


# Some mechanisms leading to underdispersion: Old and new proposals

Pedro Puig<sup>1</sup>  | Jordi Valero<sup>2</sup> | Amanda Fernández-Fontelo<sup>3</sup>

<sup>1</sup>Departament de Matemàtiques, Universitat Autònoma de Barcelona, Centre de Recerca Matemàtica (CRM), Barcelona, Spain

<sup>2</sup>Escola Superior d'Agricultura de Barcelona, Universitat Politècnica de Catalunya, Barcelona, Spain

<sup>3</sup>School of Business and Economics, Humboldt University of Berlin, Berlin, Germany

## Correspondence

Pedro Puig, Departament de Matemàtiques, Universitat Autònoma de Barcelona, Centre de Recerca Matemàtica (CRM), Barcelona, Spain.  
Email: [ppuig@mat.uab.cat](mailto:ppuig@mat.uab.cat)

## Funding information

Agència Estatal de Investigació, Grant/Award Number: CEX2020-001084-M; Ministerio de Ciencia e Innovación, Grant/Award Number: PID2022-137414OB-I00

## Abstract

In statistical modeling, it is important to know the mechanisms that cause underdispersion. Several mechanisms that lead to underdispersed count distributions are revisited from new perspectives, and new ones are introduced. These include procedures based on the number of arrivals in arrival processes, such as renewal and pure birth processes and steady-state distributions of birth-death processes, like queues with state-dependent service rates. Weighted Poisson and other well-known underdispersed distributions are also related to birth-death processes. Classical and variable binomial thinning mechanisms are also viewed as important procedures for generating underdispersed distributions, which can also generate bivariate count distributions with negative correlation. Some example applications are shown, one of which is related to Biodosimetry.

## KEYWORDS

arrival process, binomial thinning, birth-death process, COM-Poisson distribution, count distributions, state-dependent service rates, weighted Poisson distribution

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Scandinavian Journal of Statistics* published by John Wiley & Sons Ltd on behalf of The Board of the Foundation of the Scandinavian Journal of Statistics.

## 1 | INTRODUCTION

The Poisson distribution is by far the most widely recognized and commonly used type of distribution for count data analysis. One of its key properties is that the population variance equals the population mean. However, in practice, many count data are overdispersed, that is, the variance is greater than the mean. Conversely, the case of underdispersion (variance lower than the mean) is also common but less frequent. In practice, dispersion is measured with the Fisher dispersion index, defined as the ratio of variance to mean. There have also been some attempts to generalize this index to multivariate count distributions (see Kokonendji & Puig, 2018).

Why does over(under)dispersion occur? Of course, there are many count distributions in the literature that are over or underdispersed, but we are not interested in these distributions themselves but in the mechanisms leading to them. There are many mechanisms explaining overdispersion. For instance, the mechanism of compounding (or clustering) leads to the large family of Compound-Poisson distributions. Similarly, the mechanism of mixing (or clustering) leads to the other large family of Mixed-Poisson distributions. Both mechanisms are understandable and meaningful in many practical situations.

In contrast, the mechanisms leading to underdispersion are few, not very well-known, and not always able to explain the cause of the underdispersion observed in a particular dataset. Consider for instance the biodosimetry data described in Pujol et al. (2014). The authors studied the number of a certain type of chromosome aberrations, called dicentrics, in samples of blood lymphocytes when exposed to several doses of X-rays. Table 1 shows the frequency of dicentrics for three different doses. Note that the variances are lower than their respective means. Underdispersion is an extraordinary occurrence in Biodosimetry, as explained by the manual of the International Atomic Energy Agency (IAEA) (IAEA, 2011) on p. 48: *Biologically, underdispersion is very unlikely to occur. (underdispersion) may be indicative of a problem in data sampling.*

However, Pujol et al. (2014) claimed that the underdispersion detected in their datasets was caused by a real and unexplained phenomenon, and was not due to a sampling problem. They used a weighted-Poisson distribution to fit their data as an empirical solution, but they were unable to explain why underdispersion occurred.

In this paper, we will revisit some previously considered mechanisms leading to underdispersion in the literature, as well as introduce some new ones. Section 2 reviews such mechanisms connected with arrival processes, in particular those based on renewal and pure-birth processes. These are the best known mechanisms although, as we shall see, some of their results have not been correctly understood. Procedures based on the steady state distributions of birth-death processes are analyzed in Section 3. We show that an almost unknown and forgotten proposition by Wise (1962) can explain many of the underdispersed count distributions studied in the past by several authors, including the COM-Poisson and weighted-Poisson distributions (see Sellers & Morris, 2017). Thinning mechanisms and their relation with underdispersion are analyzed in Section 4. In particular, Proposition 6 presents a very meaningful variable thinning mechanism leading to underdispersion. In addition, Section 4.2 explores the relationship between variable thinning, underdispersion and negative correlation. Finally, some conclusions, comments and further research are presented in Section 5.

TABLE 1 Frequencies of dicentrics among cells for several doses of rays.

Dose (grays)	Number of dicentrics									$\bar{x}$	$s^2$	
	0	1	2	3	4	5	6	7	8			9
3	213	192	85	9	1						0.786	0.641
5	3	23	58	38	15	10	2	1			2.547	1.578
7		4	23	35	35	29	10	9	4	1	4.027	2.697

## 2 | MECHANISMS BASED ON ARRIVAL PROCESSES

An arrival process is a sequence  $\{T_n\}$  of nonnegative and increasing random variables  $0 < T_1 < T_2 < \dots$ , representing the times when some phenomenon occurs, called arrival times. Interarrival times are defined as  $X_1 = T_1$ , and  $X_i = T_i - T_{i-1}$  for  $i > 1$ . The process starts at time 0 and, in principle, multiple arrivals cannot occur at the same time. Let  $N(t)$  be the number of arrivals in the interval  $(0, t]$ . This count random variable will be our source of interest. The literature on arrival processes and the distribution of  $N(t)$  is extensive; see, for example, Cox and Isham (1980) and Ross (1995). We are going to focus on two specific arrival processes that will be described in the following subsections.

### 2.1 | Renewal processes

A renewal process is an arrival process in which the interarrival times  $X_i$  are iid random variables. Let  $F$  be the distribution of  $X_0$ , and let  $F^{(k)}$  be the distribution of the  $k$ th arrival time  $T_k = X_0 + X_1 + \dots + X_{k-1}$ . It is clear that  $F^{(k)}$  is the  $k$ -fold convolution of  $F$  with itself. Consider  $N(t)$ , the number of arrivals in the interval  $(0, t]$ . Note that  $N(t) < k \Leftrightarrow T_k \geq t$ . Then,  $P(N(t) < n) = 1 - F^{(n)}(t)$ , and the probability function of the number of arrivals or events is,

$$P(N(t) = n) = P(N(t) < n + 1) - P(N(t) < n) = F^{(n)}(t) - F^{(n+1)}(t). \tag{1}$$

To be coherent with this notation we define  $F^{(0)}(t) = 1$  for  $t \geq 0$  and  $F^{(0)}(t) = 0$  otherwise. According to (1), the distribution of the interarrival times determines the probability function of the number of arrivals. On the other hand, knowledge of the probability function also determines the distribution of the interarrival times because  $P(N(t) = 0) = 1 - F(t)$ .

When the interarrival times are exponentially distributed, the number of arrivals follows a Poisson distribution and, in fact, this is an important and meaningful characterization of the Poisson process. Other distributions have been used for the interarrival times, like Gamma (Winkelmann, 1995; Zeviani et al., 2014), Weibull (McShane et al., 2008; Moriña et al., 2019) or Inverse Gaussian (Levine, 1991; and Seshadri, 1998). In general, if  $F$  belongs to a family of distributions closed under convolutions (like Gamma and Inverse Gaussian), then the expression for the probability function (1) is tractable.

The hazard function (or failure rate function) of an interarrival time distribution  $F$  is defined as,

$$r(t) = \frac{f(t)}{1 - F(t)},$$

where  $f(t)$  is the density of  $F$ . It is worth mentioning that the hazard function  $r(t)$ , as well as  $F(t)$  and  $f(t)$ , characterizes the distribution. This function, widely used in Reliability and Survival Analysis, expresses the force of occurrence of events at time  $t$ . Distributions where  $r(t)$  is increasing(decreasing) for all  $t$ , are called Increasing(Decreasing) Failure (or Hazard) Rate distributions, denoted as IFR(DFR). Exponential distribution has a constant hazard function, and hence belongs to both families IFR and DFR.

There is an interesting relationship between the behavior of the hazard function of  $F$  and the overdispersion or underdispersion of the number of arrivals, stated in the following proposition (Barlow & Proschan, 1965, p. 54):

**Proposition 1.** *If  $F$  is IFR (DFR), then  $\text{Var}(N(t)) \leq (\geq)E(N(t))$ .*

A similar result was rediscovered by Winkelmann (1995), which is only valid asymptotically, and uses different proof. In fact, this proof is very interesting because it permits a different and useful proposition based on the coefficient of variation of  $F$ :

**Proposition 2.** *Let  $\mu$  and  $\sigma$  be the expectation and standard deviation of  $F$ , and  $c = \sigma/\mu$  its coefficient of variation. Then, if  $c < (>)1$ , there exists  $t_0 \geq 0$  such that  $\text{Var}(N(t)) \leq (\geq)E(N(t))$ , for  $t > t_0$ .*

*Proof.* The proof, outlined in Winkelmann (1995), is a direct consequence of the fact that, as  $t$  tends to infinity,  $E(N(t)) \approx t/\mu$  and  $\text{Var}(N(t)) \approx t\sigma^2/\mu^3$ . ■

Proposition 1 can be directly applied to explore the number of arrivals when  $F$  is Gamma or Weibull distributed because their hazard functions are monotone, IFR or DFR being dependent on the value of their shape parameter. However if  $F$  follows an Inverse Gaussian distribution, its hazard function is not monotone, because its profile first increases and then decreases (Seshadri, 1998), so Proposition 1 cannot be applied. However, its coefficient of variation has a simple expression so Proposition 2 is useful in this case.

Although Proposition 1 only shows a sufficient condition for underdispersion of the number of arrivals, this condition is physically very meaningful and highly explanatory: When the events are more likely to happen as the time between occurrences passes, an underdispersed number of arrivals (occurrences) is produced. Unfortunately, in general, the probability function of the number of arrivals (1) associated with this mechanism does not have a closed form. On the other hand, it is not always possible to express any count distribution as the number of arrivals (for instance at  $t = 1$ ) of a certain renewal process, and, even when possible, it is very difficult.

## 2.2 | Pure birth processes

A pure birth process is an arrival process where the interarrival times  $X_i$  are independent and exponentially distributed random variables with parameters  $\lambda_i$  (the mean is  $1/\lambda_i$ ). Note that the case where  $\lambda_0 = \lambda_1 = \dots = \lambda$  corresponds to the Poisson process. A remarkable result of Ball (1995), previously conjectured by Faddy (1994), establishes that the distribution of the number of arrivals is under(over)-dispersed depending on the behavior of the rates  $\lambda_i$ :

**Proposition 3.** *If the sequence  $\lambda_0, \lambda_1, \dots$  is decreasing (increasing) then,  $\text{Var}(N(t)) < (>)E(N(t))$ .*

The probabilities of the number of arrivals can be calculated from the Kolmogorov forward differential equations, obtaining  $P(N(t) = 0) = \exp(-\lambda_0 t)$  and for  $n \geq 1$ ,

$$P(N(t) = n) = \int_0^t \lambda_{n-1} \exp(-\lambda_n(t-x))P(N(x) = n-1) dx.$$

Considering that time scale is arbitrary and fixing  $t = 1$ , we denote  $p_n = p_n(1) = P(N(1) = n)$ . This distribution can be expressed in terms of an exponential matrix of the form:

$$(p_0, p_1, \dots, p_n) = (1, 0, 0, \dots, 0) \exp(Q), \tag{2}$$

where,

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & \cdot & 0 \\ 0 & -\lambda_1 & \lambda_1 & \cdot & 0 \\ 0 & 0 & -\lambda_2 & \cdot & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda_n \end{pmatrix}$$

It is clear that any sequence of  $\lambda_n$ 's determines a count distribution of the number of arrivals  $p_n$ 's. For instance, a linearly increasing sequence of the form  $\lambda_n = a(b + n)$  generates the Negative Binomial distribution. A linearly decreasing sequence of the form  $\lambda_n = a(M - n)$ ,  $n = 0, \dots, M$ , generates a Binomial distribution with size parameter  $M$ . Unfortunately, except for these linear sequences, the distribution of arrivals  $p_n$  does not generally have a closed form for any arbitrary sequence of  $\lambda_n$ 's. However, a simple two-parameter distribution can be obtained considering a simple pure-birth process with rates  $\lambda_0 = a$  and  $\lambda_1 = \lambda_2 = \dots = b$  (Ball (1995)). When  $a = b$  this is a Poisson process and, according to Proposition 3, when  $a > (<)b$  the distribution of the number of arrivals is under(over)dispersed. Straightforward calculations show that,

$$p_0 = e^{-a}, \quad p_1 = \frac{a}{a-b}(e^{-b} - e^{-a})$$

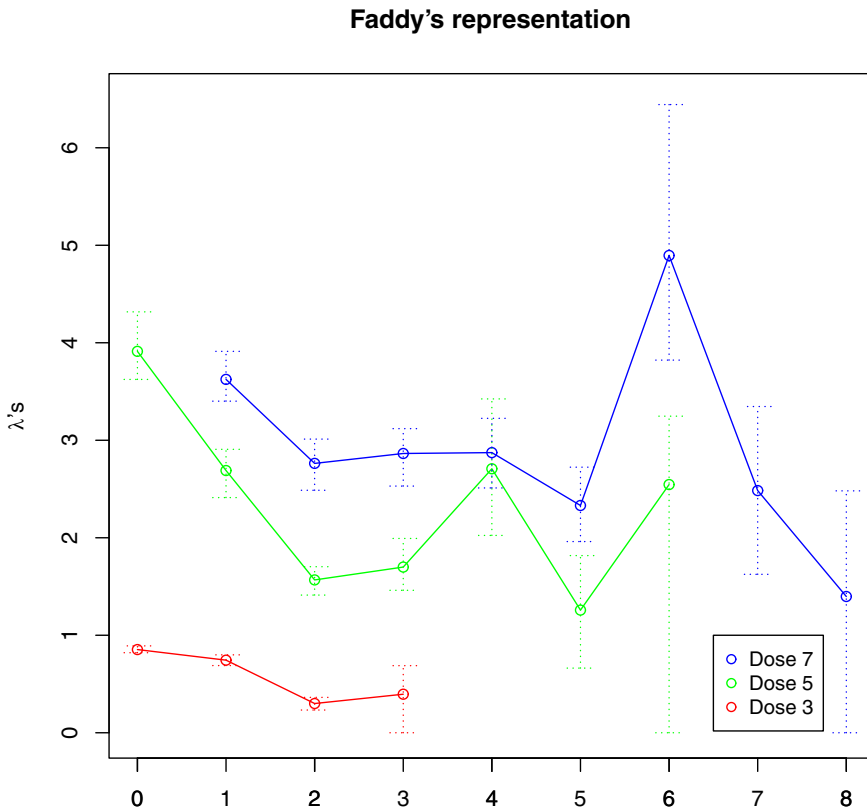
$$p_n = \frac{-b}{a-b}p_{n-1} + \frac{ab^{n-1}e^{-b}}{(a-b)(n-1)!}, \quad n \geq 2.$$

Underdispersion occurs ( $a > b$ ) when the first birth arises at a higher rate than those coming after.

Faddy (1997) considered a family of count distributions given by sequences of the form,  $\lambda_n = a(b + n)^c$ . When  $c = 0$  this is a constant sequence leading to a Poisson distribution. When  $c < 0 (c > 0)$ , the sequence is decreasing (increasing) and  $p_n$  is underdispersed (overdispersed). The case of  $c = 1$  corresponds to the Negative Binomial distribution. Podlich et al. (2004) used this family to construct semi-parametric count models that can describe several examples like the number of surviving fetal implants, and the number of cases of Toxoplasmosis.

Faddy (1997) showed that for any count distribution  $\pi_n$  there is a sequence of  $\lambda_n$ 's such that  $\pi_n = P(N(1) = n)$ . This allows us to represent any count distribution by the number of arrivals of a pure birth process. From the empirical version of Equation (2) we have the system of equations,

$$(\hat{p}_0, \hat{p}_1, \dots, \hat{p}_n) \exp(-Q) = (1, 0, 0, \dots, 0),$$



**FIGURE 1** Representation of the distributions of dicentrics for three doses. The error bars represent the first and third quartiles of the bootstrap values.

which can be solved numerically for the estimated rates  $\hat{\lambda}_n$ 's. Figure 1 shows the values of  $\hat{\lambda}_n$ 's obtained for the distributions of dicentrics shown in Table 1. The uncertainty of the  $\hat{\lambda}_n$ 's is estimated by resampling with 1000 replications. The vertical bars in Figure 1 extend from the first to the third quartiles of the bootstrap values. Note that the interquartile range is large for those  $\hat{\lambda}_n$ 's corresponding to low frequency observations.

In the example shown in Figure 1, Faddy's representation provides no further information other than a slightly decreasing trend for each dose, which is consistent with the observed underdispersion. Nonetheless, for other uses, Faddy's representation can be a meaningful and valuable modeling tool.

### 3 | STEADY STATE DISTRIBUTIONS OF BIRTH-DEATH PROCESSES

A birth-death process is a special case of a continuous-time Markov process where the states take values equal to 0, 1, 2, ... These processes are used to describe the size of a population, or the number of individuals in a queue, over time. When the process increases by 1, a *birth* event happens, and when the process decreases by 1, a *death* event occurs. When the process is at state  $k$  (the number of individuals in the queue is  $k$ ) there are two independent exponential occurrence

times running simultaneously: one with rate  $\lambda_k$  that will take us to state  $k + 1$  (birth) if it happens first, and the other with rate  $\mu_k$  which will take us to state  $k - 1$  (death) if it happens first. Let  $p_n(t)$  the probability that the birth-death process is in state  $n$  at time  $t$ . These probabilities satisfy the infinite system of differential equations,

$$\begin{aligned} p'_0(t) &= \mu_1 p_1(t) - \lambda_0 p_0(t) \\ p'_n(t) &= \lambda_{n-1} p_{n-1}(t) + \mu_{n+1} p_{n+1}(t) - (\lambda_n + \mu_n) p_n(t), \quad n = 1, 2, \dots, \end{aligned}$$

and the normalizing condition  $\sum_{n=0}^{\infty} p_n(t) = 1$ . Analytical solutions of these differential equations are difficult and often an impossible task. Nevertheless, it is possible to estimate the parameters of a general birth-death process by observing data at discrete or continuous times (see Crawford et al. (2014) and the references therein). However, consistently with our applications, we will focus the attention on their steady state distributions.

A well known result of the theory of birth-death processes establishes that there is a stationary distribution iff,

$$\sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \dots \lambda_{n-1}}{\mu_1 \mu_2 \dots \mu_n} < \infty. \tag{3}$$

In this case, the probabilities of the stationary distribution are,

$$p_n = \frac{\lambda_0 \lambda_1 \dots \lambda_{n-1}}{\mu_1 \mu_2 \dots \mu_n} p_0, \tag{4}$$

and,

$$p_0 = \left( 1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \dots \lambda_{n-1}}{\mu_1 \mu_2 \dots \mu_n} \right)^{-1},$$

where  $p_n = \lim_{t \rightarrow \infty} p_n(t)$ . In general, this distribution is overparameterized as it depends on the ratios  $\lambda_n / \mu_{n+1}$ . But when is this equilibrium distribution underdispersed? To answer this question let us define

$$\theta_r = (r + 1) p_{r+1} / p_r = (r + 1) \lambda_r / \mu_{r+1}, \quad r = 0, 1, 2, \dots \tag{5}$$

The behavior of  $\theta_r$  has been used to characterize some families of distributions. For instance, this is the case of the Katz family defined by a linear relationship of the form,  $\theta_r = \alpha + r\beta$ , which comprises Poisson ( $\alpha > 0, \beta = 0$ ), Binomial ( $\alpha > 0, \beta < 0$ ) and Negative Binomial ( $\alpha > 0, 0 < \beta < 1$ ) distributions. The sequence  $\theta_r$  is also related to the Böhning (2016) *ratio plot* for the case of Poisson distribution.

A little known result of Wise (1962) shows an important relationship between the behavior of  $\theta_r$  and the dispersion of  $p_n$ .

**Proposition 4.** *Let  $N$  be a random variable following the equilibrium distribution of the birth-death process, such that  $E(N) = \mu$  and  $Var(N) = \sigma^2$ . Then,  $cov(N, \theta_N) = \sigma^2 - \mu$ .*

A direct consequence of Proposition 4 is a sufficient condition on  $\theta_r$  (5) to assure underdispersion:

**Corollary 1.** *Consider a decreasing sequence  $\theta_0 \geq \theta_1 \geq \theta_2 \geq \dots$ . Then,  $\sigma^2 \leq \mu$ . When the sequence is increasing the inequality gets reversed ( $\sigma^2 \geq \mu$ ).*

The Wise (1962) result can be used to model many of the underdispersed count distributions studied in the past by several authors, as we shall analyze in the following subsections.

### 3.1 | Queues with state-dependent service rates: the COM-Poisson distribution and others

The classic and simplest queuing model, denoted as M/M/1, consists of a queue where a single server attends to certain events (births, jobs, persons, etc) arriving according to a Poisson process at a constant rate  $\lambda$ . It is usually assumed that service times follow exponential distributions that are independent of the interarrival times. Suppose that service times are state dependent, with rates given by  $\mu_n = f(n)$ , where  $n$  indicates the number of events waiting in the queue.

Conway and Maxwell (1961) introduced a distribution that would later be called *COM-Poisson* (named after R. W. Conway, W. L. Maxwell, and S. D. Poisson), considering  $f(n) = an^c$ , being studied later by Shmueli et al. (2005). According to (4), the probabilities take the form,

$$p_n = \frac{\rho^n}{(n!)^c} p_0, \quad n \geq 1,$$

and,

$$p_0 = \left( \sum_{i=0}^{\infty} \frac{\rho^i}{(i!)^c} \right)^{-1},$$

where  $\rho = \lambda/a$ . In this case, sequence (5) remains,  $\theta_r \propto (r+1)^{1-c}$ . Consequently, using Corollary 1, for  $c > 1$  ( $c < 1$ ) the distribution is underdispersed (overdispersed). Note that for  $c = 1$  this is just the Poisson distribution, which has a linear service-rate function ( $f(n) = an$ ). Under (over)dispersion is therefore related with a more (less) stressed service-rate function than that of Poisson distribution. Note that COM-Poisson distribution has the property that  $\log(\theta_n)$  and  $\log(p_{n+1}/p_n)$  are linear functions of  $\log(n)$ . In fact, Shmueli et al. (2005) proposed plotting the logarithms of the ratios of successive frequencies against  $\log(n)$ , as a method of validation and parameter estimation (using least squares) of the COM-Poisson distribution.

Insufficient attention has been paid to this fruitful methodology for generating underdispersed count distributions. For instance, consider now service rates of the form  $f(n) = ae^{bn}$ . Using (4) we obtain the probabilities,

$$p_n = \frac{\rho^n}{e^{\frac{n(n+1)}{2}b}} p_0 = e^{\beta_0 n - \beta_1 n^2} p_0, \quad n \geq 1, \quad (6)$$

and,

$$p_0 = \left( \sum_{i=0}^{\infty} e^{\beta_0 i - \beta_1 i^2} \right)^{-1},$$



where  $\rho = \lambda/a$ ,  $\beta_0 = \log(\rho) - b/2$  and  $\beta_1 = b/2$ . The domain of the parameters is  $\beta_1 > 0$  and  $\beta_0 \in \mathbb{R}$ , with the boundary case  $\beta_1 = 0$  and  $\beta_0 < 0$ , corresponding to the geometric distribution. This distribution was first introduced by Gelfand and Dalal (1990), although they wrongly commented that this distribution was always overdispersed. Observe that the sequence  $\theta_r \propto (r + 1)e^{-2\beta_1 r}$ ,  $r = 0, 1, \dots$ , is decreasing for  $\beta_1 > \log(2)/2$ , so, according to Corollary 1, the distribution is underdispersed for this range of  $\beta_1$ . For  $\beta_1 \leq \log(2)/2$  the sequence is not monotonic so Corollary 1 cannot be applied. Anyway, numerical calculation show that when  $\beta_1$  is close to 0 the distribution is overdispersed. Of course the boundary case  $\beta_1 = 0$  and  $\beta_0 < 0$  corresponds to the geometric distribution that is overdispersed. Note that for this distribution the log-ratio  $\log(p_{n+1}/p_n)$  is just a linear function of  $n$ .

The following example considers service rates of the form  $f(n) = ane^{bn}$ , which increase with  $n$  much more than in the preceding example. According to (4) we calculate the probabilities,

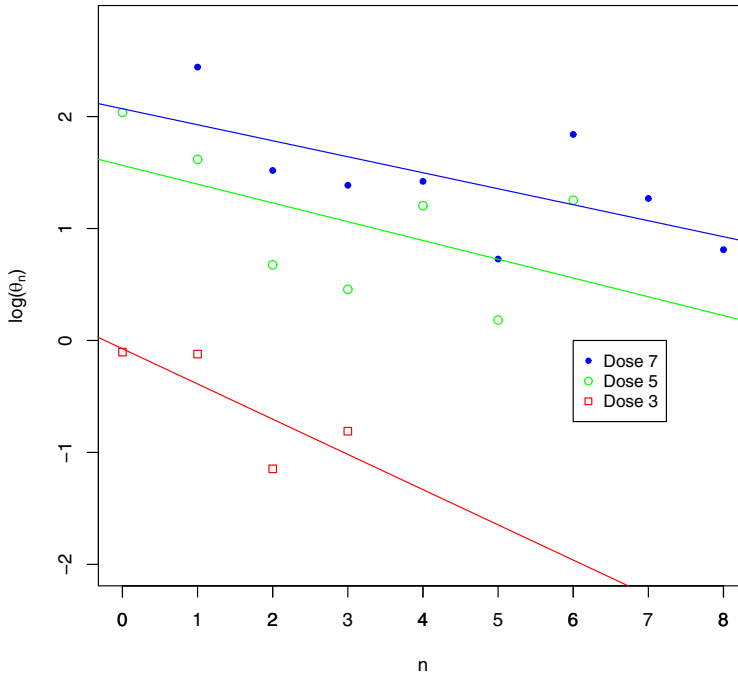
$$p_n = \frac{\rho^n}{n!e^{\frac{n(n+1)}{2}b}} p_0 = \frac{e^{\beta_0 n - \beta_1 n^2}}{n!} p_0, \quad n \geq 1, \tag{7}$$

with,

$$p_0 = \left( \sum_{i=0}^{\infty} (e^{\beta_0 i - \beta_1 i^2})/i! \right)^{-1}.$$

In this case the domain of the parameters is  $\beta_1 > 0$  and  $\beta_0 \in \mathbb{R}$ , with the boundary case  $\beta_1 = 0$  and  $\beta_0 < 0$ , corresponding to the Poisson distribution. Now the sequence  $\theta_r \propto e^{-2\beta_1 r}$ ,  $r = 0, 1, \dots$ , is always decreasing except for the boundary case  $\beta_1 = 0$ , so the distribution is underdispersed. This distribution has the property that  $\log(\theta_n)$  is a linear function of  $n$ . Figure 2 shows the empirical version of  $\log(\theta_n)$  against  $n$ , for the distributions of dicentrics described in Table 1. The corresponding least-squares regression lines point out their linear trends. Consequently, we have fitted the data using the probability function in (7), obtaining the maximum likelihood estimates and the chi-square statistics shown in Table 2. In order to avoid expected frequencies of less than one, the two last cells within doses 5 and 7 were collapsed in the chi-square computation. As usual, the degrees of freedom (df) are calculated as the number of cells minus 1, minus 2 (the number of estimated parameters). If the model does not fit with experimental data, the validity of the assumed service rates is put into question. Observe that  $\hat{\beta}_0$  increases and  $\hat{\beta}_1$  decreases with dose. It is interesting to point out that the MLE of the population mean and variance are exactly equal to the sample mean and variance. It happens because the distribution (7) used in this example, and the distribution (6) by Gelfand and Dalal (1990), are two-parameter exponential families with sufficient statistic for both the sum and the sum of the squares of the observations. For characterizations of distributions where the MLE of the population mean is the sample mean see Puig (2003) and Puig and Valero (2006) and the references therein. To include the dose as a covariate, we would need to analyze more experiments with a wide range of doses to investigate the dose dependence of the coefficients  $\beta_0$  and  $\beta_1$ . To guarantee that  $\beta_1 > 0$ , a suitable link function could be used. Anyway, the interpretation of these results from the point of view of queues with state-dependent service rates is very meaningful. This is the summary:

1. When the cell is irradiated, the particles (x rays) arrive following a Poisson process, producing damages in the cell.
2. The cell repair mechanism (service) tries to repair the damages.



**FIGURE 2** Empirical  $\log(\theta_n)$  of the distributions of dicentric chromosomes against  $n$ , with least-squares regression lines, for three X-ray doses.

**TABLE 2** MLE of the parameters with standard errors (in brackets) and expected frequencies of the distribution of dicentric chromosomes, using the probability function (7); values of the  $\chi^2$  statistic and df (in brackets) for several doses.

Dose (grays)	MLE		Number of dicentric chromosomes: expected frequencies										$\chi^2$ (df)		
	$\hat{\beta}_0$	$\hat{\beta}_1$	0	1	2	3	4	5	6	7	8	9			
3	0.1431 (0.1331)	0.1613 (0.0542)	208.6	204.8	72.8	12.5	1.2								3.93 (2)
5	1.6691 (0.2179)	0.1288 (0.0381)	5.4	25.1	45.3	42.1	22.7	7.6	1.6	0.2					9.30 (4)
7	1.9529 (0.1940)	0.0642 (0.0221)		6.5	18.8	32.1	36.1	28.5	16.5	7.2	2.4	0.6			6.41 (5)

- If the mechanism is able to repair a damage, the corresponding aberration leaves the queue. The chromosome aberrations that are finally observed (dicentric chromosomes) are the number of events in the queue.
- It is assumed that we have reached a point in the process where the distributions of the number of events will no longer change with time.

Then, underdispersion means that the repair mechanism is strongly accelerated by the number of cases of damage. We conjecture that the cell-repair mechanism is then described through function  $\mu_n = f(n)$ . In-depth experiments are required to prove this conjecture.

### 3.2 | Weighted Poisson distributions

A count random variable is weighted-Poisson distributed with weight  $w(k; b)$  if its probability function has the form,

$$P(X = k) = \frac{w(k; b) e^{-\lambda} \lambda^k}{C(\lambda, b) k!}, \quad k = 0, 1, 2, \dots,$$

where  $C(\lambda, b)$  is a normalizing constant and  $b$  is a parameter (or a vector of parameters). The weight  $w(k; b)$  can be interpreted as a kind of recording or sampling mechanism. In fact, it is immediate to see that any arbitrary count distribution can be expressed as a weighted Poisson distribution with a suitable weight. Weighted distributions go back a long way in the literature. See, for instance, Patil et al. (2004) and the references therein.

Weighted Poisson distributions have been frequently used to model underdispersed data, using weights for which the normalizing constant  $C(\lambda, b)$  takes a closed form. For instance, Pujol et al. (2014) use the weight  $w(k; b) = 1 + bk^2$  leading to a normalizing constant  $C(\lambda, b) = 1 + b(\lambda + \lambda^2)$ , and Cameron and Johansson (1997) use amenable polynomial weights of the form  $w(k; \mathbf{b}) = (1 + \sum_{i=1}^r b_i k^i)^2$ . del Castillo and Pérez-Casany (2005) and Kokonendji et al. (2008), using different methods and notation, state the following result:

**Proposition 5.** *Let the weight  $w(x; b)$  be a log-concave(convex) function in  $x \in \mathbb{R}$ . Then, the corresponding weighted Poisson distribution is under(over)-dispersed.*

Interestingly, Weighted Poisson distributions can also be viewed as stationary distributions of birth-death processes, such that

$$\theta_r = (r + 1)\lambda_r / \mu_{r+1} = \lambda \frac{w(r + 1; b)}{w(r; b)}. \tag{8}$$

Therefore, a new and simple proof of Proposition 5 arises from the fact that if  $w(x; b)$  is a log-concave(convex) function then  $w(x + 1; b)/w(x; b)$  is decreasing(increasing), and the corresponding weighted Poisson distribution remains under(over)-dispersed, accordingly to Corollary 1.

There is also a direct relation between Weighted Poisson distributions and queues with state-dependent service rates, evidenced by considering  $\lambda_0 = \lambda_1 = \dots = \lambda$  in (8). Then we obtain,

$$\mu_n = f(n) = \frac{w(n - 1; b)n}{w(n; b)}. \tag{9}$$

In fact, any service rate function proportional to  $f(n)$  will lead to the same equilibrium distribution family. Kokonendji et al. (2009) presented a nonparametric, kernel-based approach for estimating the weight function. It is important to note that this method would additionally provide a nonparametric way to estimate the  $\mu_n$  service-rate function using expression (9).

As an example of how these concepts are related, let us consider the zero-modified Poisson distribution (Johnson et al., 2005), with probability function,

$$p_0 = \gamma, \tag{10}$$

$$p_k = (1 - \gamma) \frac{e^{-\lambda} \lambda^k}{k!(1 - e^{-\lambda})}, \quad k = 1, 2, \dots$$

This distribution is widely used to model both underdispersed and overdispersed data. When  $\gamma = e^{-\lambda}$ , that is  $p_0$  is equal to the proportion of zeros of the Poisson distribution, (10) is just the Poisson probability function. Now the sequence (5) is  $\theta_0 = \phi\lambda$ ,  $\theta_1 = \theta_2 = \dots = \theta_r = \dots = \lambda$ , where  $\phi = (1 - \gamma)e^{-\lambda}/(\gamma(1 - e^{-\lambda}))$ . When  $\gamma < (>)e^{-\lambda}$  it means zero-deflation(inflation),  $\phi > (<)1$  and the sequence is decreasing(increasing), and the distribution is under(over)dispersed accordingly to Corollary 1. Note that the zero-modified Poisson distribution is a weighted-Poisson distribution with weight function  $\omega(0) = 1/\phi$  and  $\omega(k) = 1$ ,  $k = 1, 2, \dots$ . Then, from (9), we find that  $\mu_1 = 1/\phi$  and  $\mu_n = n$ ,  $n = 2, 3, \dots$ . There is an interesting interpretation: Poisson distribution arises as the equilibrium distribution of the number of customers in a queue, in a process with constant-rate exponential arrival times and exponential service-time rates proportional to the number of individuals in the queue. The alteration in the service-time rate when there is only 1 individual in the queue ( $n = 1$ ) has significant consequences. When the service-time is reduced(augmented)  $1/\phi < (>)1$ , that is the rate is augmented(reduced), the equilibrium distribution is Poisson zero-deflated(inflated), being under(over)dispersed.

In the following section we are going to present other mechanisms leading to underdispersion that are not directly related with arrival or birth-death processes.

#### 4 | THINNING MECHANISMS

Many observations in experimental sciences are related to the problem of thinning a random variable  $X$ . In this case  $X$  itself is not observed, rather a proportion  $\alpha$  or sampled version of  $X$ ,  $X_\alpha$ , is observed. The standard binomial thinning is defined as follows,

**Definition 1.** Let  $X$  be a count random variable, and let  $\xi_1, \xi_2, \dots$  be iid Bernoulli random variables with probability of success  $\alpha \in (0, 1]$ , all of them independent of  $X$ . The count random variable,

$$X_\alpha = \sum_{i=1}^X \xi_i \quad (X_\alpha = 0 \quad \text{if} \quad X = 0),$$

is called a binomial  $\alpha$ -thinning of  $X$ .

Note that  $X_\alpha$  condition on the value of  $X$  is binomially distributed, that is,  $X_\alpha | X = x \sim \text{Bin}(x, \alpha)$ . Poisson and Mixed-Poisson distributions are closed under independent binomial  $\alpha$ -thinning, in the sense that if  $X$  is Poisson or Mixed-Poisson distributed, then  $X_\alpha$  belongs to the same family. Several properties regarding closure under binomial thinning are shown in Puig and Valero (2007). The expectation and variance of  $X_\alpha$  are,

$$E(X_\alpha) = \alpha\mu, \quad \text{Var}(X_\alpha) = \alpha^2\sigma^2 + \mu\alpha(1 - \alpha), \quad (11)$$

where  $E(X) = \mu$  and  $\text{Var}(X) = \sigma^2$ .

It is plain to see that,

$$\frac{\text{Var}(X_\alpha)}{E(X_\alpha)} = \alpha \left( \frac{\sigma^2}{\mu} - 1 \right) + 1, \quad \forall \alpha \in (0, 1]$$

and consequently if  $X$  is underdispersed(overdispersed) so is  $X_\alpha$ . Moreover, the dispersion index of  $X_\alpha$  shrinks toward 1.

### 4.1 | Variable thinning

Consider now a mechanism of sight such that the proportion of observed events, recorded in the random variable  $Y$ , changes with respect to the true value of the hidden variable  $X$ . A simple model consists of assuming a variable binomial thinning as follows,

**Definition 2.** Let  $X$  be a count random variable, and let  $\alpha = \{\alpha_1, \alpha_2, \dots\}$  be a sequence such that,  $0 < \alpha_i < 1$ . The count random variable,  $Y$ , such that  $Y|X = n \sim \text{Bin}(n, \alpha_n)$ , denoted as  $Y = X_\alpha$  is called a variable binomial  $\alpha$ -thinning of  $X$ .

Note that when  $\alpha = \{\alpha_1, \alpha_1, \alpha_1, \dots\}$ , it corresponds to the standard binomial thinning  $X_{\alpha_1}$ . The probability function of  $Y$  is,

$$P(Y = k) = \sum_{n=0}^{\infty} P(Y = k|X = n)P(X = n) = \sum_{n=k}^{\infty} \binom{n}{k} \alpha_n^k (1 - \alpha_n)^{n-k} P(X = n). \tag{12}$$

The probability generating function (pgf) can be directly derived obtaining,

$$\phi_Y(t) = E(t^Y) = \sum_{n=0}^{\infty} E(t^Y|X = n)P(X = n) = \sum_{n=0}^{\infty} (1 + \alpha_n(t - 1))^n P(X = n). \tag{13}$$

The following result connects the behavior of  $\alpha_n$  with the under(over)dispersion of the observed events recorded by  $Y$ .

**Proposition 6.** Let  $X$  be a random variable following a Poisson distribution with expectation  $E(X) = \lambda$ , and  $Y = X_\alpha$ .

1. Let  $\alpha_n$  be increasing in  $n$ . Then  $Y$  is overdispersed.
2. Let  $\alpha_n$  be decreasing, and  $n\alpha_n$  increasing in  $n$ . Then  $Y$  is underdispersed.

*Proof.* Stein-Chen’s identity will be used in the proof. It states that if  $X$  follows a Poisson distribution, and  $f$  is any function for which the expected value  $E(f(X))$  exists, then  $E(Xf(X)) = E(X)E(f(X + 1))$ .

Evaluating the derivatives of the pgf (13) at  $t = 1$ , we obtain,

$$\mu_Y = \phi'_Y(1) = \sum_{n=0}^{\infty} n\alpha_n \frac{e^{-\lambda} \lambda^n}{n!} = E(X\alpha_X) = E(X)E(\alpha_{X+1}),$$

where  $X$  is a Poisson- $\lambda$  random variable. Similarly,

$$\sigma_Y^2 + \mu_Y^2 - \mu_Y = \phi''_Y(1) = \sum_{n=0}^{\infty} n(n - 1)\alpha_n^2 \frac{e^{-\lambda} \lambda^n}{n!} = E(X(X - 1)\alpha_X^2) = E(X)E(X\alpha_{X+1}^2).$$

Because  $\sigma_Y^2 - \mu_Y = \phi''_Y(1) - (\phi'_Y(1))^2$ , overdispersion or underdispersion depends of the sign of  $E(X\alpha_{X+1}^2) - E(X)E^2(\alpha_{X+1})$ . Direct algebra shows that

$$E(X\alpha_{X+1}^2) - E(X)E^2(\alpha_{X+1}) = \text{cov}(X\alpha_{X+1}, \alpha_{X+1}) + E(\alpha_{X+1})\text{cov}(X, \alpha_{X+1}). \tag{14}$$

**TABLE 3** MLE of the parameters with standard errors (in brackets) and expected frequencies of the distribution of dicentric, using the  $\alpha$ -thinning model with  $X \sim \text{Poisson}(\lambda)$  and  $\text{logit}(\alpha_n) = b/n$ ; values of the  $\chi^2$  statistic and df (in brackets) for several doses.

Dose (grays)	MLE		Number of dicentric: expected frequencies									$\chi^2$ (df)		
	$\hat{b}$	$\hat{\lambda}$	0	1	2	3	4	5	6	7	8		9	
3	4.1487 (1.2628)	0.8685 (0.0643)	213.8	197.6	72.5	14.1	1.8							4.54 (2)
5	4.1246 (1.1611)	3.4392 (0.3443)	5.9	27.5	45.2	38.0	20.9	8.6	2.8	0.8				7.78 (4)
7	5.5545 (1.7332)	5.6134 (0.6002)		6.0	20.3	33.8	35.3	26.5	15.5	7.4	3.0	1.1		3.95 (6)

Consequently, if  $\alpha_n$  is increasing  $\alpha_{n+1}$  and  $n\alpha_{n+1}$  are also increasing and both covariances in (14) are positive implying overdispersion.

On the other hand, if  $\alpha_n$  is decreasing and  $n\alpha_n$  is increasing,  $\alpha_{n+1}$  is decreasing and  $n\alpha_{n+1}$  is increasing. Therefore, both covariances in (14) are negative implying underdispersion. ■

The condition for underdispersion shown in Proposition 6 is very meaningful. A decreasing  $\alpha_n$  means, in some sense, that it is more difficult to count a large number of objects than a few. In addition, an increasing  $n\alpha_n$  means that, although the sight mechanism penalizes counts of a large number of objects, this penalization is not very hard, thus allowing the conditional average  $E(Y|X = n) = \alpha_n n$  to be increasing. In other words, it means that as the true number of objects  $X$  increases, the number of observed objects also increases on average.

For example, consider a model with  $\alpha_n = a + bn^{-c}$ , where  $a, b > 0$ ,  $a + b < 1$  and  $0 < c < 1$ . It can immediately be checked that  $\alpha_n$  is decreasing and  $n\alpha_n$  increasing leading to underdispersion.

It is also useful to describe models in terms of  $\text{logit}(\alpha_n)$ . For instance, the model with  $\text{logit}(\alpha_n) = b/n$ ,  $b > 0$ , also satisfies that  $\alpha_n$  is decreasing and  $n\alpha_n$  increasing. This property is also shared using other link functions like  $\text{probit}(\alpha_n)$  or  $\text{cloglog}(\alpha_n)$ . As an example of application of this distribution, we reanalyze the distributions of dicentric described in Table 1. The maximum likelihood estimates, expected frequencies, and chi-square statistics are shown in Table 3. It shows a good fit, similar to that obtained in Section 3.1 using a service-rate-model distribution (Table 2). The interpretation is also meaningful:

1. Ionizing particles arrive to the cell accordingly to a Poisson process. It is assumed that the number of damages  $X$  is Poisson distributed. The estimated mean of the number of damages increases with the dose of radiation, as expected.
2. Cell repair mechanism acts, and the final number of aberrations is  $Y = X_\alpha$ , a variable thinning of the original number of damages  $X$ . The action of the repair mechanism, that is  $\alpha_n$ , is very similar for doses 3 and 5 grays, and slightly less intense (less efficient) for 7 grays ( $\hat{b} = 5.5545$ ).
3. Cell repair mechanism is strongly stressed with the number of damages so that  $\alpha_n$  decreases with  $n$ . However, given a number of damages  $n$ , the expected number of aberrations,  $n\alpha_n$ , increases.

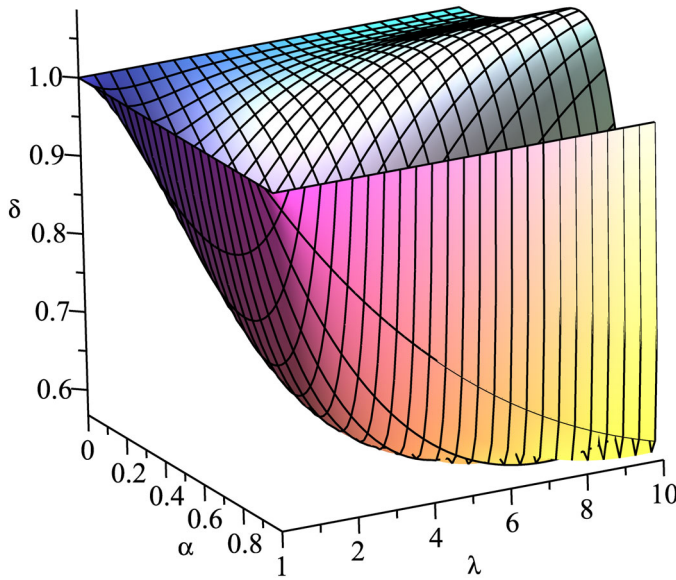


FIGURE 3 Profile of the dispersion index  $\delta$  as a function of  $\alpha$  and  $\lambda$ .

Another interesting model is that obtained from  $\alpha_n = \beta + \alpha^n$ ,  $\beta \geq 0$  and  $0 < \alpha < 1$ . It is clear that  $\alpha_n$  is decreasing, but  $n\alpha_n$  is increasing only when  $\beta > \alpha^{-2/\log(\alpha)}$ . In particular, when  $\beta = 0$ , this leads to a sight mechanism described by  $\alpha_n = \alpha^n$  producing a two-parameter count distribution with pgf,

$$\phi_Y(t) = \sum_{n=0}^{\infty} (1 + \alpha^n(t - 1))^n \frac{e^{-\lambda} \lambda^n}{n!}.$$

In this case, the expectation and dispersion index take a simple form:

$$\mu_Y = \alpha \lambda e^{\lambda(\alpha-1)} \quad \delta = \frac{\sigma_Y^2}{\mu_Y} = \alpha^3 \lambda e^{\lambda\alpha(\alpha-1)} - \mu_Y + 1.$$

Figure 3 shows the profile of the dispersion index as a function of  $\alpha$  and  $\lambda$ . Note that this model presents over-equi and underdispersion, being an example that the condition of increasing  $n\alpha_n$  cannot be removed from Proposition 6 to ensure underdispersion.

To show an example application of this distribution we analyze one of the classical fertilization datasets described in Morgan (1982) and reported in Table 4. This corresponds to an experiment conducted on sea urchins, where a batch of eggs was exposed to sperm at a fixed sperm concentration and a fixed temperature, with a dilute solution of nicotine. Table 4 shows the counts of the number of sperms that had penetrated the eggs at various time intervals for samples of 100 eggs. The row corresponding to time equal to 5 seconds has been removed from the original data here because it is not informative. This dataset was also analyzed by Ridout and Besbeas (2004).

We can observe a linear trend profile if we plot the sample means of the number of fertilizations shown in Table 4 against log-time. Moreover,  $\mu_Y \approx \lambda$  when  $\alpha$  is close to 1. Therefore, we have fitted the data considering parameter  $\lambda$  to be time-dependent according to the linear

TABLE 4 Distributions of sperm over eggs.

Time (seconds)	Number of fertilizations							$\bar{x}$	$s^2$	
	0	1	2	3	4	5	6			7
12	82	17	1						0.19	0.1757
20	33	56	9	2					0.80	0.4646
40	17	68	12	2	1				1.02	0.4642
80	9	56	25	7	2	1			1.40	0.8081
180	1	59	27	9	2	1		1	1.60	0.9899

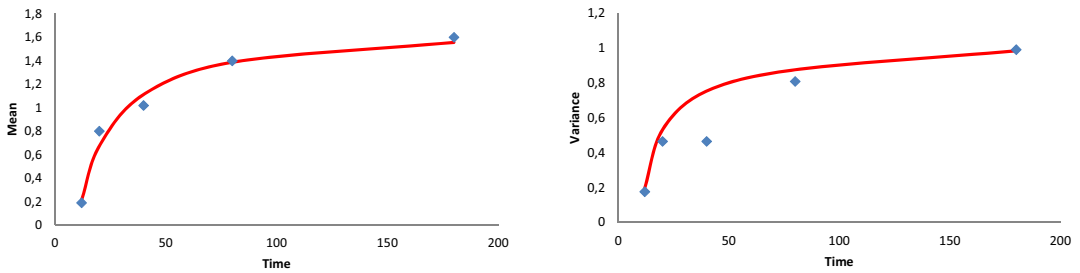


FIGURE 4 Empirical and fitted mean (left) and variance (right).

relationship,  $\lambda(t) = \beta_0 + \beta_1 \log(t)$ . The likelihood can be directly computed from the expression of the probabilities (12), and using the numerical maximization algorithm provided by function `nlm` in R we obtain the maximum likelihood estimates:  $\hat{\alpha} = 0.814$ ,  $\text{se} = 0.012$ ;  $\hat{\beta}_0 = -3.228$ ,  $\text{se} = 0.460$ ;  $\hat{\beta}_1 = 1.410$ ,  $\text{se} = 0.175$ . The interpretation of this model is meaningful:

1. The number of sperm  $X$  contacting an egg follows a Poisson distribution with an increasing mean over time.
2. The number of fertilizations is  $Y = X_\alpha$ , a variable thinning of the number of sperm  $X$ . Because  $\alpha_n = \alpha^n$ ,  $\alpha$  can be interpreted as the rate of efficiency for the first fertilization. The rate of efficiency for the second fertilization is  $\alpha^2$  and so on.

In other words, the eggs have blocking mechanisms limiting the occurrence of polyspermy and this is the cause of the underdispersion. Figure 4 shows the empirical mean and variance of the number of fertilizations for each time, together with their estimated values, exhibiting good performance.

## 4.2 | Variable thinning and negative correlation

As commented earlier, thinning mechanisms are experimentally meaningful. Sometimes, the  $\alpha$ -thinning random variable  $Y = X_\alpha$  and the “remainder”  $Z = X - X_\alpha$  are both meaningful, leading to a bivariate count random variable  $(Y, Z)$ , which is a bivariate stochastic decomposition of  $X$  since  $X = Y + Z$ .



The probability function of  $(Y, Z)$  is,

$$\begin{aligned}
 P(Y = r, Z = s) &= P(Y = r|X = r + s)P(X = r + s) = \\
 &= \binom{r + s}{r} \alpha_{r+s}^r (1 - \alpha_{r+s})^s P(X = r + s).
 \end{aligned}
 \tag{15}$$

Most bivariate count distributions have overdispersed marginals that are positively correlated. However, the following result shows the relation between the dispersion of  $X$  and the sign of the correlation of  $Y$  and  $Z$ , for the classical binomial thinning  $(\alpha_n = \alpha, \forall n)$ .

**Proposition 7.** *Let  $X$  be a count random variable with dispersion index  $\delta_X = \sigma_X^2/\mu_X$ ,  $Y = X_\alpha$ ,  $\alpha_n = \alpha, \forall n$ , and the remainder  $Z = X - Y$ . Then, the correlation coefficient is,*

$$\rho_{YZ} = \frac{\alpha(1 - \alpha)(\delta_X - 1)}{\sqrt{(\alpha^2(\delta_X - 1) + \alpha)((1 - \alpha)^2(\delta_X - 1) + 1 - \alpha)}}.
 \tag{16}$$

*Proof.* The proof is a direct consequence of (11) and the fact that  $Z = X_{1-\alpha}$ . Then,  $\delta_Y = 1 + \alpha(\delta_X - 1)$  and  $\delta_Z = 1 + (1 - \alpha)(\delta_X - 1)$ . Moreover,  $cov(Y, Z) = \alpha(1 - \alpha)\mu(\delta_X - 1)$ , and after some straightforward calculations, the proof is concluded. ■

Accordingly from expression (16), when  $X$  is overdispersed then  $Y$  and  $Z$  are both overdispersed, and their correlation is positive. Similarly, when  $X$  is underdispersed  $Y$  and  $Z$  are both underdispersed and their correlation is negative. Note that, if  $X$  is under(overdispersed), then so are  $Y$  and  $Z$ , but both  $Y$  and  $Z$  are less under(overdispersed) than  $X$ . Direct calculations show that the maximum (minimum) correlation is attained at  $\alpha = 1/2$ , obtaining

$$\max |\rho_{YZ}| = \frac{|\delta_X - 1|}{\delta_X + 1}.$$

Commonly used bivariate count models accommodate only a positive correlation between counts. The flexibility for allowing negative correlations for bivariate count distributions generated by the variable thinning stochastic decomposition is shown in the following result.

**Proposition 8.** *Let  $X$  be Poisson distributed and  $Y = X_\alpha$ , such that  $\alpha_n$  is decreasing in  $n$ . Consider the count random variable  $Z = X - Y$ , so that we obtain the decomposition  $X = Y + Z$ . Then, if  $E(\alpha_X) < 1/2$  the correlation is negative:  $Cor(Y, Z) < 0$ .*

*Proof.* Note that  $Z$  is also a variable thinning of  $X$  with probabilities  $\alpha_n^* = 1 - \alpha_n$ . Then,  $\alpha_n^*$  is increasing when  $\alpha_n$  is decreasing and vice versa. Let us now determine the sign of the covariance,

$$cov(X_\alpha, X - X_\alpha) = E(X_\alpha(X - X_\alpha)) - E(X_\alpha)E(X - X_\alpha).$$

Then,  $E(X_\alpha(X - X_\alpha)) = E(X)^2 E(\alpha_{X+2}(1 - \alpha_{X+2}))$ , according to the reasoning used in the proof of Proposition 6. Note that,  $E(X_\alpha) = E(X)E(\alpha_{X+1})$  and  $E(X - X_\alpha) = E(X)E(1 - \alpha_{X+1})$ . Therefore,

$$cov(Y, Z) = E(X)^2 (E(\alpha_{X+2}(1 - \alpha_{X+2})) - E(\alpha_{X+1})E(1 - \alpha_{X+1})),$$

and this is equal to,

$$E(X)^2(\text{cov}(\alpha_{X+2}, (1 - \alpha_{X+2})) + E(\alpha_{X+2})E(1 - \alpha_{X+2}) - E(\alpha_{X+1})E(1 - \alpha_{X+1})).$$

Note that  $\text{cov}(\alpha_{X+2}, (1 - \alpha_{X+2}))$  is negative because  $\alpha_n$  is decreasing, so that  $1 - \alpha_n$  is increasing in  $n$ . The remaining part is also negative because it can be written as the product,

$$(E(\alpha_{X+2}) - E(\alpha_{X+1}))(1 - (E(\alpha_{X+2}) + E(\alpha_{X+1}))),$$

where the first term is negative because  $\alpha_n$  is decreasing, and the second term is positive because  $E(\alpha_{X+2}) + E(\alpha_{X+1}) < 2E(\alpha_X) < 1$ . ■

For instance, condition  $E(\alpha_X) < 1/2$  can easily be checked for the model  $\alpha_n = \alpha^n$ ,  $0 < \alpha < 1$ , presented in Section 4.1. It can immediately be seen that in this case  $E(\alpha^X)$  is just the pgf of a Poisson- $\lambda$  distribution evaluated at  $\alpha$ , so condition  $E(\alpha_X) < 1/2$  is equivalent to  $e^{\lambda(\alpha-1)} < 1/2$ , or  $\alpha < 1 - \log(2)/\lambda$ .

Useful expressions for computing the expectations, variances and the covariance of  $Y$  and  $Z$  are also obtained from the proofs of Propositions 6 and 8:

$$\begin{aligned} \mu_Y &= \lambda E(\alpha_{X+1}), \quad \mu_Z = \lambda - \mu_Y \\ \sigma_Y^2 &= \lambda E(X\alpha_{X+1}^2) + \mu_Y - \mu_Y^2 \\ \sigma_Z^2 &= \lambda E(X(1 - \alpha_{X+1})^2) + \mu_Z - \mu_Z^2 \\ \text{cov}(Y, Z) &= \lambda^2 E(\alpha_{X+2} - \alpha_{X+2}^2) - \mu_Y \mu_Z \end{aligned} \quad (17)$$

Propositions 6 and 8 are useful to model bivariate count patterns, as we shall see in the following application.

The genome of coronavirus SARS-CoV-2 is about 30,000 “letters” long. The sequence of nucleotides can be downloaded from the National Centre for Biotechnology Information (NCBI) (<https://www.ncbi.nlm.nih.gov/>). We have worked with the sequence collected in Spain in 2020-03-15, accession number MT359865. Sliding window analyses are a common approach to study the local variation within a genome sequence. We want to explore the presence of pairs of specific trinucleotides in nonoverlapping chunks of size 200 (window length).

Consider for instance the number of trinucleotides **atc** ( $Y$ ) and the number of **tta** ( $Z$ ) along chunks, and the sum of both  $X = Y + Z$ . Here the number of observations for each variable is  $m = 149$  (the number of chunks). The sample means and variances are  $\bar{x} = 8.0268$ ,  $\bar{y} = 2.2550$  and  $\bar{z} = 5.7718$ , and  $s_X^2 = 7.4722$ ,  $s_Y^2 = 1.9615$  and  $s_Z^2 = 6.8935$ . The dispersion indexes are  $d_X = 0.9309$ ,  $d_Y = 0.8698$ ,  $d_Z = 1.1943$ , and the sample correlation coefficient between  $Y$  and  $Z$  is  $r_{Y,Z} = -0.1880$ . The negative sign of  $r_{Y,Z}$  is an indicator of a kind of competition between both trinucleotides. The Poisson assumption on  $X$  has been checked using the battery of 13 goodness of fit tests described by Puig and Weiss (2020). This battery includes some classical tests, like Fisher’s dispersion test, and new tests designed for a large family of overdispersed alternatives called the LC-class (Puig and Kokonendji (2018)). They explore different aspects of Poissonness. It is reasonable to assume that  $X$  follows a Poisson distribution because none of the 13 test rejected the Poisson assumption (the minimum  $p$ -value = 0.1211 was for the test  $\Delta_\infty$ ).

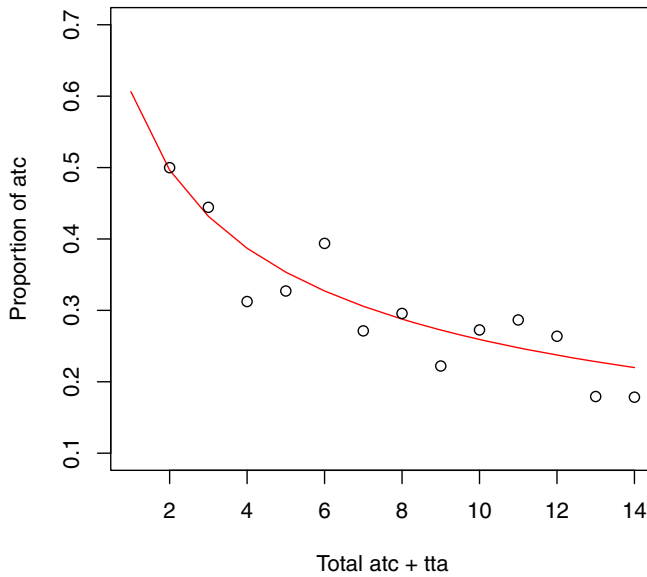


FIGURE 5 Observed and fitted values of  $\alpha_n$  for the counts of **atc**.

The values of  $\alpha_n = E(Y|X = n)/n$  can be nonparametrically estimated averaging the counts of  $Y$  for each observed value of  $X$ ,  $n$ , dividing the average by  $n$ . The dots in Figure 5 represent these estimates.

To model the bivariate distribution of  $(Y, Z)$  we assume that the odds of  $\alpha_n$  is a power function of the total number of counts ( $X$ ), that is,  $\log(\alpha_n/(1 - \alpha_n)) = a + b \log(n)$ . Interestingly, when the slope  $b$  is negative,  $\alpha_n$  is decreasing with  $n$  and  $n\alpha_n$  is increasing with  $n$ . Therefore, Proposition 6 ensures that  $Y$  is underdispersed and  $Z$  is overdispersed.

The likelihood function has the expression,

$$L(Y, Z) = \prod_{i=1}^m \binom{y_i + z_i}{y_i} \left( \frac{1}{1 + e^{-a-b \log(y_i+z_i)}} \right)^{y_i} \left( \frac{e^{-a-b \log(y_i+z_i)}}{1 + e^{-a-b \log(y_i+z_i)}} \right)^{z_i} \frac{e^{-\lambda} \lambda^{y_i+z_i}}{(y_i + z_i)!} \tag{18}$$

Note that parameter  $\lambda$  can be estimated separately because its MLE is just that of the Poisson distribution obtained with the observations of  $X$  ( $x_i = y_i + z_i$ ), that is  $\hat{\lambda} = \bar{x} = 8.0268$ . The part of the likelihood function corresponding to parameters  $a$  and  $b$  is that of a logistic regression model, with logit link function, with one covariate (the values of  $\log(X)$ ). Then, standard software for Generalized linear models can be used, obtaining the MLE,  $\hat{a} = 0.4318$ ,  $se = 0.4117$  and  $\hat{b} = -0.6434$ ,  $se = 0.1918$ . Although the intercept is not significant, note the negative sign of  $\hat{b}$  as expected. Figure 5 also shows the values of  $\hat{\alpha}_n = 1/(1 + e^{-\hat{a}-\hat{b} \log(n)})$  (red line) reflecting a good performance. The estimated value of  $E(\alpha_X)$  is  $\sum_{n=0}^{\infty} \hat{\alpha}_n e^{-\hat{\lambda}} \hat{\lambda}^n / n! = 0.2994 < 1/2$ , in agreement with Proposition 8, ensuring that both  $(Y, Z)$  are negatively correlated. Expectations, variances, and the correlation coefficient can be estimated using (17) obtaining,  $\hat{\mu}_Y = 2.2559$ ,  $\hat{\mu}_Z = 5.7710$ ,  $\hat{\sigma}_Y^2 = 1.8142$ ,  $\hat{\sigma}_Z^2 = 7.2872$  and  $\rho_{Y,Z} = -0.1478$ .

There are globally 2016 different pairs of combinations of trinucleotides. In SARS-CoV-2, only 316 pairs are significantly negatively correlated ( $\rho < -.15$ ) for chunks of size 200. Some of them can be analyzed using the methodology of the previous example (**atc** and **tta**), but the Poisson assumption on  $X$  is rejected for some other pairs. In these cases, although Propositions 6 and 8

could not be illustrated or applied, a similar analysis could be done by using suitable alternatives to the Poisson distribution in the likelihood function (18).

## 5 | DISCUSSION

The main message of this paper might be that: *if your data are underdispersed try to find out why this is happening, as the answer could be very enriching*. We have pointed out two major families of mechanisms leading to underdispersed count data distributions: stochastic processes (arrival and birth-death) and binomial thinning (constant and variable).

Renewal processes, such that the interarrival times follow an IFR distribution or have a coefficient of variation lower than 1, are an important source of underdispersed distributions of the number of arrivals. In Section 2.1 we reviewed an important result by Barlow and Proschan (1965), pointing out the true meaning of the result by Winkelmann (1995).

Pure birth processes are another meaningful source, producing underdispersed distributions when the sequence of birth-rates is decreasing (Ball (1995)). On the other hand, any arbitrary count distribution admits a representation as the number of arrivals (births) of a specific pure birth process (Faddy's representation). We have revisited these results in Section 2.2 pointing out the importance of Faddy's representation as a tool for modeling.

The steady state distributions of birth-death processes are undoubtedly important and meaningful sources of underdispersed distributions. We saw in Section 3 that decreasing sequences of ratio-plot elements  $\theta_n$  provide underdispersed distributions. In particular, queues with state-dependent service rates provide interesting models leading to distributions like COM-Poisson and that introduced by Gelfand and Dalal (1990). It is worth noting that representing the COM-Poisson distribution as a queuing model with service rate  $\mu_n = an^c$  provides for a simple proof, using Corollary 1, that the distribution is underdispersed (overdispersed) for  $c > 1$  ( $c < 1$ ). The theory of the steady state distributions of queues with state-dependent service rates, and Corollary 1, allows for construction of new underdispersed distributions in a simple way, for instance considering  $\log(\theta_n)$  as a linear function of  $n$ . This new distribution fits the frequencies of dicentrics very well, providing a biological explanation for why they are underdispersed.

Several authors have applied Weighted Poisson distributions for dealing with underdispersion. In Section 3.2 we showed that Weighted Poisson distributions can be interpreted as the steady state distribution of a birth-death process. This makes it possible to prove some theoretical results, also using Corollary 1.

Binomial thinning (or subsampling) is a meaningful operation that models sight or selection mechanisms in which an underlying count random variable  $X$  is partially observed or selected. We showed in Section 4 that if  $X$  is underdispersed, then so is any  $\alpha$ -thinning  $X_\alpha$ . However, the variable thinning mechanism defined in Section 4.2 is more flexible and realistic. We show that if the underlying variable  $X$  is Poisson distributed, and the sequence of thinnings  $\alpha_n$  is decreasing but  $n\alpha_n$  is increasing, then the variable  $\alpha$ -thinning  $X_\alpha$  is underdispersed. The choice of  $\alpha_n$  is a source of new underdispersed statistical models. We have reanalyzed the frequencies of dicentrics with the new distribution generated by  $\text{logit}(\alpha_n) = b/n$  providing meaningful results. Another new variable thinning model ( $\alpha_n = \alpha^n$ ) makes it possible to fit the classic fertilization datasets (Morgan (1982)) providing an interesting interpretation.

Two models, one using the service-rate distribution (7) and the other using the Poisson-variable-thinning distribution with  $\text{logit}(\alpha_n) = b/n$ , have been used to fit the frequencies of dicentrics. Which model is preferable? Looking at the results presented in Tables 2 and 3, we

can conclude that, statistically speaking, the Poisson-variable-thinning model is slightly better because the value of its global  $\chi^2$  statistic is 16.27 (12 df), which is lower than the value for the service-rate model, 19.64 (11 df). However, because both models attempt to explain the cell-repair mechanism, extensive biological experiments are required to determine which model is the most accurate.

Finally, Binomial thinning operators provide a natural bivariate stochastic decomposition of the form  $(X_\alpha, X - X_\alpha)$ . In Section 4.2, we explored some relationships between the underdispersion of  $X_\alpha$  and the sometimes negative correlation between  $X_\alpha$  and  $X - X_\alpha$ . In particular, Propositions 6 and 8 allow to construct new bivariate count models, where one marginal is underdispersed, the other is overdispersed and the correlation is negative. We have illustrated some of the results of this section exploring the presence of a specific pair of trinucleotides in nonoverlapping chunks of size 200 of the genome of coronavirus SARS-CoV-2. We have used a new bivariate count distribution generated by a variable thinning where the odds of  $\alpha_n$  is a power function of  $n$ . Interestingly, this model can be fitted with standard software for Generalized linear models.

Many topics remain open to further research. One of these is how to implement regression models to new distributions, like the one presented in (7). One issue is that its parameters cannot be directly interpreted as mean and variance. The COM-Poisson distribution has the same problem. A possible option is to parametrize the distribution in terms of the mean and a suitable dispersion parameter, albeit this may need the numerical solution of a nonlinear system of equations. In general, each distribution should be investigated to see how its parameters may be affected by the covariates.

It can be interesting to investigate order statistics of count distributions as a potential source of underdispersion. A recent article by Badiella et al. (2023), for instance, demonstrates how the Poisson order statistics are underdispersed.

Another interesting aspect is that most count distributions linked to capture–recapture problems are overdispersed and belong to the LC-class, a wide family of distributions endowed with a log-convex pgf (Puig and Kokonendji (2018)). Similarly, it can be directly checked that all count distributions with a log-concave pgf are underdispersed, but this property has no practical interpretation. What kind of mechanism is behind a log-concave pgf? This challenge is a motivation for further research.

**ACKNOWLEDGEMENTS**

The authors thank the two anonymous referees for their valuable comments and suggestions. This work was partially funded by the grant PID2022-137414OB-I00 from the Spanish Ministry of Science, Innovation and Universities and by the Spanish State Research Agency, through the Severo Ochoa and María de Maeztu Program for Centers and Units of Excellence in R&D (CEX2020-001084-M).

**ORCID**

Pedro Puig  <https://orcid.org/0000-0002-6607-9642>

**REFERENCES**

Badiella, L., del Castillo, J., & Puig, P. (2023). Ultra log-concavity of discrete order statistics. *Statistics & Probability Letters*, 201, 109900. <https://doi.org/10.1016/j.spl.2023.109900>  
 Ball, F. (1995). A note on variation in birth processes. *Mathematical Scientist*, 20, 50–55.

- Barlow, R. E., & Proschan, F. (1965). *Mathematical theory of reliability, with contributions by Larry C. Hunter. Reprint of the 1965 original. Classics in Applied Mathematics* (Vol. 17). Society for Industrial and Applied Mathematics (SIAM), 1996.
- Böhning, D. (2016). Ratio plot and ratio regression with applications to social and medical sciences. *Statistical Science*, 31(2), 205–218.
- Cameron, A. C., & Johansson, P. (1997). Count data regression using series expansions: With applications. *Journal of Applied Econometrics*, 12(3), 203–223.
- Conway, R. W., & Maxwell, W. L. (1961). A queuing model with state dependent service rate. *Journal of Industrial Engineering*, 12(2), 132–136.
- Cox, D. R., & Isham, V. (1980). *Point processes*. In *Chapman & Hall/CRC Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.
- Crawford, F. W., Minin, V. N., & Suchard, M. A. (2014). Estimation for general birth-death processes. *Journal of the American Statistical Association*, 109(506), 730–747.
- del Castillo, J., & Pérez-Casany, M. (2005). Overdispersed and underdispersed Poisson generalizations. *Journal of Statistical Planning and Inference*, 134, 486–500.
- Faddy, M. J. (1994). On variation in Poisson processes. *Mathematical Scientist*, 19, 47–51.
- Faddy, M. J. (1997). Extended Poisson process modelling and analysis of count data. *Biometrical Journal*, 4, 431–440.
- Gelfand, A. E., & Dalal, S. R. (1990). A note on overdispersed exponential families. *Biometrika*, 77(1), 55–64.
- IAEA. (2011). *Cytogenetic Dosimetry: Applications in Preparedness for and Response to Radiation Emergencies*. International Atomic Energy Agency.
- Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions*. In *Wiley Series in probability and statistics*. Wiley.
- Kokonendji, C. C., Mizère, D., & Balakrishnan, N. (2008). Connections of the Poisson weight function to overdispersion and underdispersion. *Journal of Statistical Planning and Inference*, 138, 1287–1296.
- Kokonendji, C. C., & Puig, P. (2018). Fisher dispersion index for multivariate count distributions: A review and a new proposal. *Journal of Multivariate Analysis*, 165, 180–193.
- Kokonendji, C. C., Senga Kiessé, T., & Balakrishnan, N. (2009). Semiparametric estimation for count data through weighted distributions. *Journal of Statistical Planning and Inference*, 139, 3625–3638.
- Levine, M. W. (1991). The distribution of intervals between neural impulses in the maintained discharges of chemically isolated ganglion cells in goldfish retina. *Biological Cybernetics*, 65, 459–467.
- McShane, B., Adrian, M., Bradlow, E. T., & Peter, S. (2008). Count models based on Weibull interarrival times. *Journal of Business and Economic Statistics*, 26(3), 369–378.
- Morgan, B. J. T. (1982). Modelling polyspermy. *Biometrics*, 38(4), 885–898.
- Moriña, D., Serra, I., Puig, P., & Corral, A. (2019). Probability estimation of a Carrington-like geomagnetic storm. *Scientific Reports*, 9(1), 2393. <https://doi.org/10.1038/s41598-019-38918-8>
- Patil, G. P., Rao, C. R., & Zelen, M. (2004). *Weighted distributions*. In *Encyclopedia of statistical sciences*. John Wiley & Sons Inc.
- Podlich, H. M., Faddy, M. J., & Smyth, G. K. (2004). Semi-parametric extended Poisson process models for count data. *Statistics and Computing*, 14(4), 311–321.
- Puig, P. (2003). Characterizing additively closed discrete models by a property of their MLEs, with an application to generalized Hermite distributions. *Journal of the American Statistical Association*, 98(463), 687–692.
- Puig, P., & Kokonendji, C. C. (2018). Non-parametric estimation of the number of zeros in truncated count distributions. *Scandinavian Journal of Statistics*, 45, 347–365.
- Puig, P., & Valero, J. (2006). Count data distributions: Some characterizations with applications. *Journal of the American Statistical Association*, 101(473), 332–340.
- Puig, P., & Valero, J. (2007). Characterization of count data distributions involving additivity and binomial subsampling. *Bernoulli*, 13(2), 544–555.
- Puig, P., & Weiss, C. (2020). Some goodness-of-fit tests for the Poisson distribution with applications in Biodosimetry. *Computational Statistics & Data Analysis*, 144, 1–12. <https://doi.org/10.1016/j.csda.2019.106878>
- 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. *PLOS One*, 9(12), e114137. <https://doi.org/10.1371/journal.pone.0114137>

- Ridout, M. S., & Besbeas, P. (2004). An empirical model for underdispersed count data. *Statistical Modelling*, 4(1), 77–89.
- Ross, S. M. (1995). *Stochastic processes*. In *Wiley series in probability and statistics: Probability and statistics* (2nd ed.). Wiley.
- Sellers, K. F., & Morris, D. S. (2017). Underdispersion models: Models that are “under the radar”. *Communications in Statistics—Theory and Methods*, 46(24), 12075–12086.
- Seshadri, V. (1998). *The inverse Gaussian distribution: Statistical theory and applications*. Springer.
- Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C*, 54(1), 127–142.
- Winkelmann, R. (1995). Duration dependence and dispersion in count data models. *Journal of Business and Economic Statistics*, 13, 467–474.
- Wise, J. (1962). The relationship between the mean and variance of a stationary birth-death process, and its economic application. *Biometrika*, 49, 253–255.
- Zeviani, W. M., Ribeiro, P. J., Bonat, W. H., Shimakura, S. E., & Muniz, J. A. (2014). The Gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics*, 41, 2616–2626.

**How to cite this article:** Puig, P., Valero, J., & Fernández-Fontelo, A. (2023). Some mechanisms leading to underdispersion: Old and new proposals. *Scandinavian Journal of Statistics*, 1–23. <https://doi.org/10.1111/sjos.12677>