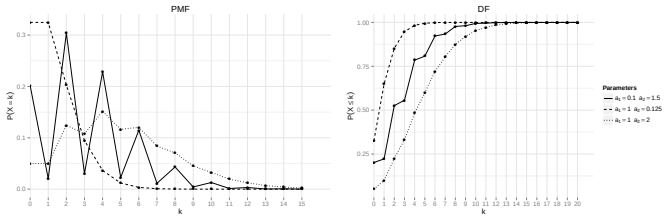


Figure A.1: Hermite probability mass and distribution functions for the indicated parameter values.

The case $m = 2$ in (A.1) is covered in detail in [47] and [48] and the resulting distribution is simply called Hermite distribution. In that case, the probability mass function, in terms of the parameters $a_1$ and $a_2$, has the expression

$$P(Y = k) = e^{-a_1 - a_2} \sum_{j=0}^{[k/2]} \frac{a_1^{k-2j} a_2^j}{(k-2j)!j!}, k = 0, 1, \ldots \tag{A.5}$$

The probability mass function and the distribution function for some values of $a_1$ and $a_2$ are shown in Figure A.1.

In [34] the authors also develop a recurrence relation that can be used to calculate the probabilities in a numerically efficient way:

$$p_k = \frac{\mu}{k(m-1)} \left( p_{k-m}(d-1) + p_{k-1}(m-d) \right), k \geq m, \tag{A.6}$$

where $p_k = P(Y = k)$ and the first values can be computed as $p_k = p_0 \frac{\mu^k}{k!} \left( \frac{m-d}{m-1} \right)^k$, $k = 1, \ldots, m-1$. Although overdispersion or multimodality are common situations when dealing with count data and Generalized Hermite distribution provides an appropriate framework to face these situations, the use of techniques based on this distribution was not easy in practice as they were not available in any standard statistical software. A description of the package `hermite` main functionalities will be given in Section A.2. Several examples of application on different fields will be discussed in Section A.3, and finally some conclusions will be commented in Section A.4.

## A.2 Package hermite

As the common distributions in `R` [81], the package `hermite` implements the probability mass function (dhermite()), the distribution function (phermite()), the quantile function (qhermite()) and a function for random generation (rhermite()) for the Generalized Hermite distribution. It also includes the function glm.hermite(), which allows to calculate, for an univariate sample of independent draws, the maximum likelihood estimates for the parameters and to perform the likelihood ratio test for a Poisson null hypothesis against a Generalized Hermite alternative. This function can also carry out Hermite regression including covariates for the population mean, in a very similar way to that of the well known `R` function glm().

### A.2.1 Probability mass function

The probability mass function of the Generalized Hermite distribution is implemented in `hermite` through the function dhermite(). A call to this function might be

```
dhermite(x, a, b, m=2)
```

The description of these arguments can be summarized as follows:

- $x$: Vector of non-negative integer values.
- $a$: First parameter for the Hermite distribution.
- $b$: Second parameter for the Hermite distribution.
- $m$: Degree of the Generalized Hermite distribution. Its default value is 2, corresponding to the classical Hermite distribution introduced in [47].

The recurrence relation (A.6) is used by dhermite() for the computation of probabilities. For large values of any parameter $a$ or $b$ (above 20), the probability of $Y$ taking $x$ counts is approximated using an Edgeworth expansion of the distribution function (A.7), i.e. $P(Y = x) = F_H(x) - F_H(x-1)$. The Edgeworth expansion does not guarantee positive values for the probabilities in the tails, so in case this approximation returns a negative probability, the probability is calculated by using the normal approximation $P(Y = x) = \Phi(x^+) - \Phi(x^-)$ where $\Phi$ is the typified normal distribution function and

$$x^{\pm} = \frac{x \pm 0.5 - a - mb}{\sqrt{a + m^2 b}}$$

are the typified continuous corrections.

The normal approximation is justified taking into account the representation of any Generalized Hermite random variable $Y$, as $Y = X_1 + mX_2$ where $X_i$ are independent Poisson distributed with population means $a$, $b$. Therefore, for large values of $a$ or $b$, the Poissons are well approximated by normal distributions.

### A.2.2 Distribution function

The distribution function of the Generalized Hermite distribution is implemented in `hermite` through the function phermite(). A call to this function might be

```
phermite(q, a, b, m=2, lower.tail=TRUE)
```

The description of these arguments can be summarized as follows:

- $q$: Vector of non–negative integer quantiles.
- *lower.tail*: Logical; if *TRUE* (the default value), the computed probabilities are $P(Y \leq x)$, otherwise, $P(Y > x)$.

All remaining arguments are defined as specified for dhermite().

If $a$ and $b$ are large enough ($a$ or $b > 20$), $X_1$ and $X_2$ are approximated by $N(a, \sqrt{a})$ and $N(b, \sqrt{b})$ respectively, so $Y$ can be approximated by a normal distribution with mean $a + mb$ and variance $a + m^2 b$. This normal approximation is improved by means of an Edgeworth expansion [8], using the following expression

$$F_H(x) \approx \Phi(x^*) - \phi(x^*) \cdot \left( \frac{1}{6}\gamma_1 He_2(x^*) + \frac{1}{24}\gamma_2 He_3(x^*) + \frac{1}{72}\gamma_1^2 He_5(x^*) \right), \tag{A.7}$$

where $\Phi$ and $\phi$ are the typified normal distribution and density functions respectively, $He_n(x)$ are the $n$th-degree probabilists' Hermite polynomials [8]

$$\begin{aligned} He_2(x) &= x^2 - 1 \\ He_3(x) &= x^3 - 3x \\ He_5(y) &= x^5 - 10x^3 + 15x, \end{aligned}$$

$x^*$ is the typified continuous correction of $x$ considered in [71]

$$x^* = 1 + \frac{1}{24(a + m^2 b)} \cdot \frac{x + 0.5 - a - mb}{\sqrt{a + m^2 b}},$$

and $\gamma_1$ and $\gamma_2$ are respectively the skewness and the excess kurtosis of $Y$ expressed in (A.3).

### A.2.3 Quantile function

The quantile function of the Generalized Hermite distribution is implemented in `hermite` through the function qhermite(). A call to this function might be

```
qhermite(p, a, b, m=2, lower.tail=TRUE)
```

The description of these arguments can be summarized as follows:

- $p$: Vector of probabilities.

All remaining arguments are defined as specified for phermite(). The quantile is right continuous: `qhermite(p, a, b, m)` is the smallest integer $x$ such that $P(Y \leq x) \geq p$, where $Y$ follows a $m$–th order Hermite distribution with parameters $a$ and $b$.

When the parameters $a$ or $b$ are over 20, a Cornish-Fisher expansion is used [8] to approximate the quantile function. The Cornish–Fisher expansion uses the following expression

$$y_p \approx \left( u_p + \frac{1}{6}\gamma_1 He_2(u_p) + \frac{1}{24}\gamma_2 He_3(u_p) - \frac{1}{36}\gamma_1^2 (2u_p^3 - 5u_p) \right) \sqrt{a + m^2 b} + a + mb,$$

where $u_p$ is the $p$ quantile of the typified normal distribution.

### A.2.4 Random generation

The random generation function rhermite() uses the relationship between Poisson and Hermite distributions detailed in Sections A.1 and A.2.1. A call to this function might be

```
rhermite(n, a, b, m=2)
```

The description of these arguments can be summarized as follows:

- $n$: Number of random values to return.

All remaining arguments are defined as specified for dhermite().

### A.2.5 Maximum likelihood estimation and Hermite regression

Given a sample $X = x_1, \ldots, x_n$ of a population coming from a generalized Hermite distribution with mean $\mu$, index of dispersion $d$ and order $m$, the log-likelihood function is

$$l(X; \mu, d) = n \cdot \mu \cdot \left(-1 + \frac{d-1}{m}\right) + log\left(\frac{\mu(m-d)}{m-1}\right) \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} log(q_i(\theta)), \quad \text{(A.8)}$$

where $q_i(\theta) = \sum_{j=0}^{[x_i/m]} \frac{\theta^j}{(x_i - mj)!j!}$ and $\theta = \frac{(d-1)(m-1)^{(m-1)}}{m\mu^{(m-1)}(m-d)^m}$.

The maximum likelihood equations do not always have a solution. It is due to the fact that this is not a regular family of distributions because its domain of parameters is not an open set. The following result gives a sufficient and necessary condition for the existence of such a solution [76]:

**Proposition A.2.1** *Let $x_1, \ldots, x_n$ be a random sample from a generalized Hermite population with fixed $m$. Then, the maximum likelihood equations have solution if and only if $\frac{\mu^{(m)}}{\bar{x}^m} > 1$, where $\bar{x}$ is the sample mean and $\mu^{(m)}$ is the $m$-th order sample factorial moment, $\mu^{(m)} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i(x_i - 1) \cdots (x_i - m + 1)$.*

If the likelihood equations do not have a solution, the maximum of the likelihood function (A.8) is attained at the border of the domain of parameters, that is, $\hat{\mu} = \bar{x}$, $\hat{d} = 1$ (Poisson distribution), or $\hat{\mu} = \bar{x}$, $\hat{d} = m$ ($m$ times a Poisson distribution). The case $\hat{\mu} = \bar{x}$, $\hat{d} = m$ corresponds to the very improbable situation where all the observed values were multiples of $m$. Then, in general, when the condition of Proposition A.2.1 is not satisfied, the maximum likelihood estimators are $\hat{\mu} = \bar{x}$, $\hat{d} = 1$. It means that data is fitted assuming a Poisson distribution.

The package `hermite` allows to estimate the parameters $\mu$ and $d$ given an univariate sample by means of the function glm.hermite():

```
glm.hermite(formula, data, link="log", start=NULL, m=NULL)
```

The description of the arguments can be summarized as follows:

- *formula*: Symbolic description of the model. A typical predictor has the form *response* $\sim$ *terms* where *response* is the (numeric) response vector and *terms* is a series of terms which specifies a linear predictor for response.
- *data*: An optional data frame containing the variables in the model.
- *link*: Character specification of the population mean link function: "log" or "identity". By default *link="log"*.

- *start*: A vector containing the starting values for the parameters of the specified model. Its default value is NULL.
- $m$: Value for parameter $m$. Its default value is *NULL*, and in that case it will be estimated as $\hat{m}$, more details below.

The returned value is an object of class *glm.hermite*, which is a list including the following components:

- *coefs*: The vector of coefficients.
- *loglik*: Log-likelihood of the fitted model.
- *vcov*: Covariance matrix of all coefficients in the model (derived from the Hessian returned by the maxLik() output).
- *hess*: Hessian matrix, returned by the maxLik() output.
- *fitted.values*: The fitted mean values, obtained by transforming the linear predictors by the inverse of the link function.
- $w$: Likelihood ratio test statistic.
- *pval*: Likelihood ratio test p-value.

If the condition given in Proposition A.2.1 is not met for a sample $x$, the glm.hermite() function provides the maximum likelihood estimates $\hat{\mu} = \bar{x}$ and $\hat{d} = 1$ and a warning message advising the user that the MLE equations have no solutions.

The function glm.hermite() can also be used for Hermite regression as described below and as will be shown through practical examples in Sections A.3.3 and A.3.4.

Covariates can be incorporated into the model in various ways (see e.g. [29]). In function glm.hermite(), the distribution is specified in terms of the dispersion index and its mean, which is then related to explanatory variables as in linear regression or other generalized linear models. That is, for Hermite regression, we assume $Y_i$ follows a generalized Hermite distribution of order $m$, where we retain the dispersion index $d$ ($> 1$) as a parameter to be estimated and let the mean $\mu_i$ for the $i$–th observation vary as a function of the covariates for that observation, i.e., $\mu_i = h(\boldsymbol{x_i}^t \beta)$, where $\boldsymbol{x_i}$ is a vector of covariates, $t$ denotes the transpose vector, $\beta$ is the corresponding vector of coefficients to be estimated and $h$ is a link function. Note that because the dispersion index $d$ is taken constant, this is a linear mean-variance (NB1) regression model.

The link function provides the relationship between the linear predictor and the mean of the distribution function. Although the log is the canonical link for count data, since ensures that all the fitted values are positive, their choice can be somewhat arbitrary or be in influenced by the data to be treated. For example, the identity link is the accepted standard in biodosimetry since there is no evidence that the increase of aberration counts with dose is of exponential shape [42]. Therefore, function glm.hermite() allows both link functions.

A consequence of using the identity–link is that the maximum likelihood estimate of the parameters obtained by maximizing the log–likelihood function of the corresponding model may lead to negative values for the mean. Therefore, in

order to avoid negative values for the mean, constraints in the domain of the parameters must be included when the model is fitted. In function glm.hermite(), this is carried out by using the function maxLik() from package `maxLik` [38], which is used internally for maximizing the corresponding log–likelihood function. This function allows to introduce constraints which are needed when the identity–link is used.

It should also be noted that results may depend of the starting values provided to the optimization routines. If no starting values are supplied, the starting values for the coefficients are computed/fixed internally. Specifically, when the log–link is specified, the starting values are obtained by fitting a standard Poisson regression model through a call to the internal function glm.fit() from package `stats`. If the link function is the identity and no initial values are provided by the user, the function takes 1 as initial value for the coefficients. In both cases, the initial value for the dispersion index $\hat{d}$ is taken to 1.1.

Regarding to the order of the Hermite distribution, it can be fixed by the user. If it is not provided (default option), when the model includes covariates, the order $\hat{m}$ is selected by discretised maximum likelihood method, fitting the coefficients for each value of $\hat{m}$ between 1 (Poisson) and 10, and selecting the case that maximizes the likelihood. In addition, if no covariates are included in the model and no initial values are supplied by the user, the naïve estimate $\hat{m} = \frac{s^2/\bar{x}-1}{1+log(p0)/\bar{x}}$, where $p0$ is the proportion of zeros in the sample is also considered. In the unlikely case the function returns $\hat{m} = 10$, we recommend to check the likelihood of the next orders ($m = 11, 12, \ldots$) fixing this parameter in the function until a local minimum is found.

When dealing with the Generalized Hermite distribution it seems natural to wonder if data could be fitted by using a Poisson distribution. Because the Poisson distribution is included in the Generalized Hermite family, this is equivalent to test the null hypothesis $H_0 : d = 1$ against the alternative $H_1 : d > 1$. To do this, an immediate solution is to use the likelihood ratio test, which test–statistic is given by $W = 2(l(X; \hat{\mu}, \hat{d}) - l(X; \hat{\mu}, 1))$, where $l$ is the log-likelihood function.

Under the null hypothesis $W$ is not asymptotically $\chi^2_1$ distributed as usual, because $d = 1$ is on the border of the domain of the parameters. Using the results of [25] and [84] it can be shown that in this case the asymptotic distribution of $W$ is a 50:50 mixture of a zero constant and a $\chi^2_1$ distribution. The $\alpha$ percentile for this mixture is the same as the $2\alpha$ upper tail percentile for a $\chi^2_1$ [76]. The likelihood ratio test is also performed through glm.hermite() function, using the maximum likelihood estimates $\hat{\mu}$ and $\hat{d}$.

A *summary* method for objects of class *glm.hermite* is included in the `hermite` package, giving a summary of relevant information, as the residuals minimum, maximum, median and first and third quartiles, the table of coefficients including the corresponding standard errors and significance tests based on the Normal reference distribution for regression coefficients and the likelihood ratio test against the Poisson distribution for the dispersion index. The AIC value for the proposed model is also reported.

| Microarthropods per sample | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Frequency | 122 | 40 | 14 | 16 | 6 | 2 |

Table A.1: Frequency distribution of Collenbola microarthropods.

| Microarthropods per sample | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Frequency | 118.03 | 49.11 | 10.22 | 14.56 | 5.61 | 1.15 |

Table A.2: Expected frequency distribution of Collenbola microarthropods.

## A.3 Examples

Several examples of application of the package `hermite` in a wide range of contexts are discussed in this section, including classical and recent real datasets and simulated data.

### A.3.1 Hartenstein (1961)

This example by [37] describes the counts of Collenbola microarthropods in 200 samples of forest soil. The frequency distribution is shown in Table A.1.

This dataset was analysed in [76] with a Generalized Hermite distribution of order $m = 3$. The maximum likelihood estimation gave a mean of $\hat{\mu} = \bar{x} = 0.75$, and an index of dispersion of $\hat{d} = 1.8906$.

Using glm.hermite() we calculate the parameter estimates:

```
R> library("hermite")
R> data <- c(rep(0,122), rep(1,40), rep(2,14), rep(3,16),
 rep(4,6), rep(5,2))
R> glm.hermite(data~1, link="log", start=NULL, m=3)$coefs

     (Intercept) dispersion.index            order
      -0.2875851        1.8905920        3.0000000
```

We can see that these parameter estimates are equivalent to those reported in [76].

The estimated expected frequencies are shown in Table A.2.

The frequencies in Table A.2 have been obtained running the following code and using the transformation

$$\begin{array}{ll} b & = \frac{\mu(d-1)}{m(m-1)}, \\ a & = \mu - mb. \end{array} \tag{A.9}$$

```
R> a <- -exp(mle1$coefs[1])*(mle1$coefs[2] -
 mle1$coefs[3])/(mle1$coefs[3] - 1)
R> b <- exp(mle1$coefs[1])*(mle1$coefs[2] -
 1)/(mle1$coefs[3]*(mle1$coefs[3] - 1))
R> exp <- round(dhermite(seq(0,5,1), a, b, m=3)*200,2)
```

Note that the null hypothesis of Poisson distributed data is strongly rejected, with a likelihood ratio test statistic $W = 48.66494$ and its corresponding p-value$= 1.518232e-12$, as we can see with

| Currency and banking crises | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Observed | 45 | 44 | 19 | 17 | 19 | 13 | 6 | 4 |
| Expected (Hermite) | 38.51 | 35.05 | 37.40 | 24.36 | 15.96 | 8.33 | 4.23 | 1.88 |
| Expected (Poisson) | 22.07 | 44.66 | 45.20 | 30.49 | 15.43 | 6.25 | 2.11 | 0.61 |

Table A.3: Observed and expected frequency distributions of currency and banking crises.

```
R> mle1$w
[1] 48.66494
R> mle1$pval
[1] 1.518232e-12
```

### A.3.2 Giles (2010)

In [30], the author explores an interesting application of the classical Hermite distribution ($m = 2$) in an economic field. In particular, he proposes a model for the number of currency and banking crises. The reported maximum likelihood estimates for the parameters were $\hat{a} = 0.936$ and $\hat{b} = 0.5355$, slightly different from those obtained using glm.hermite(), which are $\hat{a} = 0.910$ and $\hat{b} = 0.557$. The actual and estimated expected counts under Hermite and Poisson distribution assumptions are shown in Table A.3.

In this example, the likelihood ratio test clearly rejects the Poisson assumption in favor of the Hermite distribution ($W = 40.08$, p–value=$1.22e - 10$).

The expected frequencies of the Hermite distribution shown in Table A.3 have been calculated running the code,

```
R> exp2 <- round(dhermite(seq(0,7,1), 0.910, 0.557, m=2)*167, 2)
R> exp2
```

### A.3.3 Giles (2007)

In [29] the author proposes an application of Hermite regression to the 965 number 1 hits on the Hot 100 chart over the period January 1955 to December 2003. The data were compiled and treated with different approaches by Giles (see [28] for instance), and is available for download at the author website `http://web.uvic.ca/~dgiles/`. For all recordings that reach the number one spot, the number of weeks that it stays at number one was recorded. The data also allow for reentry into the number one spot after having being relegated to a lower position in the chart. The actual and predicted counts under Poisson and Hermite distributions are shown in Table A.4.

Several dummy covariates were also recorded, including indicators of whether the recording was by Elvis Presley or not, the artist was a solo female, the recording was purely instrumental and whether the recording topped the charts in nonconsecutive weeks.

The estimates and corresponding standard errors are obtained through the instructions

```
R> fit.hot100 <- glm.hermite(Weeks2 ~ Elvis+Female+Inst+NonCon,
 data=data.df, start=NULL, m=2)
R> fit.hot100$coefs
    (Intercept)            Elvis               Female
```

| Weeks | Actual | Poisson | Hermite |
|---|---|---|---|
| 0 | 337 | 166.95 | 317.85 |
| 1 | 249 | 292.91 | 203.58 |
| 2 | 139 | 256.94 | 214.60 |
| 3 | 93 | 150.26 | 109.61 |
| 4 | 47 | 65.90 | 67.99 |
| 5 | 35 | 23.12 | 29.32 |
| 6 | 21 | 6.76 | 13.78 |
| 7 | 13 | 1.69 | 5.20 |
| 8 | 9 | 0.37 | 2.04 |
| 9 | 8 | 0.07 | 0.69 |
| 10 | 5 | 0.01 | 0.24 |
| 11 | 2 | 0.00 | 0.07 |
| 12 | 2 | 0.00 | 0.02 |
| 13 | 4 | 0.00 | 0.01 |
| 14 | 0 | 0.00 | 0.00 |
| 15 | 1 | 0.00 | 0.00 |

Table A.4: Observed and expected frequency distribution of Hot 100 data.

```
      0.4578140         0.9126080             0.1913968
           Inst            NonCon      dispersion.index
      0.3658427         0.6558621             1.5947901
          order
      2.0000000
R> sqrt(diag(fit.hot100$vcov))
[1] 0.03662962 0.16208682 0.07706250 0.15787552
 0.12049736 0.02533045
```

For instance, we can obtain the predicted value for the average number of weeks that an Elvis record hits the number one spot. According to the model, we obtain a predicted value of 3.9370, while the observed corresponding value is 3.9375. The likelihood ratio test result justifies the fitting through a Hermite regression model instead of a Poisson model:

```
R> fit.hot100$w;fit.hot100$pval
[1] 385.7188
[1] 3.53909e-86
```

### A.3.4 DiGiorgio et al. (2004)

In [20] the authors perform an experimental simulation of *in vitro* whole body irradiation for high–LET radiation exposure, where peripheral blood samples were exposed to 10 different doses of 1480MeV oxygen ions. For each dose, the number of dicentrics chromosomes per blood cell were scored. The corresponding data is included in the package `hermite`, and can be loaded into the `R` session by

```
R> data(hi_let)
```

In [78] the authors apply Hermite regression (to contrast the Poisson assumption) for fitting the dose–response curve, i.e. the yield of dicentrics per cell as a quadratic function of the absorbed dose linked by the identity function (which is commonly used in biodosimetry). This model can be fitted using the glm.hermite() function in the following way:

```
R> fit.hlet.id <- glm.hermite(Dic~Dose+Dose2-1, data=hi_let,
link="identity")
```

Note that the model defined in `fit.hlet.id` has no intercept.

A summary of the most relevant information can be obtained using the *summary()* method as in

```
R> summary(fit.hlet.id)

Call: glm.hermite(formula = Dic ~ Dose + Dose2 - 1, data = hi_let,
 link = "identity")
Deviance Residuals:
        Min          1Q      Median          3Q         Max
-0.06261842 -0.03536264  0.00000000  0.10373982  1.42251927

Coefficients:
                  Estimate Std. Error    z value      p-value
Dose              0.4620671 0.03104362  14.884450 4.158952e-50
Dose2             0.1555365 0.04079798   3.812357 1.376478e-04
dispersion.index 1.2342896 0.02597687 107.824859 1.468052e-25
order             2.0000000         NA         NA           NA
(Likelihood ratio test against Poisson is reported by *z value*
 for *dispersion.index*)

AIC:  5592.422
```

We can see the maximum likelihood estimates and corresponding standard errors in the output from the *summary* output. Note also that the likelihood ratio test rejects the Poisson assumption ($W = 107.82$, p-value$=1.47e-25$).

### A.3.5 Higueras et al. (2015a)

In the first example in [39] the Bayesian estimation of the absorbed dose by Cobalt–60 gamma rays after the *in vitro* irradiation of a sample of blood cells is given by a density proportional to the probability mass function of a Hermite distribution taking 102 counts whose mean and variance are functions of the dose $x$, respectively $\mu(x) = 45.939x^2 + 5.661x$ and $v(x) = 8.913x^4 - 22.553x^3 + 69.571x^2 + 5.661x$.

The reparameterisation in terms of $a$ and $b$ as a function of the dose is given by the transformation

$$\begin{array}{rl} a(x) & = 2\mu(x) - v(x) \\ b(x) & = \frac{v(x)-\mu(x)}{2} \end{array} \tag{A.10}$$

This density only makes sense while $a(x)$ and $b(x)$ are positive. The dose $x > 0$ and consequently $b(x)$ is always positive and $a(x)$ is positive for $x < 3.337$. Therefore, the probability density outside $(0, 3.337)$ is 0.

The following code generates the plot of the resulting density,

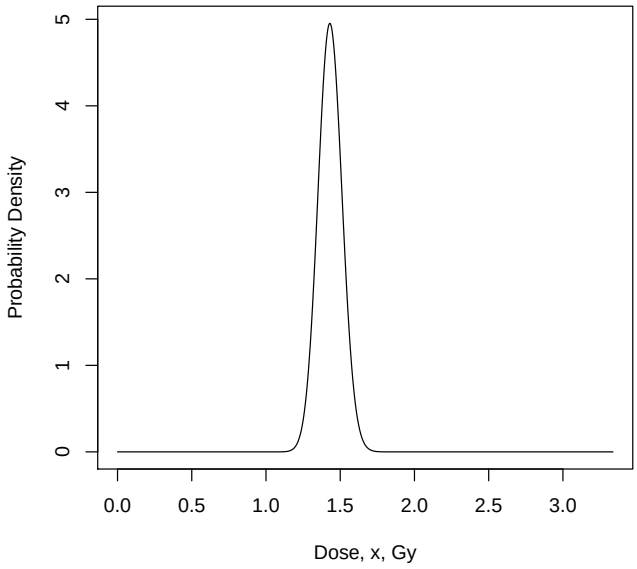


Figure A.2: Absorbed dose density plot.

```
R> u <- function(x) 45.939*x^2 + 5.661*x
R> v <- function(x) 8.913*x^4 - 22.553*x^3 + 69.571*x^2 + 5.661*x
R> a <- function(x) 2*u(x) - v(x); b <- function(x) (v(x) - u(x))/2
R> dm <- uniroot(function(x) a(x), c(1, 4))$root; dm
R> nc <- integrate(Vectorize(function(x) dhermite(102, a(x),
 b(x))), 0, dm)$value
R> cd <- function(x){ vapply(x, function(d) dhermite(102, a(d),
 b(d)), 1)/nc }
R> x <- seq(0, dm, .001)
R> plot(x, cd(x), type="l", ylab="Probability Density",
 xlab="Dose, x, Gy")
```

Figure A.2 shows this resulting density.

### A.3.6 Random number generation

A vector of random numbers following a Generalized Hermite distribution can be obtained by means of the function rhermite(). For instance, the next code generates 1000 observations according to an Hermite regression model, including Bernoulli and normal covariates `x1` and `x2`:

```
R> n <- 1000
R> #### Regression coefficients
R> b0 <- -2
R> b1 <- 1
R> b2 <- 2
```

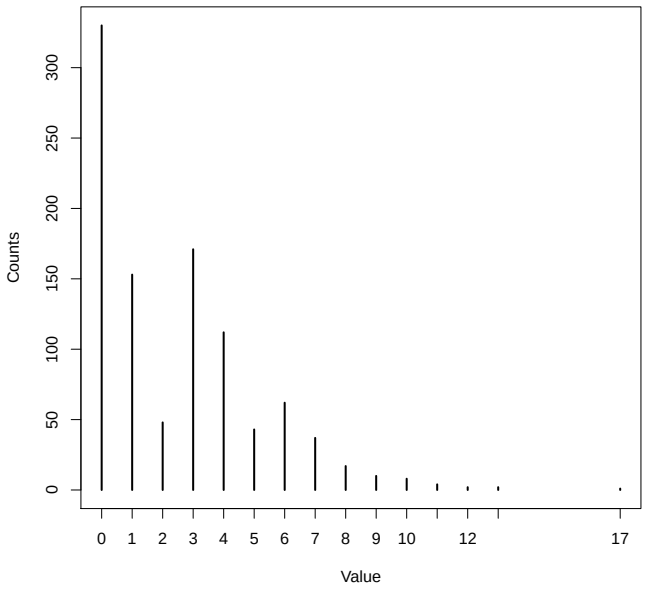


Figure A.3: Random generated Hermite values

```
R> #### Covariate values
R> set.seed(111111)
R> x1 <- rbinom(n, 1, .75)
R> x2 <- rnorm(n, 1, .1)
R> u <- exp(b0 + b1*x1 + b2*x2)
R> d <- 2.5
R> m <- 3
R> b <- u*(d - 1)/(m*(m - 1))
R> a <- u - m*b
R> x <- rhermite(n, a, b, m)
```

This generates a multimodal distribution, as can be seen in Figure A.3. The probability that this distribution take a value under 5 can be computed using the function phermite():

```
R> phermite(5, mean(a), mean(b), m = 3) [1] 0.8734357
```

Conversely, the value that has an area on the left of 0.9653287 can be computed using the function qhermite():

```
R> qhermite(0.8734357, mean(a), mean(b), m = 3) [1] 5
```

```
R> mle3 <- glm.hermite(x ~ factor(x1) + x2, m = 3)
R> mle3$coefs
     (Intercept)      factor(x1)1               x2 dispersion.index
       -1.809163         1.017805         1.839606         2.491119
```

```
          order
       3.000000
R> mle3$w;mle3$pval
[1] 771.7146
[1] 3.809989e-170
```

In order to check the performance of the likelihood ratio test, we can simulate a Poisson sample and run the function glm.hermite() again:

```
R> y <- rpois(n, u)
R> mle4 <- glm.hermite(y ~ factor(x1) + x2, m=3)
R> mle4$coefs[4]
    dispersion.index
                   1
R> mle4$w; mle4$pval [1] -5.475874e-06 [1] 0.5
```

We can see that in this case the maximum likelihood estimate of the dispersion index $d$ is almost 1 and that the Poisson assumption is not rejected (p-value= 0.5). If the MLE equations have no solution, the function glm.hermite() will return a warning:

```
R> z <- rpois(n, 20)
R> mle5 <- glm.hermite(z ~ 1, m=4)
    Warning message:
    In glm.hermite(z ~ 1, m = 4) : MLE equations have no solution
```

In that case, we have $\frac{\mu^{(4)}}{\bar{z}^4} = 0.987$ and therefore the condition of Proposition A.2.1 is not met.

## A.4 Conclusions

Hermite distributions can be useful for modeling count data that presents multimodality or overdispersion, situations that appear commonly in practice in many fields. In this article we present the computational tools that allow to overcome these difficulties by means of the Generalized Hermite distribution (and the classical Hermite distribution as a particular case) compiled as an `R` package. The `hermite` package also allows the user to perform the likelihood ratio test for Poisson assumption and to estimate parameters using the maximum likelihood method. Hermite regression is also a useful tool for modeling inflated count data, and it can be carried out by the `hermite` package in a flexible framework and including covariates. Currently, the `hermite` package is also used by the `radir` package [65] that implements a Bayesian innovative method for radiation biodosimetry introduced in [39].

# Bibliography


[1] Ainsbury EA, Lloyd DC. Dose estimation software for radiation biodosimetry. *Health Physics* 2010; **98**:290–295.

[2] Ainsbury EA, Vinnikov VA, Maznyk NA, Lloyd DC, Rothkamm K. A Comparison of six statistical distributions for analysis of chromosome aberration data for radiation biodosimetry. *Radiation Protection Dosimetry* 2013; **155**:253–267.

[3] Ainsbury EA, Vinnikov VA, Puig P, Maznyk NA, Rothkamm K, Lloyd DC. CytoBayesJ: Software tools for Bayesian analysis of cytogenetic radiation dosimetry data. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* 2013; **756**:184–191.

[4] Ainsbury EA, Vinnikov VA, Puig P, Higueras M, Maznyk NA, Lloyd DC, Rothkamm K. Review of Bayesian statistical analysis methods for cytogenetic radiation biodosimetry, with a practical example. *Radiation Protection Dosimetry* 2014; **162**(3):185–196.

[5] Aitchison J, Shen SM. Logistic–normal distributions, Some properties and uses. *Biometrika* 1980; **67**:261–272.

[6] Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; **19**:716–723.

[7] Alfò M, Maruotti A. Two–part regression models for longitudinal zero–inflated count data. *The Canadian Journal of Statistics* 2010; bf 38(2):197–216.

[8] Barndorff–Nielsen OE, Cox DR. *Asymptotic techniques for use in statistics.* Chapman & Hall, 1989.

[9] Barquinero JF, Barrios L, Caballín MR, Miró R, Ribas M, Subias A, Egozcue J. Establishment and validation of a dose–effect curve for $\gamma$–rays by cytogenetic analysis. *Mutation Research* 1995; **326**:65–69.

[10] Barquinero JF, Barrios L, Caballín MR, Miró R, Ribas M, Egozcue J. Biological dosimetry in simulated in vitro partial irradiations. *International Journal of Radiation Biology* 1997; **71**:435–440.

[11] Bayarri MJ, Berger JO, Datta GS. Objective Bayes testing of Poisson versus inflated Poisson models. *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K.*

*Ghosh* 105–121, Institute of Mathematical Statistics, 2008. See http://projecteuclid.org/euclid.imsc/1209398464.

[12] Bender MA, Awa AA, Brooks AL, Evans HJ, Groer PG, Littlefield LG, Pereira C, Preston RJ, Wachholz BW. Current status of cytogenetic procedures to detect and quantify previous exposures to radiation. *Mutation Research* 1998; **196**:103–159.

[13] Böhning D, Dietz E, Schlattmann P, Mendoca L, Kirchner U. The zero–inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society A* 1999; **162**, 195–209.

[14] Brame RS, Groer PG. Bayesian analysis of overdispersed chromosome aberration data with the negative binomial model. *Radiation Protection Dosimetry* 2002; **102**:115–119.

[15] Brame RS, Groer PG. Bayesian methods for chromosome dosimetry following a criticality accident. *Radiation Protection Dosimetry* 2003; **104**:61–63.

[16] Christensen R, Johnson W, Branscum A, Hanson TE. *Bayesian ideas and data analysis*. Chapman and Hall/CRC Press, 2011.

[17] Demidenko E, Williams BB, Flood AB, Swartz HM. Standard error of inverse prediction for dose–response relationship: approximate and exact statistical inference. *Statistics in Medicine* 2012; **32**:2048–2061.

[18] Deperas J, Szluinska M, Deperas–Kaminska M, Edwards A, Lloyd D, Lindholm C, Romm H, Roy L, Moss R, Morand J, Wojcik A. CABAS: a freely available PC program for fitting calibration curves in chromosome aberration dosimetry. *Radiation Protection Dosimetry* 2007; **124**(2):115–123.

[19] DiCarlo AL, Maher C, Hick JL, Hanfling D, Dainiak N, Chao N, Bader JL, Coleman CN, Weinstock DM. Radiation injury after a nuclear detonation: medical consequences and the need for scarce resources allocation. *Disaster Medicine and Public Health Preparedness* 2011; **5**:S32-S44.

[20] Di Giorgio M, Edwards AA, Moquet J, Finnon P, Hone PA, Lloyd DC, Kreiner AJ, Schuff JA, Taja MR, Vallerga MB, López FO, Burlón A, Debray ME, Valda A. Chromosome aberrations induced in human lymphocytes by heavycharged particles in track segment mode. *Radiation Protection Dosimetry* 2004; **108**:47-53.

[21] Di Giorgio M, Zaretzky A. Biological dosimetry – A Bayesian approach for presenting uncertainty on biological dose estimates. *Annals of 'II Encuentro de Docentes e Investigadores de Estadstica en Psicologa'*. University of Buenos Aires, 2011.

[22] Edwards AA, Lloyd DC, Purrott RJ. Radiation Induced Chromosome Aberrations and the Poisson Distribution. *Radiation and Environmental Biophysics* 1979; **16**:89–100.

[23] Frome EL, DuFrain RJ. Maximum likelihood estimation for cytogenetic dose-response curves. *Biometrics* 1986; **42**:73–84.

[24] Frühwirth–Schnatter, S. *Finite Mixture and Markov Switching Models*. Springer Series in Statistics, 2006.

[25] Geyer CJ. On the asymptotics of constrained M–estimation. *The Annals of Statistics* 1994; **22**(4):1993-2010.

[26] Ghosh JK, Delampady M, Samanta T. *An Introduction to Bayesian Analysis, Theory and Methods*. Springer Texts in Statistics, 2006.

[27] Ghosh SK, Mukhopadhyay P, Lu JC. Bayesian analysis of zero–inflated regression models. *Journal of Statistical Planning and Inference* 2006; **136**:1360–1375.

[28] Giles DE. Superstardom in the US popular music industry revisited. *Economics Letters* 2006; **92**(1):68-74.

[29] Giles DEA. Modeling inflated count data. *Proceedings of the MODSIM 2007 International Congress on Modelling and Simulation*. Modelling and Simulation Society of Australia and New Zealand, 2007.

[30] Giles DE. Hermite regression analysis of multi-modal count data. *Economics Bulletin* 2010; **30**(4):2936-2945.

[31] Groer PG. Exact and approximate Bayesian estimation of net counting rates. *Radiation Protection Dosimetry* 2002; **102**:265–268.

[32] Groer PG, Carnes BA. Bayesian estimation of dose thresholds. *Radiation Protection Dosimetry* 2003; **103**:111–115.

[33] Groer PG, Pereira CABD. Calibration of a radiation detector: chromosome dosimetry for neutrons. In: *Probability and Bayesian Statistics*. Ed. R. Viertl, Plenum Publishing Corporation, 1987; 225–232.

[34] Gupta R, Jain G. A generalized hermite distribution and its properties. *SIAM Journal on Applied Mathematics* 1974; **27**(2):359-363.

[35] Gurland J. Some interrelations among compound and generalized distributions. *Biometrika* 1957; **44**:265–268.

[36] Hall EJ, Giaccia AJ. *Radiobiology for the radiologist*, 7$^{\text{th}}$ edition. Lippincott Williams & Wilkins, 2012.

[37] Hartenstein R. On the distribution of forest soil microarthropods and their fit to 'contagious' distribution functions. *Ecology* 1961; **42**(1):190–194.

[38] Henningsen A, Toomet O. maxLik: A package for maximum likelihood estimation in R. *Computational Statistics* 2011; **26**(3):443-458.

[39] Higueras M, Puig P, Ainsbury EA, Rothkamm K. A new inverse regression model applied to radiation biodosimetry. *Proceedings of the Royal Society A* 2015, DOI: 10.1098/rspa.2014.0588.

[40] Higueras M, Puig P, Ainsbury EA, Vinnikov VA, Rothkamm K. A new Bayesian model applied to cytogenetic partial body irradiation estimation. *Radiation Protection Dosimetry* 2015, DOI: 10.1093/rpd/ncv356.

[41] Hürlimann W. A Characterization of the Compound Multiparameter Hermite Gamma Distribution via Gauss's Principle. *The Scientific World Journal* 2013, DOI: 10.1155/2013/468418.

[42] IAEA. *Cytogenetic dosimetry: applications in preparedness for and response to radiation emergencies*. International Atomic Energy Agency, 2011.

[43] ISO. *Radiation protection performance criteria for service laboratories performing biological dosimetry by cytogenetics*. International Organization for Standardization, 2004.

[44] Johnson NL, Kemp AW, Kotz S. *Univariate discrete distributions*, 3rd edition. John Wiley & Sons, 2005.

[45] Kass RE, Raftery AE. Bayes Factors. *Journal of the American Statistical Association* 1995; **90**(430):773–795.

[46] de Keizer B, Hoekstra A, Konijnenberg MW, de Vos F, Lambert B, van Rijk PP, Lips CJM, de Klerk JMH. Bone marrow dosimetry and safety of high 131i activities given after recombinant human thyroid–stimulating hormone to treat metastatic differentiated thyroid cancer. *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine* 2004; **45**:1549–1554.

[47] Kemp CD, Kemp AW. Some Properties of the 'Hermite' Distribution. *Biometrika* 1965; **52**:381–394.

[48] Kemp AW, Kemp CD. An alternative derivation of the Hermite distribution. *Biometrika* 1966; **53**:627–628.

[49] Kottas A, Branco MD, Gelfand AE. A nonparametric Bayesian modelling approach for cytogenetic dosimetry. *Biometrics* 2002; **58**:593–600.

[50] Krnjajic M, Kottas A, Draper D. Parametric and nonparametric Bayesian model specification, A case study involving models for count data. *Computational Statistics & Data Analysis* 2008; **52**:2110–2128.

[51] Lloyd DC, Edwards AA, Prosser JS, Corp MJ. The dose response relationship obtained at constant irradiation times for the induction of chromosome aberrations in human lymphocytes by Cobalt–60 gamma rays. *Radiation and Environmental Biophysics* 1984; **23**:179–189.

[52] Lukacs E. *Characteristic Functions*. Hafner Publishing Company, 1970.

[53] Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325–337.

[54] Madruga MR, Pereira CAdB, Rabello–Gay M. Bayesian Dosimetry, Radiation dose versus frequencies of cells with aberrations. *Environmetrics* 1994; **5**:47–56.

[55] Madruga MR, Ochi–Lohmann TH, Okazaki K, Pereira CAdB, Rabello–Gay MN. Bayesian Dosimetry II, Credibility intervals for radiation dose. *Environmetrics* 1996; **7**:325–331.

[56] Matthews RAJ. Methods for assessing the credibility of clinical trial outcomes. *Drug Information Journal* 2001; **35**:1469–1478.

[57] Maznik NA, Vinnikov VA, Maznik VS. The distribution of individual radiation doses in liquidators of Chernobyl accident using cytogenetic analysis. *Radiatsionnaia biologiia, radioecologiia/Rossiĭskaia akademiia nauk* 2003; **43**:412–419.

[58] Maznik NA. Long–term follow–up cytogenetic survey and biological dosimetry in persons evacuated from 30–km Chernobyl NPP zone. *Radiatsionnaia biologiia, radioecologiia/Rossiĭskaia akademiia nauk* 2004; **44**:566–573.

[59] Maznik NA and Vinnikov VA. The retrospective cytogenetic Dosimetry using the results of conventional chromosomal analysis in Chernobyl cleanup workers. *Radiatsionnaia biologiia, radioecologiia/Rossiĭskaia akademiia nauk* 2005; **45**:700–708.

[60] McCullagh P, Nelder J. *Generalized Linear Models*, 2nd edition. Chapman & Hall/CRC, 1989.

[61] Merkle W. Statistical methods in regression and calibration analysis of chromosome aberration data. *Radiation and Environmental Biophysics* 1983; **21**:217–233.

[62] Miller G, Inkret WC, Martz HF. Bayesian detection analysis for radiation exposure. *Radiation Protection Dosimetry* 1993; **48**:251–256.

[63] Miller G, Inkret WC, Martz HF. Bayesian detection analysis for radiation exposure, II. *Radiation Protection Dosimetry* 1995; **58**:115–125.

[64] Morand J, Deperas–Standylo J, Urbanik W, Moss R, Hachem S, Sauerwein W, Wojcik A. Confidence limits for Neyman type–A–distributed events. *Radiation Protection Dosimetry* 2008; **128**:437–443.

[65] Moriña D, Higueras M, Puig P, Ainsbury EA, Rothkamm K. `radir` package: An `R` implementation for cytogenetic biodosimetry dose estimation. *Journal of Radiological Protection* 2015, DOI: 10.1088/0952-4746/35/3/557.

[66] Moriña D, Higueras M, Puig P, Oliveira M. Generalized Hermite distribution modelling with the R package hermite. *The R Journal*, submitted.

[67] Mukhopadhyay S. Bayesian nonparametric inference on the dose level with specified response rate. *Biometrics* 2000; **56**:220–226.

[68] Nelson SJ. A stochastic model of the effects of ionizing radiation on mammalian cells *in vitro*. *Bulletin of Mathematical Biology* 1984; **46**:423–446.

[69] Neyman J. On a new class of 'contagious' distribution, applicable in entomology and bacteriology. *The Annals of Mathematical Statistics* 1939; **10**:35–55.

[70] Oliveira M, Ainsbury EA, Einbeck J, Higueras M, Rothkamm K, Puig P. Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study. *Biometrical Journal*, accepted under revision 30/06/2015.

[71] Pace L, Salvan A. *Principles of statistical inference from a neo–Fisherian perspective*. World Scientific Publishing, 1997.

[72] Panjer HH. Recursive evaluation of a family of compound distributions. *ASTIN Bulletin* 1981; **12**:22–26.

[73] Pereira CAdB, Stern JM. Evidence and credibility, full Bayesian significance test for precise hypotheses. *Entropy* 1999; **1**:69–80.

[74] Pereira CAdB, Stern JM. Model selection, full Bayesian approach. *Environmetrics* 2001; **12**:559–568.

[75] Piper J, Sprey J. Adaptive classifiers for dicentric chromosomes. *Journal of Radiation Research* 1992; **33**:159–70.

[76] Puig P. Characterizing additively closed discrete models by a property of their maximum likelihood estimators, with an application to generalized hermite distributions. *Journal of the American Statistical Association* 2003; **98**(463):687-692.

[77] Puig P, Valero J. Count data distributions: some characterizations with applications. *Journal of the American Statistical Association* 2006; **101**:332–340.

[78] Puig P, Barquinero JF. An application of compound Poisson modelling to biological dosimetry. *Proceedings of the Royal Society A* 2011; **467**:897–910.

[79] Pujol M, Barquinero JF, Puig P, Puig R, Caballín MR, and Barrios L. A New Model of Biodosimetry to Integrate Low and High Doses. *PLOS ONE*, DOI: 10.1371/journal.pone.0114137 (2014).

[80] Rao CR, Chakravarti IM. Some small sample tests of significance for a Poisson distribution. *Biometrics* 1956; **12**:264–282.

[81] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, 2014.

[82] Romm H, Ainsbury EA, Barnard S, Barrios L, Barquinero JF, Beinke C, Deperas M, et al. Automatic scoring of dicentric chromosomes as a tool in large scale radiation accidents. *Mutation Research-Genetic Toxicolgy and Enviromental Mutagenesis* 2013; **756**:174–183.

[83] Sasaki MS. Chromosomal biodosimetry by unfolding a mixed Poisson distribution: a generalized model. *International Journal of Radiation Biology* 2003; **79**:83-97.

[84] Self SG, Liang K-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 1987; **82**(398):605–610.

[85] Serfling RJ. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 1980.

[86] Serna A, Alcaraz M, Navarro JL, Acevedo C, Vicente V, Canteras M. Biological dosimetry and Bayesian analysis of chromosomal damage in thyroid cancer patients. *Radiation Protection Dosimetry* 2008; **129**:372–380.

[87] Siebert BR. Uncertainty in radiation dosimetry, basic concepts and methods. *Radiation Protection Dosimetry* 2006; **121**:3–11.

[88] Sivia S, Skilling J. *Data analysis: a Bayesian tutorial*. Oxford University Press, 2006.

[89] Straume T, Bender MA. Issues in cytogenetic biological dosimetry, emphasis on radiation environments in space. *Radiation Research* 1997; **148**:60–70.

[90] Stiratelli RG, McCarthy KL, Scribner HE. Parametric approaches to the analysis of in vivo cytogenetics studies. *Environmental and Molecular Mutagenesis* 1985; **7**:43–54.

[91] Suto Y, Hirai M, Akiyama M, Kobashi G, Itokawa M, Akashi M, Sugiura N. Biodosimetry of restoration workers for the Tokyo Electric Power Company (TEPCO) Fukushima Daiichi nuclear power station accident. *Health Physics* 2013; **105**:366–373.

[92] Szluinska M, Edwards A, Lloyd D. Presenting statistical uncertainty on cytogenetic dose estimates. *Radiation Protection Dosimetry* 2006; **123**:443–449.

[93] van Dijk JW. Uncertainties in personal dosimetry for external radiation, a Monte Carlo approach. *Radiation Protection Dosimetry* 2006; **121**:31–39.

[94] Vinnikov VA, Ainsbury EA, Maznyk NA, Lloyd DC, Rothkamm K. Limitations associated with analysis of cytogenetic data for biological dosimetry. *Radiation Research* 2010; **174**:403–414.

[95] Vinnikov VA, Maznyk NA. Cytogenetic dose–response *in vitro* for biological dosimetry after exposure to high doses of gamma–rays. *Radiation Protection Dosimetry* 2013; **154**(2):186–197.

[96] Virsik RP, Harder D. Statistical Interpretation of overdispersed distribution of radiation-induced dicentric chromosome aberrations at high LET. *Radiation Research* 1981; **85**:13–23.

[97] Voisin P, Roy L, Hone PA, Edwards AA, Lloyd DC, Stephan G, Romm H, Groer PG, Brame R. Criticality accident dosimetry by chromosomal analysis. *Radiation Protection Dosimetry* 2004; **110**:443–447.

[98] Wang H, Liu Q, Wan D, Xiang J, Du L, Wang Y, Cao J, Fu Y, Fan F, Hecker M. BioDoser: improved dose-estimation software for biological radiation dosimetry. *Computer Methods and Programs in Biomedicine* 2012; **108**:402–406.

[99] Weise K, Hbel K, Rose E, Schlger M, Schrammel D, Tschner M, Michel R. Bayesian decision threshold, detection limit and confidence limits in ionising-radiation measurement. *Radiation Protection Dosimetry* 2006; **121**:52–63.