

AN EXACT GOODNESS-OF-FIT TEST BASED ON THE OCCUPANCY PROBLEMS TO STUDY ZERO-INFLATION AND ZERO-DEFLATION IN BIOLOGICAL DOSIMETRY DATA

Amanda Fernández-Fontelo ^{a,*}, Pedro Puig ^a, Elizabeth A. Ainsbury ^b, Manuel Higuera ^{c,d}

^aDepartament de Matemàtiques, Universitat Autònoma de Barcelona, 08193, Bellaterra (Barcelona), Spain.

^b Public Health England, Centre for Radiation, Chemical and Environmental Hazards, Chilton, Didcot, Oxon OX11 0RQ, UK.

^c BCAM - Basque Center for Applied Mathematics, 48009, Bilbao, Spain.

^d Institute of Health and Society, Newcastle University, Newcastle upon Tyne NE1 4LP, UK.

Received on April 2, 2003, revised on November 19, 2003, accepted on December 22, 2003

The goal in biological dosimetry is to estimate the dose of radiation that a suspected irradiated individual has received. For that, the analysis of aberrations (most commonly dicentric chromosome aberrations) in scored cells is performed and dose response calibration curves are built. In whole body irradiation with X and gamma-rays, the number of aberrations in samples is properly described by the Poisson distribution, although in partial body irradiation the excess of zeros provided by the non-irradiated cells leads, for instance, to the Zero-Inflated Poisson distribution. Different methods are used to analyse the dosimetry data taking into account the distribution of the sample. In order to test the Poisson distribution against the Zero-Inflated Poisson distribution, several asymptotic and exact methods have been proposed which are focused on the dispersion of the data. In this work we suggest an exact test for the Poisson distribution focused on the zero-inflation of the data developed by Rao and Chakravarti (1956), derived from the problems of occupancy. An approximation based on the standard Normal distribution is proposed in those cases where the computation of the exact test can be tedious. A Monte Carlo Simulation study was performed in order to estimate empirical confidence levels and powers of the exact test and other tests proposed in the literature. Different examples of applications based on *in vitro* data and also data recorded in several radiation accidents are presented and discussed. A Shiny application which computes the exact test and other interesting goodness-of-fit test for the Poisson distribution is presented in order to bring them to all the researchers.

INTRODUCTION

The goal in biological dosimetry is to estimate the dose of ionising radiation absorbed by an overexposed individual by using chromosome damage in peripheral lymphocytes as a biomarker of exposure. When radiation exposure occurs (as X-rays or gamma-rays, for instance), damage in DNA is randomly distributed between cells producing DNA single and double strand breaks which can be repaired or misrepaired to form aberrations such as dicentrics, centric rings, acentrics, translocations, among others. Dicentrics are the most used biomarkers of radiation exposure because they are reliable and well-established in certain well defined exposure scenarios (IAEA, 2011). The radiation dose that an overexposed individual has received is estimated by means of the dose response calibration curve which is created by exposures of human blood cells to different and appropriate doses of radiation. Full details on how these curves are created and calculated are in the manual of the International Atomic Energy Agency (IAEA, 2011).

The endpoint of interest in this case is the number of aberrations (generally dicentrics) observed. Commonly, in whole body irradiation under low-LET (Linear

Energy Transfer) radiation exposures, the number of recorded dicentrics per cell follows a Poisson distribution whose rate depends on the dose considering a linear quadratic function with identity link (IAEA, 2011 and Oliveira et al., 2016). However, in partial body irradiation which occurs when a fraction of the body is exposed to an homogeneous radiation dose, the Poisson distribution does not make sense (Oliveira et al., 2016, Higuera et al., 2016 and Vinnikov et al., 2010). In fact, the number of aberrations recorded in partial body irradiation can be explained as a mixture of a Poisson distribution and structural zeros (Oliveira et al., 2016, Higuera et al., 2016 and Vinnikov et al., 2010). In other words, the distribution of the aberrations of the non-irradiated scored cells provides an excess of zeros comparing with the distribution of those aberrations produced in an homogeneous whole body irradiation. It leads to a mixture of a Poisson and structural zeros which is called Zero Inflated (or modified) Poisson (ZIP) distribution. The probability mass function of a ZIP distribution is:

$$P(X = x) = \begin{cases} \omega + (1 - \omega)e^{-\theta} & x = 0 \\ \frac{(1-\omega)e^{-\theta}\theta^x}{x!} & x = 1, 2, \dots \end{cases} \quad (1)$$

*to whom correspondence should be addressed

denoting X as the number of aberrations, the parameter $0 < \omega < 1$ is the proportion of extra zeros and the parameter θ is the intensity of the underlying Poisson distribution. Here, ω represents the proportion of non irradiated scored cells in the sample and θ is the yield of chromosomal aberrations in the irradiated fraction. More about the ZIP distribution can be found in Johnson, Kemp and Kotz (2005).

According to the manual of the IAEA, the standard procedure to detect partial body irradiation uses the well-known u-test to determine whether the ratio of the sample variance to the sample mean (the sample dispersion index) is significantly different from 1 (the Poisson assumption holds that the mean and the variance are equal). Dolphin's method (Dolphin, 1969) is recommended in the manual of the IAEA to estimate both the absorbed dose and the fraction of the body irradiated, based on the maximum likelihood estimation of (1). However, other approaches have been proposed including the Bayesian method introduced by Higuera *et al.* (2016), and Sasaki and Miyata (1968) introducing the Qdr method. In addition, some other well-known statistical tests can be appropriate to check whether the data come from a Poisson against ZIP distribution, such as the likelihood ratio test (LRT), the Wald test and the score test (van den Broek 1995). All these tests are appropriate for large samples. An exact version of the u-test when dealing with small samples is also known (Rao and Chakravarti, 1956).

The main limitation of the u-test is that it is just focused on the dispersion of the data, detecting partial body irradiation when the data is overdispersed (sample dispersion index greater than 1). However, other characteristics of the data such as the zero-inflation can also lead with the rejection of the hypothesis of whole body irradiation, in cases where this is not rejected when just the u-test is applied. Accordingly, in this paper we suggest to use an exact zero-inflation or zero-deflation test for the Poisson distribution in biological dosimetry. This test, that we term the CR-test, was firstly proposed by Rao and Chakravarti (1956) based on the theory of occupancy problems. We also show an approximation of this test for large samples based on the standardised Normal distribution (based-Normal CR-test).

As commented above, data from individuals exposed to PBI can be overdispersed. Conversely, underdispersed data are unusual, and their observation is sometimes indicative that something during the experiment did not work (IAEA, 2011), or perhaps due to other mechanisms currently not very well-known (Pujol *et al.*, 2014). However, there are other possible causes producing overdispersion. For instance, this is the case for whole body low-LET-irradiation from different doses (heterogeneous exposures), which can be modelled using Mixed-Poisson distributions (Sasaki, 2003, Pujol *et al.*, 2016). It is important to point out that whole body high-LET-irradiation also produces

overdispersion, which can be modelled by Compound-Poisson distributions (Puig and Barquero, 2011). In the same way, zero-inflated data are also found in these scenarios. In fact, it can be mathematically shown that Mixed-Poisson and Compound-Poisson distributions are always zero-inflated with respect Poisson. In order to measure the degree of zero-inflation, the index z_i provided by Puig and Valero (2006) can be used. For a count variable X it is defined as $z_i = 1 + \log(p_0)/\mu$ where p_0 is the probability that X is equal to 0 and μ is its expectation (population mean). Naturally, the sample version of the index is obtained by replacing p_0 by the relative frequency of zeros in the sample f_0 , and μ by the sample mean \bar{x} . When X is Poisson distributed then $z_i = 0$ and $z_i > 0$ ($z_i < 0$) means that X is zero-inflated (zero-deflated).

In order to demonstrate the usefulness of the suggested CR-test and show its applicability, three examples will be discussed. The first one is based on an *in vitro* data from Sasaki (2003). The second concerns the data obtained from one of the patients heavily affected in the Tokaimura criticality accident. The last one is based on the data from one of the overexposed individuals in the accident in Stamboliyski (Bulgaria) in an industrial sterilization unit.

This paper is organised as follows. In Section 2 we review the theoretical aspects of the CR-test in terms of the occupational problems, and its usefulness in biological dosimetry. Also, in this section, an approximation of the CR-test for large samples is proposed and studied. Section 3 presents a Monte Carlo simulation study for exploring the empirical confidence levels and powers of several goodness-of-fit tests for the Poisson distribution, included those proposed in this work. In Section 4 the results of some examples of application, an *in vitro* experiment and some real radiation accidents, are shown and discussed. Also, a Shiny application for computing the CR-test in an easy way is presented. Finally, Section 6 is dedicated to discussion of the main results of the work.

CR-TEST FOR THE POISSON DISTRIBUTION

The CR-test of goodness-of-fit for the Poisson distribution, focused on the properties of zero-inflation and zero-deflation, was firstly presented by Rao and Chakravarti (1956) on the basis of a problem related to occupancy distributions. Previously, Fisher (1950) provided an exact goodness-of-fit test by considering the conditional probability of the observations given their total (the statistic containing all the information of the Poisson parameter θ in the sample called the *sufficient statistic*), and Rao and Chakravarti (1956) decided to follow-up that work by introducing the test presented in this paper. They considered an experiment with n boxes and S equiprobable balls, in the sense that each box has the same probability to contain a ball, and they asked for the probability of getting n_0 empty boxes.

Let the random variable X_j be the number of balls in the box j . The probability mass function (pmf) of $X = (X_1, X_2, \dots, X_n)$ behind this experiment, is that of the well-known Multinomial distribution with parameters $(1/n, 1/n, \dots, 1/n, S)$. In this sense, Fisher (1950) introduced this Multinomial distribution by supposing a sample $X = (X_1, X_2, \dots, X_n)$ distributed as a $\text{Poisson}(\theta)$, and $S = \sum_{i=1}^n X_i$ (sufficient statistic). The distribution of X conditioned to this sufficient statistic $S(X|S)$ is just the Multinomial distribution with the above parameters. Full details can be found in Casella and Berger (2002).

As previously mentioned, one of the main interests in the previous experiment consists on study the random variable N_0 which represents the number of empty boxes. This problem has been well studied in the literature by many renowned authors such as Tukey (1949), Feller (1957) and Kemp (1978). In addition, other authors have proposed interesting extensions of this classical problem like Parzen (1960), who dealt with the problem of specified occupancy problems, or Johnson and Kotz (1977) who studied randomised occupancy problems.

Focusing here on the classical problem, Rao and Chakravarti (1956) duly studied the distribution of N_0 resulting in the following expression, for the probability of observing exactly n_0 empty boxes:

$$P(N_0 = n_0) = \sum_{j=n_0}^n (-1)^{j-n_0} \binom{n}{j} \binom{j}{n_0} \left(1 - \frac{j}{n}\right)^S, \quad (2)$$

Here n is the number of boxes and S is the number of balls. Taking into account the expression (2), it is straightforward to compute the probability of observing more than n_0 empty boxes (or less than n_0 empty boxes) by summing all the corresponding probabilities, that is, $P(N_0 > n_0) = \sum_{i=n_0+1}^n P(N_0 = i)$. This is the exceedance function (also called survival or reliability function) of the random variable N_0 . On the other hand, the cumulative function of N_0 , that is, the probability of observing at most n_0 empty boxes in the previous experiment, is easily computed since $P(N_0 \leq n_0) = 1 - P(N_0 > n_0)$. Full details related to this distribution such as its moments and factorial moments, its probability generating function (pgf), among others, can be found in Johnson, Kemp and Kotz (2005).

Asymptotic approach for large values

Although $P(N_0 > n_0)$ can be computed in all cases, as n and S increase its computation is more tedious since the factorial numbers intensify the calculation of the expression (2). A first and naïve solution for solving the computational problems consists on applying logarithms to the corresponding expression of the probabilities to make more efficient its computation. However, in many cases this solution is not enough, and then some approximations need to be used. One of these approximations is based on the standard Normal distribution which can also be found in the work of Rao

and Chakravarti (1956).

Let ϵ_i be an indicator taking the value 1 when the i -box is empty, and 0 otherwise. Because ϵ_i is a Bernoulli random variable it is immediate to see that, $E(\epsilon_i) = E(\epsilon_i^2) = P(\epsilon_i = 1) = \left(\frac{n-1}{n}\right)^S$. Moreover, some algebra shows that $E(\epsilon_i, \epsilon_j) = \left(\frac{n-2}{n}\right)^S$ when $i \neq j$. Notice that $E(\epsilon_i, \epsilon_j) \neq E(\epsilon_i)E(\epsilon_j)$. Because $N_0 = \sum_{i=1}^n \epsilon_i$, the expectation and variance of N_0 can be directly computed as follows:

$$E(N_0) = E\left(\sum_{i=1}^n \epsilon_i\right) = \sum_{i=1}^n E(\epsilon_i) = n \left(\frac{n-1}{n}\right)^S. \quad (3)$$

$$\begin{aligned} V(N_0) &= \sum_{i=1}^n V(\epsilon_i) + \sum_{i \neq j} \text{Cov}(\epsilon_i, \epsilon_j) \\ &= \sum_{i=1}^n \left(E(\epsilon_i^2) - E(\epsilon_i)^2\right) \\ &\quad + \sum_{i \neq j} \left(E(\epsilon_i \epsilon_j) - E(\epsilon_i)E(\epsilon_j)\right) = n(n-1) \left(\frac{n-2}{n}\right)^S + \\ &\quad + n \left(\frac{n-1}{n}\right)^S \left(1 - n \left(\frac{n-1}{n}\right)^S\right). \end{aligned} \quad (4)$$

Additionally, it is fairly well-known that the normalised version of N_0 , that is, $Z_0 = (N_0 - E(N_0))/\sqrt{V(N_0)}$ follows approximately a standard Normal distribution. Based on the expressions (3) and (4),

$$P(N_0 > n_0) \approx 1 - \Phi(z_0), \quad (5)$$

where $\Phi(z_0)$ represents the cumulative distribution function of the standard Normal evaluated at $z_0 = (n_0 - E(N_0))/\sqrt{V(N_0)}$.

Testing zero-inflation or zero-deflation

Given a sample $X = (X_1, X_2, \dots, X_n)$, of independent Poisson counts, let n_0 be the number of observed zeros. The CR-test to contrast the null hypothesis H_0 : *Data are Poisson distributed*, against the alternative H_1 : *Data are zero-inflated (zero-deflated)* consists in calculating the p-value $P(N_0 \geq n_0) (P(N_0 \leq n_0))$ using expressions (2) or (5), rejecting H_0 when the p-value is lower than a suitable significance level, usually established as $\alpha = 0.05$. Concretely, the exact p-value of the CR-test for zero-inflation is computed as follows:

$$\begin{aligned} P(N_0 \geq n_0) &= \sum_{i=n_0}^n \sum_{j=i}^n (-1)^{j-i} \binom{n}{j} \binom{j}{i} \left(1 - \frac{j}{n}\right)^S \\ &= \sum_{i=n_0}^n (-1)^{i-n_0} \frac{1}{(n_0-1)!(j-n_0)!} \binom{n}{i} \left(1 - \frac{i}{n}\right)^S. \end{aligned} \quad (6)$$

Similarly, the exact p-value of the test for zero-deflation is the following:

$$\begin{aligned}
 P(N_0 \leq n_0) &= \sum_{i=0}^{n_0} \sum_{j=i}^n (-1)^{j-i} \binom{n}{j} \binom{j}{i} \left(1 - \frac{j}{n}\right)^S \\
 &= 1 - \sum_{i=(n_0+1)}^n (-1)^{i-n_0} \frac{1}{(n_0-1)!(j-n_0)!} \binom{n}{i} \left(1 - \frac{i}{n}\right)^S
 \end{aligned}
 \tag{7}$$

On biological dosimetry data

The ideas behind the previous occupancy problem can be easily translated to the biological dosimetry framework. Accordingly, suppose that $X = (X_1, X_2, \dots, X_n)$ is a dosimetry sample where X_j represents the number of aberrations (usually dicentrics and/or rings) found in the cell j for $j = 1, \dots, n$. Then, $S = \sum_{j=1}^n X_j$ is the total number of aberrations found in the sample. Aberrations are randomly distributed between cells with the same probability $1/n$.

As commented above, the features of the sampling distribution allows estimation of the dose of radiation received by an exposed individual and the type of exposure. Thus, deviations from the Poisson distribution can lead, for instance, to rejection of the hypothesis of whole body irradiation in low-LET exposures. One of these departures comes from the excess number of cells free of aberrations (N_0), which are frequently found in partial body irradiation. Therefore, as has been previously noted, the variable N_0 and the associated index z_i , seem to be suitable measures for exploring whether the data come from a Poisson distribution.

The reason why the CR-test is suitable to test the Poisson distribution is because it allows the researchers to evaluate and detect exactly the possible problem of zero-inflation (or zero-deflation) in the data, which it cannot be detected using only the u-test (just testing overdispersion and underdispersion), score-test, Wald test or Likelihood Ratio Test (LRT). Indeed, dispersion problems in the data (overdispersion and underdispersion) can finally mean the rejection of the Poisson distribution. It is important to remark that the CR-test can be seen as to complement the widely used u-test for dispersion, being useful in scenarios where the dispersion is not significant but the number of zeros is anomalous.

SIMULATION RESULTS

A Monte Carlo simulation study was performed in order to describe the empirical confidence levels $\hat{P}(\text{not reject } H_0 | H_0 \text{ is true}) = 1 - \hat{\alpha}$, and the empirical powers $\hat{P}(\text{reject } H_0 | H_1 \text{ is true}) = 1 - \hat{\beta}$ of CR, and the asymptotic tests explored in this work under different scenarios. Notice that the empirical confidence level of

the CR-test, assuming a significance level α , is just $1 - \alpha$ since this is exactly the p-value. The simulations only consider alternatives ZIP-distributed (1) because this is the most suitable distribution for partial body low-LET radiation exposures.

Several values for $\omega : 0, 0.01, 0.25, 0.4, 0.5, 0.75, 0.90$, sample size $n : 50, 100, 200, 300$ and $\theta : 0.1, 0.5, 1, 2$, have been used, simulating for each combination $M = 2000$ replicates. A true significance level of $\alpha = 0.05$ were considered.

In those cases where $\omega = 0$ (Poisson samples) the best test is the one in which the value of the empirical confidence level is the closest to the real value (0.95). Otherwise, a confidence level greater than the real level means that the corresponding test is conservative, while a lower value indicates a greater type I error than α . In the other cases where $\omega > 0$ (zero-inflated Poisson samples), the best test is the one with the greatest empirical power since we are interested in powerful tests.

Table 1 shows the empirical confidence levels and powers resulting from the Monte Carlo simulation study performed to compare those tests introduced in this work (the CR-test and its approximated version), the u-test and the score test. As expected, as larger are sample sizes n and the intensities of the underlying Poisson distribution θ , confidence levels are closer to the real value (0.95), and also the tests seem to be more powerful.

Specifically, when $\omega = 0$ (Poisson samples), the best test is the CR-test since their confidence levels are always 0.95. However, the widely used u-test shows inferior results in all simulated scenarios since all of their empirical confidences levels are lower than 0.95. This last means that we are observing more type I error than that expected (0.05). The based-Normal CR-test shows results not as good as those from the CR-test, especially when θ is really small. The results from the score test are also good in all simulation scenarios. It happens because the score test is specially designed to test Poisson distribution against a ZIP alternative.

This Monte Carlo simulation study was performed in R for the u-test, the score test and also for the based-Normal CR-test. However, the simulations for the CR-test were computed using the software `Maxima`, since it is computationally intensive. The corresponding syntax codes are available as supplemental material.

EXAMPLES OF APPLICATION

In this section three examples are presented in which the results based on the classical u-test and the CR-test suggested in this work lead to different conclusions.

In vitro irradiation

Sasaki (2003) proposed a method to estimate the dose over a wide dose-range, not only limited to low or

EXACT TEST FOR POISSON DISTRIBUTION

n	ω	θ	u-test				score test				based-Normal CR-test				CR-test			
			.1	.5	1	2	.1	.5	1	2	.1	.5	1	2	.1	.5	1	2
50	0		.916	.939	.934	.939	.952	.955	.950	.947	.912	.944	.950	.934	.950	.950	.950	.950
	.10		.092	.096	.140	.261	.057	.050	.080	.242	.094	.091	.151	.372	.012	.049	.108	.284
	.25		.116	.182	.332	.728	.080	.098	.240	.792	.118	.186	.386	.891	.020	.100	.288	.829
	.40		.109	.266	.542	.944	.072	.178	.416	.971	.110	.280	.620	.991	.021	.142	.530	.984
	.50		.132	.333	.684	.989	.098	.249	.562	.990	.132	.362	.733	.998	.019	.204	.630	.995
	.90		.238	.538	.788	.978	.235	.485	.739	.968	.238	.539	.790	.978	.041	.183	.511	.891
100	0		.922	.944	.936	.936	.953	.953	.952	.948	.923	.954	.952	.951	.950	.950	.950	.950
	.10		.088	.120	.20	.404	.056	.068	.120	.440	.086	.111	.211	.584	.017	.086	.160	.538
	.25		.123	.240	.492	.923	.75	.146	.419	.974	.124	.229	.571	.991	.031	.172	.509	.988
	.40		.152	.392	.799	.999	.112	.272	.764	1.000	.152	.383	.856	1.000	.025	.278	.798	.999
	.50		.158	.474	.916	1.000	.122	.368	.888	1.000	.159	.493	.939	1.000	.032	.353	.898	1.000
	.90		.154	.629	.970	1.000	.148	.552	.956	1.000	.155	.632	.976	1.000	.049	.446	.936	1.000
200	0		.940	.940	.946	.944	.962	.948	.950	.942	.929	.950	.950	.945	.950	.950	.950	.950
	.10		.078	.130	.271	.604	.052	.064	.192	.747	.092	.120	.306	.842	.028	.108	.279	.808
	.25		.096	.348	.742	.998	.068	.240	.732	1.000	.118	.360	.830	1.000	.032	.304	.792	1.000
	.40		.142	.577	.964	1.000	.111	.472	.966	1.000	.171	.612	.985	1.000	.054	.506	.983	1.000
	.50		.164	.702	.990	1.000	.128	.612	.990	1.000	.206	.736	.996	1.000	.062	.631	.995	1.000
	.90		.184	.857	.998	1.000	.168	.814	.998	1.000	.194	.868	.999	1.000	.093	.736	.999	1.000
300	0		.928	.928	.940	.941	.958	.945	.948	.946	.930	.940	.940	.943	.950	.950	.950	.950
	.10		.084	.164	.348	.750	.045	.108	.282	.879	.088	.170	.410	.942	.031	.127	.348	.931
	.25		.116	.425	.883	.999	.070	.317	.893	1.000	.130	.452	.946	1.000	.052	.419	.929	1.000
	.40		.184	.724	.996	1.000	.126	.636	.997	1.000	.198	.751	.998	1.000	.062	.679	.999	1.000
	.50		.198	.826	.999	1.000	.138	.774	1.000	1.000	.206	.844	1.000	1.000	.081	.813	1.000	1.000
	.90		.286	.928	1.000	1.000	.242	.904	1.000	1.000	.286	.936	1.000	1.000	.111	.878	1.000	1.000
			.166	.889	.999	1.000	.166	.850	.998	1.000	.166	.890	.999	1.000	.117	.763	.996	1.000

Table 1. Results of Monte Carlo simulation study based on different scenarios. Bold lines represent the empirical confidence levels of every test according to different values of n, ω and θ , assuming theoretically a confidence level of 0.95. Normal lines represent the empirical powers of every test according to different values of n, ω and θ .

moderate low dose. In order to exemplify his method, an *in vitro* experiment was performed covering a wide dose-range. Two types of simulated *in vitro* experiment were taken into account using ^{137}Cs gamma rays. Full details can be found in Sasaki (2003).

Table 2 shows the results obtained using the u-test and the CR-test for one of the higher doses data sets included in the Table 1 of Sasaki (2003). According to the result of the u-test (Table 2), the conclusion would be that the sample is under-dispersed. However, based on the results of the CR-test (Table 2), the sample seems not to be zero-deflated and, of course, not zero-inflated. Pujol *et al.* (2014) also detected underdispersion for higher doses, using weighted-Poisson distributions for fitting their data.

Tokaimura criticality accident

In 1999 in Tokaimura (Japan) three individuals received high doses of radiation in a small plant when they were preparing a small batch of fuel for an experimental

reactor. It was the first batch of fuel for that reactor in three years, so these workers were not qualified neither well prepared for the work. Unfortunately, they brought together too much uranium enriched to a quite high level, causing a limited uncontrolled nuclear chain reaction (criticality), which continued for 20 hours. A total of 119 people received a relatively important dose, but only three operators received doses above permissible limits. Two of them died hours or days later to the accident. After several analysis, the conclusion was that they apparently received whole body irradiation. Hayata *et al.* (2001) explained that in non-uniform exposures, the chromosome aberration population follows a mixture of Poisson distributions since each subpopulation was exposed to different dose (in PBI the population is a mixture of a Poisson plus structural zeros). They studied the dose distribution of dicentric plus ring aberrations scored from the blood of the three workers involved in the accident, suspecting that the distribution of two of them was a mixture

dose	n	S	n_0	mean	variance	z_i	statistic u	CR p-value	u p-value
10	200	705	3	3.525	2.924	-.191	-1.702	.147	.044

Table 2. Results of one sample of Sasaki data which can be found in Sasaki (2003).

of two Poisson distributions since they were very heterogeneous.

Here, we focus on the cell-aberration distribution of one of these workers. As the mean number of cell aberrations was 3.069 and the variance was 3.214, the sample dispersion index was a little bit greater than 1 (Poisson distribution). Accordingly, the u-test and the CR-test were performed considering a right-tailed test. Based on the results in Table 3, the cell-aberrations distribution of this worker does not seem overdispersed (p-value=0.329), but it seems zero-inflated (p-value=0.022).

Bulgarian accident

In 2011 in Stamboliyski (Bulgaria) five people suffered an important radiation accident in an industrial sterilisation facility used to sterilise equipment, food and other agricultural products. They were irradiated with a synthetic radioactive isotope of cobalt (^{60}Co) from a subterranean shielded area where ^{60}Co was contained in up to six cylindrical vertical tubes. After the accident, blood samples were taken from the five patients and were sent to the Institut de Radioprotection et de Sûreté Nucléaire (IRSN, Paris), and also to the National Centre of Radiobiology and Radiation Protection (NCRRP, Bulgaria). In addition, the NCRRP sent electronic images of 22 metaphases from the individual 1 to the IRSN laboratory for prior evaluation.

Grégorie *et al.* (2013) presented the cytogenetic data and the estimated doses for all the five patient based on the information send to NCRRP and IRSN laboratories. They concluded that all the patients were totally irradiated, providing 95% confidence levels of the estimated doses for each individual. However, some contradictory results were found for the individual 1. In fact, in both analysis (blood samples and images) performed by the IRSN the conclusion was that the individual was totally irradiated, while the analysis performed in NCRRP concluded that the individual 1 was partially irradiated (PBI).

The aim of this example is to explore the data of individual 1 using the CR-test, and compare with the results reported in Grégorie *et al.* (2013) which were performed based on the classical analysis proposed by the IAEA manual.

Results in Table 4 show that sample 1^a was almost significantly underdispersed (p=0.056), and for 1^c , although it was slightly underdispersed, according the u-test the null hypothesis of Poissonity can not be rejected. Conversely, sample 1^b was significantly overdispersed (p=0.021). One tailed CR-tests have also

been performed, looking for zero-inflation or zero-deflation according to the sign of the z_i statistic. According to the results shown in Table 4, sample 1^a seems to be zero-deflated, while for the other samples the number of zeros are compatible with the Poisson assumption.

Shiny-based application

In order to make available the use of the CR-test to all researchers, a shiny application based on R language has been built. It is available through https://manu2h.shinyapps.io/gof_poisson/.

This is a user-friendly application which allows the user to compute the CR-test, and also other tests as the classical u or the score. In addition, this application computes a meaningful plot showing the probability mass function and the cumulative probability function of the number of zeros for a Poisson sample of size n and sample sum S . In more detail, the input of the application is the numerical frequencies of counts, from zero to the maximum of the sample (*i.e.* $\ll 100\ 10\ 5 \gg$), means that in the sample there are 100 zeros, 10 ones and 5 twos). According to the sample information, the Shiny application gives the user a basic summary of the sample (size, sum, mean and variance), the CR p-value, the u test statistic and p-value, the score test statistic and p-value and the Bayes factor testing Poisson vs. ZIP distributions (Higuera *et al.*, 2016). At least for now, the application computes one-tailed tests, looking for overdispersion or underdispersion and zero-inflation or zero-deflation accordingly with the signs of u and z_i . Additionally, the application gives the option to the researcher to present the previous results in table format, and also to compute the previous mentioned plot. Finally, some relevant references related to goodness-of-fit tests for the Poisson distribution and this current work are also available from the application.

DISCUSSION

The CR-test is presented in this paper as a tool for biological dosimetry research, for analysing goodness-of-fit in scored samples, being useful to identify exposure patterns (homogeneous or inhomogeneous) and low or high-LET radiation. Although some classical tests, as u and score tests, are used to investigate the most reliable distribution of chromosome aberrations, they are focused on the dispersion of the data (u-test) or they are designed for an specified alternative (score-test). In addition, they are build based on asymptotic theory and consequently they

EXACT TEST FOR POISSON DISTRIBUTION

n	S	n_0	mean	variance	z_i	statistic u	CR p-value	u p-value
175	537	14	3.069	3.214	.177	.441	.022	.329

Table 3. Results of scored blood samples from one individual accidentally irradiated in the Tokaimura accident. Data can be found in Hayata *et al.* (2001).

Case	n	S	n_0	mean	variance	z_i	statistic u	CR p-value	u p-value
1 ^a	22	31	2	1.409	.729	-.702	-1.589	.030	.056
1 ^b	284	272	117	.958	1.122	.074	2.042	.070	.021
1 ^c	64	100	13	1.563	1.488	-.020	-0.269	.544	.394

Table 4. Results of the individual I based on the number of dicentric plus centric rings. ^a: IRSN analysis based on transmitted images sent by NCRRP ^b: Data from NCRRP and ^c: Data from IRSN. Data are available in Grégoire *et al.* (2013).

provide approximated p-values. In general, for cell-aberrations data, the highest number of observations (highest frequency) corresponds to the value equal to zero. Consequently, it is important to decide whether or not the observed number of zeros is compatible with the Poisson assumption.

In this work the CR-test is proposed as a complement of the classical u-test, in order to study the possible zero-inflation and zero-deflation. This test provides exact p-values, providing also a normal approximation to be used when dealing with large sample sizes. The normal approximation is very precise when $n \geq 200$ and $S \geq 50$. Data coming from the accidents of Tokaimura (Japan, 1999) and Stambolyski (Bulgaria, 2011) are analyzed, highlighting the usefulness of this test.

The u-test is basic in the analysis of dosimetry data, being the test of reference in the manual of the IAEA, in part due to its simplicity and the fact that all the calculations can be done using a worksheet. However, the approximated CR-test (based-Normal CR-test) is also very simple and it can also be performed using a worksheet. We think that to implement this test in the IAEA manual would be a good suggestion.

REFERENCES

1. Casella, G. and Berger, R.L. (2002) *Statistical Inference*. The Wadsworth Group, California.
2. Dolphin, G.W. (1969). Biological dosimetry with particular reference to chromosome aberration analysis. A review of methods. Handling of Radiation Accidents (Proc. Int. Symp. Vienna), IAEA, Vienna, 215-224.
3. Feller, W. (1957). *An Introduction to Probability Theory and its Applications*. Wiley Series, USA.
4. Grégoire, E., Hadjidekova, V., Hristova, R., Gruel, FG., Roch-Lefevre, S., Voisin, P., Staynova, A., Deleva, S., Ainsbury, E.A., Lloyd, D.C. and Barquinero, J.F. (2013). Biological dosimetry assessments of a serious radiation accident in

5. Bulgaria in 2011. *Radiation Protection Dosimetry* **155(4)**, 418-422.
5. Hayata, I., Kanda, R., Minamihisamatsu, M., Furukawa, A. and Sasaki, M.S. (2001). In "Proceedings of the International Symposium on the Criticality Accident in Takaimura: Medical aspects of Radiation Emergency". Tsuji, H. and Akushi, M, eds., National Institute of Radiological Science, Chiba, Japan, 82-89.
6. Higuera, M., Puig, P., Ainsbury, E.A., Vinnikov, V.A. and Rothkamm, K. (2016). A new bayesian model applied to cytogenetic partial body irradiation estimation. *Radiation Protection Dosimetry* **168(3)**, 330-336.
7. IAEA (2011). *Cytogenetic Dosimetry: Applications in Preparedness for and Response to Radiation Emergencies*. International Atomic Energy Agency.
8. Johnson, N. L., Kemp, A. W. and Kotz, S. (2005). *Univariate Discrete Distributions*. Wiley Series in Probability and Statistics, New Jersey.
9. Kemp, A. W. (1978b). On probability generating functions for matching and occupancy distributions. *Zastosowania Matematyki* **16**, 207-213.
10. Nicholson, W. L. (1961). Occupancy probability distribution critical points. *Biometrika* **48**, 175-180.
11. Oliveira, M., Einbeck, J., Higuera, M., Ainsbury, E., Puig, P. and Rothkamm, K (2016). Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study. *Biometrical Journal* **58(2)**, 259-279.
12. Parzen, E. (1960). *Modern Probability Theory and Its Applications*. Wiley Series, USA.
13. Puig, P. and Valero, J. (2006). Count data distributions. *J. Am. Statist. Assoc.* **101**, 3323-340.
14. Puig, P. and Barquinero, J.F. (2011). An application of compound Poisson modelling to biological dosimetry. *Proceedings of The Royal Society A*

- Mathematical Physical and Engineering Sciences* **467(2127)**, 897-910.
15. Pujol M., Barquinero J.F., Puig P., Puig R., Caballín M.R., Barrios L. (2014). A New Model of Biodosimetry to Integrate Low and High Doses, *PlosOne*, 1-19, DOI: 10.1371/journal.pone.0114137.
 16. Pujol M., Barrios L., Puig P., Caballín M.R., Barquinero J.F. (2016). A new model for biological dose-assessment in cases of heterogeneous exposures to ionizing radiation. *Radiation Research*, **185(2)**, 151-162.
 17. Rao, C.R. and Chakravarti, I.M. (1956). Some small sample tests of significance for a Poisson distribution. *Biometrics* **12**, 264-282.
 18. Sasaki, M.S. and Miyata, H. (1968). Biological dosimetry in atom bomb survivors. *Nature* **220**, 1189-1193.
 19. Sasaki, M.S. (2003). Chromosomal biodosimetry by unfolding a mixed Poisson distribution: a generalized model. *Int.J.Radiat.Biol.* **79(2)**, 83-97.
 20. Tukey, J.W. (1949). Moments of random group size distributions. *Annals of Mathematical Statistics*, **20**, 523-539.
 21. van den Broek, J. (1995). A score test for zero inflation in a poisson distribution. *Biometrics* **51(2)**, 738-743.
 22. Vinnikov, V. A., Ainsbury, E., Maznyk, N. A. and Rothkamm, K. (2010). Limitations associated with analysis of cytogenetic data for biological dosimetry. *Radiation Research* **174(4)**, 403-414.