Journal de la Société Française de Statistique
Vol. 157 No. 2 (2016)

Distributions to model overdispersed count data

Titre: Distributions pour le comptage de données surdispersées

Sylvain Coly[1,2], Anne-Franoise Yao[2], David Abrial[1] and Myriam Charras-Garrido[1]

Abstract: In the early twentieth century, only a few count distributions (binomial and Poisson distributions) were commonly used in modeling. These distributions fail to model bimodal or overdispersed data, especially data related to phenomena for which the occurrence of a given event increases the chance of additional events occurring. New count distributions have since been introduced to address such phenomena; they are named "contagious" distributions. This group of distributions, which includes the negative binomial, Neyman, Thomas and Pólya-Aeppli distributions, can be expressed as mixture distributions or as stopped-sum distributions. They take into account bimodality and overdispersion, and show a greater flexibility with regards to value distributions. The aim of this literature review is to 1) explain the introduction of these distributions, 2) describe each of these overdispersed distributions, focusing in particular on their definitions, their basic properties, and their practical utility, and 3) compare their strengths and weaknesses by modeling overdispersed real count data (bovine tuberculosis cases).

Résumé : Au début du vingtième siècle, seules quelques lois de comptage (loi binomiale, loi de Poisson) sont couramment utilisées en modélisation. Elles trouvent leur limite dans la modélisation de données bimodales ou surdispersées, notamment celles associées à un phénomène dont la survenue engendre d'autres occurrences. Pour les modéliser, de nouvelles lois sont élaborées, dites "lois contagieuses". Ces distributions, telles que la loi binomiale négative, la loi de Neyman, la loi de Thomas ou encore la loi de Pólya-Aeppli, peuvent s'exprimer sous forme de mélange de lois usuelles ou encore de somme de variables aléatoires dont l'indice est borné par une variable aléatoire. Elles permettent d'ajuster des répartitions bimodales, surdispersées et très irrégulières. L'objectif de notre revue de la littérature est de décrire l'apparition de ces lois surdispersées, d'effectuer un descriptif de chacune de ces distributions, en s'intéressant en particulier à leurs différentes caractérisations, à leurs propriétés élémentaires et à leur utilisation potentielle, et enfin de les comparer en les appliquant à des données réelles surdispersées (cas de tuberculose bovine).

Keywords: Count Distributions, Overdispersion, Mixture Distributions, Stopped-Sum Distributions, Negative Binomial Distribution, Neyman Distribution, Thomas Distribution, Pólya-Aeppli Distribution
Mots-clés : Distributions discrètes, Surdispersion, Mélange de lois, Somme finie de distributions, Loi Binomiale Négative, Loi de Neyman, Loi de Thomas, Loi de Pólya-Aeppli
AMS 2000 subject classifications: 60E05, 60-00, 62E15, 92-00, 92C60

1. Introduction

In the early twentieth century, a very small number of discrete probability distributions were used in modeling. Indeed, in their work, Feller and Bliss only discuss binomial and Poisson distributions, which were the only ones used at the time, according to Corbet (albeit alongside the normal distribution, which is not a count distribution) (Feller, 1943; Fisher et al., 1943; Bliss and Fisher, 1953). Although few other discrete distributions existed, such as the discrete

[1] Centre INRA Auvergne / Rhône-Alpes, Unité d'Épidmiologie Animale.
E-mail: sylvain.coly@clermont.inra.fr
[2] Université Blaise Pascal, Laboratoire de Mathématiques, UMR 6620, CNRS

 ISSN: 2102-6238

uniform distribution, they were of limited use in modeling biological data. The binomial distribution describes the number of "successful" events that occur when a yes/no experiment is repeated multiple times. The binomial distribution can be approximated by the Poisson distribution (Le Cam, 1960), which is more simple and frequently employed in modeling (Satterthwaite, 1942).

During this same time period, several studies carried out in different fields of biology showed that these commonly used distributions were not adapted to all situations (Gurland, 1958) and underscored the necessity of employing more flexible and complex distributions, especially when dealing with overdispersed data (Bliss and Fisher, 1953). Consequently, starting in 1920s, many biologists came up with novel probability distributions that were inspired by their specific fields of study. Some of these distributions were referred to as "contagious" distributions; this descriptor appears multiple times in the literature between 1930 and 1960. Oddly enough, the term is used by ecologists, entomologists, bacteriologists, electronics researchers, accidentologists and actuaries, but not by epidemiologists. In general, it was applied in situations that were "contagious" in nature, namely involving phenomena in which the occurrence of a given event (called the primary case) led to the occurrence of additional such events (called the secondary cases) over a circumscribed time period and geographical area. However the definition of mathematical contagion long remained vague. For instance, researchers such as Greenwood and Yule, and later Feller, distinguished between two types of contagion that manifest themselves in the same way: "false" contagions (also referred to as "apparent" contagions) and "true" contagions (Greenwood and Yule, 1920; Feller, 1943; Gurland, 1958).

False contagions are in fact artifacts of population heterogeneity. In false contagions, population features (e.g., age demographics, sex ratios) can explain the clustering of events in certain places; no mechanism of contagion is actually at work. Differences in event risk within a given region can also be explained by environmental factors, such as pollution levels or the local climate. The general consensus in the literature is that, to deal with "false contagion," one should use a Poisson distribution where the parameter is allowed to vary to account for heterogeneity in risk levels within the study area (Satterthwaite, 1942). Thus, this parameter can take on any number of fixed values, with defined probabilities, or act as a random variable following a given distribution. The idea of coupling the Poisson distribution to other distributions was first proposed by Greenwood and Yule (Greenwood and Yule, 1920). In such cases, the resultant distribution is referred to as a "compound" Poisson distribution, a term coined by Feller (Feller, 1943). Compound distributions are now mainly known as mixture distributions.

In the case of a "true" contagion, we harken back to the epidemiological roots and the more generally accepted sense of the word, i.e. an initial primary case results in secondary cases. For instance, Gurland makes mention of a "model of random colonies" (Gurland, 1963). In the literature, several modeling approaches are described; however all these models involve summing or integrating random variables, with bounds that are themselves often random variables. In most cases, the process involves the summation of random variables whose cardinality follows a Poisson distribution. The resulting distribution is described in the literature as a generalized Poisson distribution (Feller, 1943) and is now mainly known as Poisson stopped-sum distribution. It should be noted that the concept of the generalized Poisson distribution discussed here may

 ISSN: 2102-6238

differ from that encountered in more recent publications and in the end of this paper (Section 7) (Consul and Jain, 1973; Joe and Zhu, 2005).

In the literature, the Pólya urn model also emerges as a source of novel contagious distributions. Pólya's distribution model, which is based on the idea of drawing balls from an urn, is one of the applications of his research on integration (Polya, 1930). Pólya began with the simple idea that an individual infected by a given contagious disease has a chance of transmitting the disease to those around him. Thus this theoretical urn is adapted to a situation of contagion: drawing an "infected" individual, under certain conditions, increases the chances of subsequently drawing other "infected" individuals. Pólya felt that a symmetrical representation of the situation would facilitate the mathematics; thus, similarly, drawing a "healthy" individual increases the probability of drawing other "healthy" individuals. Both Woodbury and Rutherford developed distributions inspired by this model (Woodbury, 1949; Rutherford, 1954).

However, the dichotomy between "true" contagion and "false" contagion is actually a false one. Coming from two opposing philosophical standpoints, both from an applied and a modeling perspective, Greenwood and Yule ("falsely contagious" compound models) (Greenwood and Yule, 1920) and Eggenberger and Pólya ("truly contagious" generalized models) (Eggenberger and Polya, 1923) ultimately arrived at the same distribution, as Feller underscored (Feller, 1943). Furthermore, this distribution can also be viewed as emerging from the Pólya urn model. Consequently, most contagious distributions can be viewed as compound and/or generalized Poisson distributions. The large number of distributions that can be obtained using these two approaches can be explained by both: 1) the Gurland theorem (Gurland, 1959, 1963); 2) the fact that the functions that generate Poisson, binomial, or negative binomial distributions share similarities, which means that they readily lend themselves to the creation of stopped-sum distributions that can also appear in the form of mixture distributions.

Feller's assessment that "true" and "false" contagions in fact form a false dichotomy raises questions about how contagious distributions are defined and the very nature of contagion itself. Underlying all contagious distributions is the idea that if one event occurs, the probability of other such events occurring in the future is greatly increased (i.e., $\mathrm{P}(X > 1|X \geq 1) \gg \mathrm{P}(X > 1)$). This idea, which is vague and never formally defined in the literature, is oft repeated in publications on the topic. The definition of a contagious distribution can also be more formal and adopt a specific form (e.g., sum, mixture, etc.), as in the case of compound Poisson models or generalized Poisson models. However, while the Zero-Inflated Poisson (ZIP) distribution can be viewed as Poisson-Bernoulli mixture, it is not considered to be contagious. Another feature that is clearly relevant when discussing situations of contagion, but that is not mentioned in any of the literature on this topic, is the non-unimodal nature of the count distributions. Indeed, a contagious disease with a low incidence could result in small clusters containing a few cases each. In such a situation, distributions with a mode of 0 and a mode that is the average size of these small clusters would provide a good fit. That said, a ZIP distribution can be bimodal without being contagious, while a negative binomial distribution is inherently unimodal. While the ability to have one or more modes is certainly useful, it unfortunately cannot be used as one of definitive traits of contagious distributions. Establishing a formal definition of contagious distributions is challenging, and it is

 ISSN: 2102-6238

difficult to identify features that they share exclusively.

A basic definition of contagion can be found in both the Oxford Dictionary and the "Dictionnaire de l'Acadmie franaise" (Dictionary of the French Academy) (Académie_Française, 1935; Stevenson and Lindberg, 2015): 1) *The communication of disease from one person or organism to another by close contact;* 2) *A disease spread by close contact;* 3) *The spreading of a harmful idea or theory*. From an epidemiological perspective, contagion is described as a situation in which an infected patient, referred to as the primary case, subsequently infects others, referred to as secondary cases, via direct, indirect, or vector-borne transmission. A primary case can also be defined as a case of unknown origin, a *de facto* patient zero because no preceding cases can be identified. It can also be a case whose occurrence is independent and random, statistically speaking, from other cases in the dataset. Alternatively, it might be the first case caused by a specific pathogenic strain, which means that the evolutionary time scale comes into play. Actually the definition of contagion is rather dependent on which infections are deemed to be primary versus secondary, a challenge that is tied to the notion of proximity (e.g., direct versus indirect contact). Because proximity is a relative concept, it is difficult to establish a concrete definition for contagion, and the way in which contagion manifests itself also depends on temporal and spatial scales. For instance, depending on the scale chosen, the primary case and the secondary cases may or may not be located in the same area and/or time period. Therefore, contagion can lead to the following patterns in epidemiological count data associated with a given space: 1) atypical case clustering (i.e., local overdispersion); 2) case aggregation (i.e., spatial structure); 3) case persistence (i.e., temporal structure); and 4) case diffusion or propagation (i.e., spatiotemporal structure). This latter pattern provides the most evident example of contagion, since propagation and contagion are strongly related. In purely spatial analyses, however, only local overdispersion and case aggregation provide evidence of contagion.

The definition of contagion is dependent on the notions of proximity and contact, which are both relative in nature. Contagion can involve spatial and/or temporal autocorrelations and local overdispersion, but only the overdispersion can be modeled with the count distribution exclusively. Statistically, dispersion is defined as the variance divided by the mean. Data are said to be overdispersed when this value is greater than one, which means that the variance is greater than the mean. For instance, when modeling the occurrence of a quite uncommon phenomenon, the Poisson distribution most often used has a mode of zero and is strictly decreasing and convex; therefore, when overdispersion is present, the distribution will become relatively more convex. In contrast to a set of Poisson-distributed data with the same mean, overdispersed data will contain a greater number of zeros and extreme values and fewer intermediate values. Overdispersion can have several different root causes: 1) sampling issues (e.g., under- or oversampling, reporting rates, data quality); 2) population demographics and structuration (e.g., based on age, sex, profession); 3) environmental factors (e.g., weather, rate of urbanization, vegetation); 4) situations of contagion, i.e. a case potentially results in further cases in the same area and during the same period.

A shared feature of the contagious distributions is that they can account for overdispersion. However, contagious distributions themselves are never clearly defined nor is a definitive link established with the two aforementioned concepts (contagion and overdispersion) even if they are

 ISSN: 2102-6238

more commonly referenced in contexts of overdispersion. However, many distributions allow for overdispersion but do not necessarily fall into the category of contagious distributions. One notable example is zero-inflated (ZI) distributions, in which the proportion of null values is increased to model situations in which data are missing or censored, as well as those in which phenomena are underdetected or underreported. Ultimately, this overdispersion can result from other factors (e.g., missing data, population demographics, or environmental factors). Consequently, it is difficult to assign any characteristic traits to contagious distributions. Given this intellectual impasse, we turn our attention to overdispersed distributions.

First, we will address the negative binomial distribution (Section 2, page 43), paying particular attention to Fisher's work. Next, we will focus on the distributions developed by Neyman (Section 3, page 46), and Thomas (Section 4, page 49). We will then move on to the Pólya-Aeppli distribution and the distributions derived from the Pólya urn model, notably the Woodbury and Rutherford distributions (Section 5, page 50). Lastly, we will briefly list other overdispersed mixture distributions (Section 6, page 53), as well as simple overdispersed distributions (Section 7, page 56), that have not been clearly been identified as contagious, despite their inclusion of overdispersion. A numerical application is also presented; an overdispersed dataset (bovine tuberculosis cases in France between 2001 and 2010) is modeled by 7 of the distributions mentioned above (Section 8). This illustration provides an interesting comparison of these distributions and give some clues about their strengths and shortcomings. Finally a conclusion highlights the properties of the overdispersed distributions, and the context in which each of them can be used.

2. Negative binomial distribution

Poisson-Pearson type III distribution. Greenwood and Yule (Greenwood and Yule, 1920), as well as Eggenberger and Polya (Eggenberger and Polya, 1923; Eggenberger, 1924) came up with the same distribution, which is a mixture of a Poisson distribution and a Pearson type III distribution (Pearson, 1895). The Pólya distribution has been applied in such contrasting fields as epidemiology and accident prevention (Lundberg, 1940; Rosenblatt, 1940; Kitagawa and Huruya, 1941). Satterthwaite comments that the Poisson-Pearson type III distribution is the most frequently used mixture distribution and that it provided a good fit for heterogeneous data (Satterthwaite, 1942). As the gamma distribution is a special case of the Pearson type III distribution, then the negative binomial distribution is a special case of the Poisson-Pearson type III mixture distribution. It has many practical applications, and it will be the subject of focus in this section.

Negative binomial distribution. The negative binomial distribution, noted $\mathscr{NB}$, was formalized by Rémond de Montmort in 1714 (Rémond de Montmort, 1713). If $X \sim \mathscr{NB}(n,p)$ when $n \in \mathbb{N}$ and $0 < p < 1$,

$$\mathrm{P}(X=k) = \left(\begin{array}{c} n+k-1 \\ k \end{array} \right) p^n (1-p)^k$$

when $k \in \mathbb{N}$. Feller named it Pascal distribution (Feller, 1957; Gurland, 1957), a name still used for the distribution in situations in which $n \in \mathbb{N}$. Indeed, this distribution can be extended to $n \in \mathbb{R}+$.

 ISSN: 2102-6238

In this case, the negative binomial distribution is defined by

$$\mathrm{P}(X=k)=\frac{\Gamma(n+k)}{k!\Gamma(n)}p^n(1-p)^k,$$

for $k\in\mathbb{N}$ and can be named Pólya distribution.

In the context of a Bernoulli process, where the probability of success is $1/(1+p)$, this distribution can be used to describe the distribution of X, the number of trials to conduct above and beyond n to obtain n successes (Gurland, 1959). Here, the mean and variance of X are found using $\mathbb{E}(X)=n(1-p)/p$ and $\mathrm{Var}(X)=n(1-p)/p^2$, respectively, which means that the variance is $\mathrm{s}(X)=1/p$ times greater than the mean, allowing overdispersion. The single mode is 0 if $n\leq 1$, and $\lfloor(n-1)(1-p)/p\rfloor$ if $n>1$.

Both parameters can be estimated with the moments method: $\tilde{n}=\bar{X}/(s(X)-1)$ and $\tilde{p}=1/s(X)$. An alternative method is the method of mean-and-zero-frequency. *In extenso* p^* is obtained solving $p^*\ln(p^*)/(1-p^*)=\ln(f_0)/\bar{X}$, where f_0 is the empirical proportion of null values, then n^* is computed using the expression of the mean: $n^*=\bar{X}p^*/(1-p^*)$. The parameters can also be computed with the maximum likelihood estimation. However the differentiate of the likelihood function with respect to n results in an equation which has no closed-form solution, thus the root is numerically computed using for instance Brent's algorithm (Brent, 2002). Even if these methods provide robust estimations of the parameters, there are regions where the methods based on the mean and the variance or the zero frequency are less efficient (Anscombe, 1950). Sophisticated methods have been proposed to estimate parameters in these cases, such as the digamma method or the generalized minimum Chi-Squared method.

In his first publications, Gurland viewed the negative binomial distribution as a false contagious distribution; indeed, it is considered to be a Poisson distribution for which the parameter varies according to population demographics and, in this case, follows a gamma distribution (Feller, 1943; Gurland, 1959). The result is therefore a mixture of a Poisson distribution and a gamma distribution, which is expressed as

$$X\sim\mathscr{P}(L)\text{ where }L\sim\Gamma\left(p,\frac{p^2}{n}\right).$$

This hierarchical Poisson-gamma model was developed by Clayton (Clayton and Kaldor, 1987). However, it should be noted that it is actually just a special case of a Poisson-Pearson type III mixture distribution.

The negative binomial distribution can also be used to describe data in the fields of entomology and bacteriology (Jones et al., 1948). A classic example of a contagious distribution involves the production of larvae: groups of eggs appear randomly and each results in a random number of larvae (Neyman, 1939; Skellam, 1952; Evans, 1953). Generally, the frequency of egg groups can be modeled using a Poisson distribution, while larvae number follows a logarithmic distribution. Fisher introduced the logarithmic distribution to model the size of larvae populations. More precisely, he proposed to model the number of species with different numbers of representatives

 ISSN: 2102-6238

(i.e., 1, 2, 3, 4, and so on) with the terms of a Taylor series for the logarithmic function. If X follows a logarithmic distribution with parameter τ, considering $\alpha = -1/\ln(1-\tau)$,

$$\mathrm{P}(X=k) = \begin{cases} 0 & \text{if } k=0, \\ \alpha \frac{\tau^k}{k} & \text{when } k \in \mathbb{N}^*. \end{cases}$$

The total number of larvae produced therefore follows a negative binomial distribution, which is actually a generalized Poisson distribution (according to Feller's definition). This distribution was also used to describe counts of insects and yielded a better fit than the Neyman (type A) and Pólya-Aeppli distributions. Anscombe identified examples of study systems in which the negative binomial distribution could be applied (Anscombe, 1950). He showed how several different approaches, with very different starting points, could be used to arrive at the same result. In particular, he mentioned that the negative binomial distribution could result from the summation of logarithmic variables whose index is constrained by a variable following a Poisson distribution (von Luders, 1934; Quenouille, 1949). The negative binomial also appears in descriptions of growing populations, when rates of mortality, natality, and migration are assumed to be constant (McKendrick, 1914; Kendall, 1949). This approach can also be extended to describe the spread of an infectious disease within a community. In this case, the negative binomial distribution appears in the form of a stopped-sum distribution, which is the sum of independent and identically distributed random variables, the number of which is also a random variable. Therefore,

$$X = \sum_{i=0}^{Z} Y_i \text{ where } Z \sim \mathscr{P}(\lambda) \text{ and } Y_i \sim \mathrm{Log}(p),$$

where $\lambda > 0$ and $0 < p < 1$.

Gurland cites one of Feller's later publications to show that a negative binomial distribution can have "contagious" features, and Feller's work was originally inspired by Pólya (Polya, 1930; Feller, 1957; Gurland, 1959). Starting with a Pólya urn model, the negative binomial distribution emerges from the Pólya distribution (equation (1), page 50) as $n \to +\infty$, $p \to 0$ and $\alpha \to 0$, using the notations defined in Section 5. Gurland deduced from his discovery that it is not always possible to distinguish between true and false contagious distributions based on the data alone.

Gurland considered the negative binomial to be a contagious distribution, as is clear from the title of one of his papers (Gurland, 1959). He cites several examples of data in ecology, bacteriology, microbiology, epidemiology, and accidentology (Adelstein, 1952; Evans, 1953) for which the negative binomial distribution, among other discrete distributions, is particularly well suited. This distribution is also recommended in econometrical regression models of overdispersed count data (Grogger and Carson, 1991; Kennedy, 1994). According to Gurland, the most commonly encountered count distributions are the binomial, Poisson, and negative binomial (Gurland, 1959). Fisher felt that the negative binomial provided a poor fit if an excessive number of zeros were present and that if the proportion of unobserved species was known, then it was preferable to include it in the model (Fisher et al., 1943). In such situations, a ZI distribution may better serve.

 ISSN: 2102-6238

3. Neyman distribution

This distribution was developed by Neyman to describe certain biological datasets, including counts of larvae treated with lethal chemicals and bacteria growing in Petri dishes. Although these two examples come from the fields of entomology and bacteriology, respectively, Feller noted that the distribution could be applied in other contexts, such as risk theory, insurance science, and traffic control. Although the Neyman distribution addresses some of the shortfalls of the Poisson distribution, that was not its original intent. Rather, the challenge was to come up with a means of carrying out Student's t-tests (Hsu, 1938) and Fisher's exact tests when sample variances were unequal. Neyman's goal was therefore to come up with a distribution class that could be used with data demonstrating a high level of variance. Neyman was working with datasets in which zeros were very frequent but ones were sparse. The Poisson distribution therefore provided a poor fit, and Neyman decided to take a closer look at contagious distributions (as per Polya, 1930). The distribution Neyman came up with was shaped by several hypotheses related to larva traits (e.g., social behavior, life span, and degree of mobility). He then simplified his model by utilizing characteristic densities for some of the distributions involvedusual distributions in the development of the general equation describing his distribution. His work ultimately yielded several classes of distributions (e.g., type A, type B, type C, etc.) that differed in parameter number.

Two-parameter type A Neyman distribution. For a random variable X following this type of distribution, which has parameters m_1 and m_2, and where k is a natural integer:

$$\mathrm{P}(X=k)=\mathrm{e}^{-m_1}\frac{m_2^k}{k!}\sum_{i=0}^{+\infty}\frac{(m_1\mathrm{e}^{-\mathrm{m}_2})^i}{i!}i^k.$$

This distribution can be viewed as an extension of the Poisson distribution. It can be seen that:

$$\mathrm{P}(X=0)=\mathrm{e}^{-m_1(1-\mathrm{e}^{-m_2})},\ \mathrm{P}(X=1)=\mathrm{e}^{-m_1(1-\mathrm{e}^{-m_2})}m_2m_1\mathrm{e}^{-m2},$$

$$\mathrm{P}(X=2)=\mathrm{e}^{-m_1(1-\mathrm{e}^{-m_2})}\frac{m_2^2}{2}(m_1^2\mathrm{e}^{-2m_2}+m_1\mathrm{e}^{-m_2}).$$

These expressions clearly reveal that when a given event occurs, it increases the probability of other such events occurring as well.

A strong recurrence relation was established by Beall:

$$\mathrm{P}(X=k+1)=\frac{m_1m_2\mathrm{e}^{-m_2}}{k+1}\sum_{i=0}^{k}\frac{m_2^i}{i!}\mathrm{P}(X=k-i).$$

According to Neyman, this relation was used several times with satisfactory results (Neyman, 1939). The two-parameter type A Neyman distribution is known as the Neyman distribution. The parameters m_1 and m_2 are in general noted λ and θ.

It would appear that Neyman type A distributions are in fact a type of Poisson-Poisson mixture, based on Gurland's theorem (Massé and Theodorescu, 2005), which means that they can also be expressed in the following way:

$$X\sim\mathscr{P}(L\theta)\text{ when }L\sim\mathscr{P}(\lambda),$$

 ISSN: 2102-6238

where $\lambda > 0$ and $\theta > 0$.

It is assumed that the presence of a larva at a given location is associated with an increased probability of other larvae occurring nearby, with the data following a Poisson distribution. This situation can be described as follows:

$$X = \sum_{i=0}^{Z} Y_i \text{ where } Z \sim \mathscr{P}(\lambda) \text{ and } Y_i \sim \mathscr{P}(\theta).$$

Here $\lambda > 0$ and $\theta > 0$ are positive parameters that depend, respectively, on the expected occurrence of the primary case and on the contagion dynamics that will lead to secondary cases. The resulting distribution is indeed a Neyman distribution.

Keeping with the notation system used for the compound and generalized Poisson-Poisson distributions, $\mathbb{E}(X) = \lambda\theta$ et $\mathrm{Var}(X) = \lambda\theta(\theta+1)$. As expected in a situation of contagion, Neyman type A distributions only represent overdispersion using a Poisson model: overdispersion is described by $\mathrm{s}(X) = \theta + 1$. Neyman distributions can have different numbers of modes, either a single mode equal to 0 or another value, two modes, or multiple modes if the model's two parameters have large values. In the latter case, the distribution will adopt a very uneven shape. The computation of the order 2 moments provides an estimation of the parameters : $\tilde{\lambda} = \bar{X}/(s(X)-1)$ and $\tilde{\theta} = s(X) - 1$.

Three-parameter type A Neyman distribution. A more flexible distribution can be generated if the study system is separated into two parts, each with a different *a priori* probability associated with encountering a larva. This distribution can be used to model study systems with two very different environments or two very different subpopulations. For a random variable X described by this type of distribution, which has parameters m_1, m_2, and m_3, there is the following strong recurrence relation (Neyman, 1939) :

$$\mathrm{P}(X=0) = \mathrm{e}^{-m_1(1-\frac{1}{2}(\mathrm{e}^{-m_2}+\mathrm{e}^{-m_3}))},$$

$$\mathrm{P}(X=k+1) = \frac{m_1}{2(k+1)} \sum_{i=0}^{k} \frac{m_2^{i+1}\mathrm{e}^{-m_2} + m_3^{i+1}\mathrm{e}^{-m_3}}{i!} \mathrm{P}(X=k-i),$$

with $k \in \mathbb{N}^*$.

***k*-parameter type A Neyman distribution.** The distribution can be further generalized to include any number of parameters based on the framework for the three-parameter model, in order to handle a heterogeneous environment in which encounter probabilities differ (in Neyman's case, for larvae) (Gurland, 1958; Subrahmaniam, 1966). From an epidemiological perspective, Neyman type A distributions can be used to examine primary cases with different probabilities of occurrence; in contrast, the frequency of secondary cases follows a Poisson distribution whose parameter is constant. Neyman type A distributions, as well as the Thomas distribution (Section 4) and a large number of other count distributions, converge to a normal distribution. Furthermore, the sum of random variables following a Neyman type A distribution in turn follows a Neyman

 ISSN: 2102-6238

type A distribution whose mean and variance are equal to the sums of the term means and variances, respectively (Teich, 1981).

Neyman type A distributions have been applied by ecologists to describe the distribution of plants in randomly selected quadrats. In these situations, the Poisson distribution fails to model clustering of individuals, a pattern that is observed in several species (Archibald, 1948). As a result, such systems display overdispersion, where the variance is much greater than the mean (Thomas, 1949). In Archibald's study, parameters λ and θ of the two-parameter Neyman type A distribution are proportional to the number of clusters and the mean number of plants per cluster, respectively. Evans concluded that the Neyman distribution provided a better fit for plant count data than did the negative binomial and Pólya-Aeppli distributions (Evans, 1953). He then proposed that plant overcrowding and strong levels of competition could help explain the good fit of this distribution in this context. Teich underscored the widespread applicability of this distribution and specifically cited the example of galaxy distribution in space (Teich, 1981). Moreover, he considered that this distribution was well suited to describing counts of events generated by multiplicative Poisson processes, as long as said events occur a limited number of times. This distribution has also demonstrated its usefulness in electronics research on semi-conductors. Semi-conductors have highly inconsistent yields resulting from variation in defect density – clusters of defect can be modeled with contagious distributions (Park and Jun, 2000, 2002). Massé and Theodorescu have suggested that unimodal distributions could be useful when dealing with counts of infectious disease cases (Massé and Theodorescu, 2005). In their paper, they study Neyman distributions; they focus specifically on the two-parameter distribution and discuss contexts in which it is unimodal.

Type B and type C Neyman distributions. Using slightly different characteristic functions (based on *a priori* differences in larva behavior), Neyman arrived at type B and C distributions (Neyman, 1939). However, based on the literature, they appear to have been used and studied much less often than their type A counterparts. *In fact, no information is provided on the situations in which they were used, their strengths, or their weaknesses. It is only mentioned that their relevance could be examined by comparing the results they provide with real observational data.*

Generalized Neyman distributions. Beall and Rescia developed a series of distributions for which the three types of Neyman distributions are the three first terms (Beall and Rescia, 1953). This series converges to a Pólya-Aeppli distribution (Gurland, 1958). Gurland adopted a different approach, inspired by the observation that not all the larvae in a Petri dish can be easily viewed. The idea was that one can approximate the number of groups of individuals with a Poisson distribution where group size follows a Poisson distribution with rate parameter λ. Such a situation is functionally equivalent to a Neyman type A distribution. For example, if one assumes that in each group of eggs, each egg has a probability p of being randomly observed, then one is in the approximate realm of a two-parameter Neyman distribution whom second parameter is $p.\lambda$. The generalized distribution is achieved by making p a random variable that follows a beta distribution (Gurland, 1958). Gurland applied this distribution to insect counts and demonstrated its suitability. He also showed that, in certain situations, the distribution can amount to a special case of the Neyman type A or Pólya-Aeppli distributions.

 ISSN: 2102-6238

4. Thomas distribution

The Thomas distribution is an alternative to the Neyman distribution that was developed specifically for ecological data (e.g., the abundance of different plant species within quadrats) and that was influenced by Archibald's work (Neyman, 1939; Archibald, 1948). The approach is very similar to that of the Neyman distribution; however, a shifted Poisson distribution is used because it is assumed that a cluster of cases contains at least one case.

As in the case of the Neyman distribution, the Thomas distribution can be expressed as a mixture, *in extenso* the compound shifted Poisson-Poisson distribution:

$$X \sim \mathscr{P}_s(L\theta) \text{ where } L \sim \mathscr{P}(\lambda),$$

when $\lambda > 0$ and $\theta > 0$, thus facilitating implementation. Here, $\mathscr{P}_s$ is a shifted Poisson distribution with only positive values.

Using a Poisson distribution for both of the distributions, occurrence of a non-contact case and contamination), the model takes the following form:

$$X = \sum_{i=0}^{Z} Y_i \text{ where } Z \sim \mathscr{P}(\lambda) \text{ and } Y_i \sim \mathscr{P}_s(\theta),$$

when $\lambda > 0$ and $\theta > 0$. Essentially, the individuals present in a given area are split up into several groups, resulting in the following:

$$\mathrm{P}(X=0) = \mathrm{e}^{-\lambda},\ \mathrm{P}(X=1) = \lambda \mathrm{e}^{-(\lambda+\theta)},\ \mathrm{P}(X=2) = \lambda \mathrm{e}^{-(\lambda+\theta)} \left(\theta + \frac{\lambda \mathrm{e}^{-\theta}}{2}\right).$$

Using the same notation system as for the compound and generalized distributions, $\mathbb{E}(X) = \lambda(\theta+1)$ and $\mathrm{Var}(X) = \lambda(\theta^2+3\theta+1)$. Keeping with the idea of contagion, the Thomas distribution only allows overdispersion to be modeled using a Poisson series. This relationship is conveyed by $\mathrm{s}(X) = \theta + 1 + \theta/(\theta+1)$, which is very similar to the expression used for the Neyman distribution. Additionally, both the random variables and their sum follow a Thomas distribution, where the mean and the variance are equal to the sums of the means and variances of the terms, respectively (Teich, 1981). As in the case of the Neyman distribution, the number of modes for the Thomas distribution can be one (either at 0 or another value), two, or more than two (if the two parameters have large values), which results in a highly uneven distribution pattern. One of the inherent features of the Thomas distribution is that, compared to the Neyman distribution, the appearance of slightly larger extreme values is allowed. The parameters can be estimated with the moments method: $\tilde{\lambda} = 2\bar{X} / \left(s(X) + 1 + \sqrt{5 - 2s(X) + s(X)^2}\right)$ and $\tilde{\theta} = \left(s(X) + 1 + \sqrt{5 - 2s(X) + s(X)^2}\right)/2$ which are well-defined if the mean is strictly positive and if the data are overdispersed.

The Thomas distribution allows for many zeros while simultaneously displaying a mode at a much larger value. Teich considered that, like the Neyman distribution, the Thomas distribution

 ISSN: 2102-6238

could be applied in many different situations; he cited the examples of the distribution of larvae across a counting grid and of galaxies across space (Teich, 1981). He felt that the Thomas distribution was appropriate for modeling multiplicative Poisson processes, as long as event repetition was relatively limited with respect to study duration.

5. Contagious distributions due to Pólya

The Pólya-Aeppli, Woodbury and Rutherford distributions all arise from the work of Pólya. On the one hand, the Pólya-Aeppli distribution was first introduced by Aeppli, who was at that time a student of Pólya, in the context of his work about Markovian processes and then studied by Pólya (Aeppli, 1924; Polya, 1930). On the other hand, the Woodbury and Rutherford distributions arise naturally from the Pólya urn model. Like the Neyman and Thomas distributions, they are contagious distributions. However, they involve a fundamentally different approach; indeed, they do not seek to model biological phenomena. They are generated instead using probabilities calculated using models. The approximated distributions are used to build the desired model rather than to construct equations describing biological data. The goal is to generalize a Bernoulli process and its associated distributions (i.e., the Bernoulli and the binomial).

In this model, balls are being drawn from an urn that contains N balls ($N \in \mathbb{N}$) of which pN are white ($0 < p < 1$) and qN are black ($q = 1 - p$). A successful event is defined as the drawing of a white ball. When a ball is drawn $1 + N\alpha$ balls of the same color are placed in the urn. After n draws, if k successes have been recorded, the probability of drawing a white ball is $(p + k\alpha)/(1 + n\alpha)$. This probability actually depends on two parameters: 1) the initial proportion of white balls p in the urn and 2) the value of δ, whose sign and absolute value respectively indicate the direction and strength of the relationship between a given drawing event and those that preceded it. The distribution of the number of white balls drawn can be called the Pólya distribution and is defined by

$$\mathrm{P}(X = k) = \binom{n}{k} \frac{\prod_{i=0}^{k-1}(p + i\alpha)\prod_{j=0}^{n-k-1}(q + j\alpha)}{\prod_{l=0}^{n-1}(1 + l\alpha)} \tag{1}$$

when $k \in [\![0, n]\!]$; its parameters are p, α and n. In particular, successes and failures are "contagious" as soon as $\alpha > 0$. In this system, which clearly has inherent contagious properties, balls are randomly drawn from an urn following a specific resampling approach. According to Feller, this approach and, by consequence, any emergent probability distributions fall within the "true" contagion category (Feller, 1943).

5.1. *Pólya-Aeppli distribution*

The Pólya-Aeppli distribution (noted $\mathscr{PA}$) is often referred to as one of the classic contagious distributions (Aeppli, 1924; Polya, 1930). That said, it is used far less frequently than the negative binomial or the Neyman distributions. The Pólya-Aeppli distribution has particularly been used to describe systems modelled by Markovian processes (Aeppli, 1924), such as biological processes (Nuel, 2008) or traffic accidents (Özel and Ceyhan, 2010). The Pólya-Aeppli distribution, also

 ISSN: 2102-6238

referred to as the geometric Poisson distribution (Sherbrooke, 1968), is formed by the sum of a set of shifted geometric variables whose cardinality follows a Poisson distribution. It is characterized by two parameters, λ and p, and can be expressed as follows:

$$\mathrm{P}(X=k) = \mathrm{e}^{-\lambda} \sum_{j=1}^{k} \left(\begin{array}{c} k-1 \\ j-1 \end{array} \right) \frac{(p\theta)^j}{j!} (1-p)^{k-j},$$

with $k \in \mathbb{N}$. Another possibility is to arrive at a stopped-sum Poisson distribution *via* a geometric distribution:

$$X = \sum_{i=1}^{Z} Y_i \text{ where } Z \sim \mathscr{P}(\lambda) \text{ and } Y_i \sim \mathscr{G}(p),$$

when $\lambda > 0$ and $0 < p < 1$. Furthermore, the geometric distribution itself can be expressed as a Poisson-exponential mixture. However, a major inconvenience of the Pólya-Aeppli distribution is that it cannot be expressed directly as a compound distribution, thus complicating its use in certain situations.

The equations for the mean and the variance are $\mathbb{E}(X) = \lambda/(1-p)$ and $\mathrm{Var}(X) = \lambda(1+p)/(1-p)^2$, respectively. The index of dispersion is therefore $\mathrm{s}(X)(1+p)/(1-p) > 1$, which means only overdispersion can be modeled. Furthermore, it can be seen that $\mathscr{PA}(\lambda, 1) = \mathscr{P}(\lambda)$. For small values of λ and values of p that are close to 1, there will be a single mode at 0. The mode will adopt greater values as λ, increases, and for small values of p, there will be a much greater degree of overdispersion, making it possible to fit the model to multimodal data.

The parameters of this distribution cas be obtained with the moments method: $\tilde{\lambda} = 2\bar{X}/(s(X)+1)$ and $\tilde{p} = (s(X)-1)/(s(X)+1)$. The mean-and-zero-frequency method ($\lambda^* = -\ln(f_0/N)$ and $p^* = 1-\lambda^*/\bar{X}$) and the first-two-frequencies estimation ($\lambda' = -\ln(f_0/N)$ and $p' = -f_1/(f_0 \ln(f_0/N))$) can also be used (where $N = \max(k \in \mathbb{N}/f_k \neq 0)$). The parameters can also be numerically computed using the Maximum Likelihood Estimation by solving the system of equations $\hat{\lambda}/(1-\hat{p}) = \bar{X}$ and $\sum_{k=1}^{N} (f_k(k-1)\widehat{g_{k-1}})/(N\widehat{g_k}) = \bar{X}$, where f and $\hat{g}$ are respectively the empirical frequency and the estimated Pólya-Aeppli probability mass.

5.2. *Woodbury distribution*

The Woodbury distribution is a generalized case of a Bernoulli process where the probability of a successful event depends on the occurrence of previous successes (Woodbury, 1949). This distribution can also be arise from the Pólya's urn. It can be represented by a situation in which an urn contains both white and black balls; white balls are replaced after they are drawn, but black balls are not. The notation is the following: n is the number of trials, k is the possible number of successes ($k \leq n$), $\mathrm{P}_n(k)$ is the probability of k successes occurring over the course of n trials, p_k is the probability of an event occurring after k successes, and $q_k = 1 - p_k$. Here $\mathrm{P}_n(k) = 0$ when $k < 0$ or $k > n$, and $\mathrm{P}_0(0) = 1$. Woodbury started with the recurrence relation

$$\mathrm{P}_{n+1}(k+1) = p_k \mathrm{P}_n(k) + q_{k+1} \mathrm{P}_n(k+1), \tag{2}$$

 ISSN: 2102-6238

and, if $\lambda_k = p_k.n$, an approximation of $\mathrm{P}_n(k)$ is obtained using

$$\mathrm{P}(k) = \sum_{i=0}^{k} \left(\mathrm{e}^{-\lambda_i}. \prod_{j \in [\![0;n]\!] \setminus i} \frac{\lambda_i}{\lambda_i - \lambda_j} \right)$$

at large values of n

This model can be used in ecology to study the distribution of the number of plants within a given area (Woodbury, 1949). In cases where the probability of success of a given trial is low, an approximation has been developed, which is based on a Poisson distribution that is used to approximate a binomial distribution.

5.3. Rutherford distribution

Rutherford also used the Pólya's urn as Woodbury and started with the same recurrence relation (2). However, he assumed that when a white ball was drawn, it was replaced in the urn with α other white balls (Rutherford, 1954). In this approach, there is a clustering of secondary cases around the primary case (the ball drawn). In contrast to Woodbury, Rutherford used the simplification $p_k = p + \alpha k$, which implies that $n < q/\alpha$ if $\alpha > 0$, and $n < -p/\alpha$ if $\alpha < 0$. Based on Woodbury's solution,

$$\mathrm{P}_n(k) = \frac{1}{k!} \prod_{i=1}^{k-1} \left(\frac{p}{\alpha} + i \right) \sum_{i=0}^{k} (-1)^i \binom{k}{i} (q - \alpha i)^n,$$

where n satisfies the former condition and $k \in \mathbb{N}$.

For this distribution, where X follows a Rutherford distribution, noted $\mathscr{R}$, with parameters p, α and n, the mean and variance are found using $\mathbb{E}(X) = (p/\alpha).[(1+\alpha)^n - 1]$ and $\mathrm{Var}(X) = (p/\alpha)[(p/\alpha).(1+\alpha)^{2n} - (p/\alpha + 1).(1+2\alpha)^n + (1+\alpha)^n]$, respectively. The spread of this distribution depends on the values of the parameters and especially the sign of α.

Rutherford notes that $(1+\alpha)^{2n}$ and $(1+2\alpha)^n$ are fairly similar for rather small values of α and n. The Rutherford distribution can therefore be approximated by a negative binomial, or more precisely $\mathscr{NB}(n', p') \approx \mathscr{R}(p, \alpha, n)$ where $n' = p/\alpha$ and $p' = (1+\alpha)^n - 1$. The necessary conditions are that $n^2\alpha^2$ has a negligible value and/or that $p/\alpha < n - 1$. The first condition is rarely fulfilled in practice (because n tends to be large), but the second is often satisfied. It is tricky to interpret n' and p' from the perspective of the Pólya urn model. While it is theoretically possible to approximate the Rutherford distribution using a binomial distribution, the conditions that would need to be met (notably $p/|\alpha| < n - 1$) for the binomial are the opposite of the conditions that define the Rutherford distribution. Nonetheless, a satisfactory approximation can be achieved in cases where p/α where n have values that are of the same order. In such a situation, however, the number of trials of the associated Bernoulli process is, a priori, not an integer. This problem can be circumvented, in turn, by approximating the binomial distribution with a Poisson distribution, if conditions permit.

 ISSN: 2102-6238

If the Rutherford distribution is approximated using the negative and positive binomial distributions, it is the similarities between the first moments that are used to establish the relationships among parameters. Situations in which the approximation has relevance can then be studied *a posteriori*. Rutherford also examined a means of approximating his distribution using the Gram-Charlier series, via similarities between the expressions for the first cumulants.

The Rutherford distribution is of particular interest to biologists and economists (Rutherford, 1954). Rutherford tested its performance with data on accidents involving women manufacturing highly explosive munitions, where it appeared to outperform the negative binomial and Neyman distributions. However, the improvement is actually rather small if one takes into account that the distribution has three parameters while the others have two.

6. Other mixture models

Finite Poisson mixture distribution. In a review, Feller provides an overview of Poisson mixture distributions. The first distribution mentioned is built using the step function F. It is, in fact, simply a Poisson distribution whose parameters are a finite number n of fixed values, $(\lambda_i)_{i\in[\![1,n]\!]}$, which follow a certain probability distribution $(p_i)_i$, where $\sum_{i=1}^{n} p_i = 1$. This yields the following expression:

$$\mathrm{P}(X=k) = \frac{1}{k!}\sum_{i=1}^{n} p_i \mathrm{e}^{-\lambda_i}\lambda_i^k,$$

with $k \in \mathbb{N}$.

Most of the time, the two-parameter form is used (where $p_1 = p$ and $p_2 = 1-p$) and therefore the mean and the variance can be found using $\mathbb{E}(X) = p\lambda_1 + (1-p)\lambda_2$ and $\mathrm{Var}(X) = p(1-p)(\lambda_1-\lambda_2)^2 + p\lambda_1 + (1-p)\lambda_2$, respectively. The distribution is clearly characterized by overdispersion, because the index of dispersion is described by $\mathrm{s}(X) = 1 + [p(1-p)(\lambda_1-\lambda_2)^2]/[p\lambda_1 + (1-p)\lambda_2]$.

In the more general case, when X follows this type of n-parameter distribution, the mean can be found using $\mathbb{E}(X) = \sum_{i=1}^{n} p_i\lambda_i$ and the variance can be found using $\mathrm{Var}(X) = \sum_{i=1}^{n} p_i\lambda_i(1+\lambda_i) - (\sum_{i=1}^{n} p_i\lambda_i)^2$ employing the notation system above. The index of dispersion is described by $\mathrm{s}(X) = [\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} p_i p_j(\lambda_i-\lambda_i)^2]/[\sum_{i=1}^{n} p_i\lambda_i]$. Consequently, the distribution is clearly overdispersed, regardless of parameter number.

This distribution was found to satisfactorily model the number of problems experienced on telephone lines (Palm, 1937), and was also applied on epidemiological data (Lundberg, 1940). Schlattmann and Böhning suggested to use it for disease mapping. Indeed, in Bayesian hierarchical models, the basic Poisson distribution should not be used as a first step because the associated *a priori* assumptions regarding case distribution are too strong (Schlattmann and Böhning, 1993).

Compound negative binomiale-gamma distribution / Generalized negative binomiale-logarithmic distribution. A negative binomial distribution can be viewed as a Poisson distribution compounded by a gamma distribution or a stopped-sum Poisson distribution of logarithmic

 ISSN: 2102-6238

variables. In the same way, one can define a distribution where the Poisson distribution is replaced by a negative binomial distribution. However, in contrast to the standard negative binomial, this distribution, called the negative binomial-gamma, can account for the overdispersion of primary cases.

The compound negative binomial-gamma distribution has three parameters and can be written as follows:

$$X \sim \mathscr{NB}(n,p) \text{ where } n \sim \Gamma(\alpha,\beta),$$

when $0 < p < 1$, $\alpha > 0$ and $\beta > 0$. It can also take the form of a stopped-sum distribution:

$$X = \sum_{i=1}^{Z} Y_i \text{ where } Z \sim \mathscr{NB}(\alpha,p') \text{ and } Y_i \sim \text{Log}(\pi)$$

when $p' = \beta\ln(1+p)$ and $\pi = \frac{p}{1+p}$.

The mean of this distribution is found using $\mathbb{E}(X) = \alpha\beta p = \alpha p'\pi/[(\pi-1).\ln(1-\pi)]$, and the variance is equal to $\text{Var}(X) = \alpha\beta p(\beta p + p + 1) = \alpha\delta\pi(\delta\pi + \ln(1-\pi))/[(1-\pi).\ln(1-\pi)]^2$. The spread can therefore be described by the expression $\text{s}(X) = \beta p + p + 1 = [\delta\pi + \ln(1-\pi)]/[(\pi-1).\ln(1-\pi)] > 1$, which means that only overdispersion can be modeled.

Generalized Poisson-negative binomial distribution / Compound negative binomial-Poisson distribution. The generalized Poisson-negative binomial distribution was developed by Subrahmaniam, who was inspired by Neyman's work (Neyman, 1939; Gurland, 1959). It was used to describe the number of larvae distributed across a gridded Petri dish (Subrahmaniam, 1966). This distribution provides an alternative to the Neyman and Thomas distributions in situations where the overdispersion associated with secondary cases can be modeled by replacing a Poisson distribution (potentially shifted) by a negative binomial distribution. This distribution has three parameters and can be expressed as follows:

$$X = \sum_{i=0}^{Z} Y_i \text{ where } Z \sim \mathscr{P}(\lambda) \text{ and } Y_i \sim \mathscr{NB}(n,p)$$

when $\lambda > 0$, $n > 0$ and $0 < p < 1$. Given Gurland's theorem, this distribution can also be viewed as a negative binomial-Poisson mixture distribution, which means it can be described using the following equation:

$$X \sim \mathscr{NB}(Ln,p) \text{ where } L \sim \mathscr{P}(\lambda)$$

with the same notation system as above.

For the Poisson-negative binomial distribution, the mean and variance can be determined using $\mathbb{E}(X) = \lambda n(1-p)/p$ and $\text{Var}(X) = \lambda n(n-np+1)(1-p)/p^2$, respectively. Therefore, the index of dispersion, $\text{s}(X) = (n-np+1)/p$, is clearly greater than 1.

The Pólya-Aeppli distribution is a special case of this distribution class, when $k = 1$ and with $\theta = \lambda/p$. This distribution can also be approximated by other characteristic distributions, such as

 ISSN: 2102-6238

- a Neyman type A distribution with two parameters λ and θ when $k \to \infty$ and $p \to 1$ where $\theta = \lim(k(1-p)/p)$;
- a negative binomial distribution with parameters n' and p when $\lambda \to \infty$ and $n \to 0$ where $n' = \lim(n\lambda)$;
- a Poisson distribution with a single parameter, λ', when $\lambda \to \infty$ and $p \to 1$ where $\lambda' = \lim(\lambda(1-p)/p)$.

Generalized Poisson-binomial distribution / Compound binomial-Poisson distribution. The generalized binomial-Poisson distribution is one of the stopped-sum distributions which can also be expressed as the mixture of a Poisson distribution and another classical distribution. This distribution is very close to the Neyman distribution; it should be used when the binomial distribution can not be replaced by the Poisson distribution, i.e. when the population size in some areas is small and/or when the frequence of the studied phenomenon is high. This three-parameter distribution can be given by

$$X = \sum_{i=0}^{Z} Y_i \text{ where } Z \sim \mathscr{P}(\lambda) \text{ and } Y_i \sim \mathscr{B}in(n,p),$$

when $\lambda > 0$, $n > 0$ and $0 < p < 1$. This distribution is also a compound binomial-Poisson distribution, thus

$$X \sim \mathscr{B}in(Ln,p) \text{ where } L \sim \mathscr{P}(\lambda)$$

using the following equation and the same notation system as above.

The mean and the variance of this distribution are found using $\mathbb{E}(X) = \lambda np$ and $\mathrm{Var}(X) = \lambda np.(np-p+1)$, respectively. Thus the index of dispersion, $\mathrm{s}(X) = np-p+1$, is greater than 1. The generalized binomial-Poisson distribution can be multimodal depending on the parameter values.

Compound Poisson-binomial distribution / Generalized binomial-Poisson distribution. The compound binomial-Poisson distribution is given by

$$X \sim \mathscr{P}(Y\lambda) \text{ when } Y \sim \mathscr{B}in(n,p),$$

when $n > 0$, $0 < p < 1$ and $\lambda > 0$. It can also take the form of a stopped-sum distribution with the same parameters:

$$X = \sum_{i=1}^{Z} Y_i \text{ where } Y_i \sim \mathscr{B}in(n,p) \text{ and } Z \sim \mathscr{P}(\lambda).$$

As the distribution defined right before, this distribution is very close to the Neyman distribution. It should be used when the the Poisson distribution does not provide a relevant approximation of the binomial distribution, i.e. when the number of studied areas is small and/or when the occurrence of primary cases is quite frequent.

The mean is found using $\mathbb{E}(X) = \lambda np$. One of its special cases is the zero-inflated Poisson (ZIP) distribution; indeed, if $n = 1$, this distribution is equivalent to the ZIP distribution whom

 ISSN: 2102-6238

parameters are λ and $1-p$.

Other compound Poisson distributions. Many other distributions have been generated by mixing a Poisson distribution with a second distribution (Karlis and Xekalaki, 2005): Beta (and one of its special cases known as the Yule or Yule-Simon), Lindeley, inverse normal (resulting in the Sichel distribution), inverse Gamma, Pareto, uniform, truncated normal, truncated Beta, truncated Gamma, log-Student and Lomax distributions. However, it is important to note that they are referred to neither as contagious distributions nor as stopped-sum distributions. Furthermore, most were developed to describe non-epidemiological data (e.g., economic data).

7. Other overdispersed distributions

Overdispersed distributions have traditionally taken the form of sum and mixture distributions that aim to model as closely as possible the systems they were developed to describe (e.g., plant abundance or larvae development). However, other distributions have been developed that can be adapted to situationd of overdispersed data, thus to situations of contagion. Here, we specifically focus on the generalized Poisson distribution, developed to integrate overdispersion into the Poisson distribution, and the ZI distribution, which has been associated with the Poisson, negative binomial, and generalized Poisson which accounts for censored data.

Generalized Poisson distribution. The Generalized Poisson distribution (represented by $\mathscr{GP}$) is an alternative to the simple Poisson distribution that can take into account overdispersion (Consul and Jain, 1973). If $X \sim \mathscr{GP}(\lambda, s)$ when $\lambda > 0$, then

$$\mathrm{P}(X=k) = \begin{cases} \lambda(\lambda(1-s)+k.s)^{k-1}\frac{(1-s)}{k!}\mathrm{e}^{-(\lambda(1-s)+k.s)} & \text{if } k \leq m \\ 0 & \text{if } k > m \end{cases}$$

as long as $\max(-1, -\lambda/(m-\lambda)) < s < 1$ where m is the largest integer for which $\lambda(1-s)+m.s > 0$ where $s < 0$.

Therefore, the expressions for the mean, variance, and index of dispersion are $\mathbb{E}(X) = \lambda$, $\mathrm{Var}(X) = \lambda/(1-s)^2$ and $\mathrm{s}(X) = 1/(1-s)^2$, respectively. Using this distribution, overdispersion can be modeled (when $s > 0$), as can underdispersion (when $s < 0$). The λ parameter is interpreted in the same way as for a classic Poisson distribution, and provides information about the spread of the distribution *via* sign and its absolute value. This distribution is a compound Poisson distribution (Joe and Zhu, 2005). However, it cannot be implemented because the mixing distribution is not specified. Both parameters can be estimated using the moments method: $\tilde{\lambda} = \bar{X}$ and $\tilde{s} = 1 - 1/\sqrt{s(X)}$.

Zero-inflated Poisson (ZIP) distribution. Overdispersion is often the product of the presence of excessive null values compared to what would be expected for a Poisson distribution, a common outcome when data are missing, censored, or underreported. In such situations, a ZIP model can be applied (Lambert, 1992; Gates, 2002; Winkelmann, 2003). It follows a Poisson distribution while also accounting for the inflated number of zeros. The ZIP distribution has two parameters,

 ISSN: 2102-6238

λ and π, and is given by the equation

$$\mathrm{P}(X=k)=\begin{cases}\pi+(1-\pi)\mathrm{e}^{-\lambda} & \text{if } k=0\\ (1-\pi)\mathrm{e}^{-\lambda}\frac{\lambda^y}{k!} & \text{if } k\in\mathbb{N}^*\end{cases}$$

where $\lambda>0$ is the parameter of the underlying Poisson distribution and $0\leq\pi\leq 1$ is the probability of extra zeros (Agarwal et al., 2002).

Consequently, the mean can be found with $\mathbb{E}(X)=\lambda(1-\pi)$ and the variance can be found with $\mathrm{Var}(X)=\lambda(1-p)(1+\lambda\pi)$, meaning the index of dispersion is given by $\mathrm{s}(X)=(1+\lambda\pi)$. This model can therefore only deal with overdispersion. Furthermore, the level of overdispersion increases as the proportion of extra zeros, π, and the mean for the underlying Poisson distribution, λ, increase.

The parameters of the ZIP distribution can be estimated using the moments method: $\tilde{\lambda}=\bar{X}+s(X)-1$ and $\tilde{\pi}=(s(X)-1)/(\bar{X}+s(X)-1)$. As both parameters are positive, the expression of $\tilde{\pi}$ must be truncated. More precisely $\tilde{\pi}'$ and $\tilde{\lambda}'$ are respectively defined as $\tilde{\pi}$ and $\tilde{\lambda}$ in case of overdispersion and as 0 and $\bar{X}$ otherwise. Both parameters can also be computed using the Maximum Likelihood Estimation: $\hat{\lambda}$ is the root of $\hat{\lambda}\mathrm{e}^{\hat{\lambda}}/(\mathrm{e}^{\hat{\lambda}}-1)=\bar{X}/(1-f_0)$, where f_0 is the empirical proportion of null values, and $\hat{\pi}$ is defined by $\hat{\pi}=1-\bar{X}/\hat{\lambda}$.

In practice, Bayesian models are better suited to dealing with latent variables following a Bernoulli distribution with parameter p. Shifting approaches does not change the model. However, the Bayesian perspective can facilitate the model's implementation, because the ZIP distribution can be viewed as a compound Poisson-Bernoulli distribution:

$$X\sim\mathscr{P}(\delta\lambda)\text{ where }\delta\sim\mathscr{B}(\pi).$$

Although the approach is somewhat more artificial, it is also possible to view it as a sum distribution:

$$X=\sum_{i=1}^{Z}Y_i\text{ where }Z\sim\mathscr{B}(\pi)\text{ and }Y_i\sim\mathscr{P}(\lambda).$$

Zero-Inflated Binomial (ZIB) distribution. The ZIB distribution is defined similarly to the ZIP. It has three parameters, two of which are the parameters of a binomial distribution (n and p), and a third which is the proportion of additional zeros (π). It is defined as

$$\mathrm{P}(X=k)=\begin{cases}\pi+(1-\pi))(1-p)^n & \text{if } k=0\\ (1-\pi)\binom{n}{k}p^k(1-p)^{n-k} & \text{if } k\in\mathbb{N}^*\end{cases}$$

for all integers k.

 ISSN: 2102-6238

If X follows this distribution, the mean and variance are found using $\mathbb{E}(X) = (1-\pi)np$ and $\mathrm{Var}(X) = (1-\pi)np(1-p+\pi np)$, respectively. The index of dispersion can therefore be found using $\mathrm{s}(X) = 1-p+\pi np$. This distribution is overdispersed when $p \neq 0$ and $\pi \geq 1/n$.

Zero-inflated negative binomial (ZINB) distribution. The zero-inflated distribution can be used in tandem with a negative binomial distribution (Fahrmeir et al., 2006) when overdispersion seems to be linked to an excess of zeros and when, additionally, the variance associated with the non-zero values is large. Like the ZIP distribution, it can be expressed as a mixture or a sum distribution.

Zero-inflated Generalized Poisson distribution (ZIGP). Another possibility for dealing with an excess of zeros is to use a model based on a Generalized Poisson distribution. The distribution obtained has three parameters and is given by

$$\mathrm{P}(X=k) = \begin{cases} p+(1-p)(1-s)\frac{\lambda}{C}\mathrm{e}^{-C} & \text{if } k=0 \\ (1-p)\lambda C^{k-1}\frac{(1-s)}{k!}\mathrm{e}^{-C} & \text{if } k \in [\![1,m]\!] \\ 0 & \text{if } k>m. \end{cases}$$

where $k \in \mathbb{N}$, $C = \lambda(1-s)+k.s$ and m is defined as for the Generalized Poisson distribution.

The mean and variance are found using $\mathbb{E}(X) = \lambda(1-p)$ and $\mathrm{Var}(X) = \lambda(1-p).[p\lambda + 1/(1-s)^2]$, respectively. Therefore, the index of dispersion can be determined with $\mathrm{s}(X) = \lambda p + 1/(1-s)^2$. It is possible to parse out this expression to distinguish the contribution made by s $(1/(1-s)^2)$, which is the parameter describing the spread of the generalized Poisson, and the contribution made by p, which is the proportion of additional zeros.

The ZIGP distribution can be generalized to deal with excessively large proportions of a value k, where k is an integer (Famoye and Singh, 2003). In most cases where this generalization is used, $k=1$ or $k=0$.

8. Numerical Application

We test the relevance of the previously seen distributions on real count data $X = (X_i)_i$. We consider the number of bovine tuberculosis cases X_i that occurred each year between 2001 and 2010 in France, divided into 448 areas (Table 1). These data have been provided by the DGAL (the French Directorate for Food) (ANSES and DGAL, 2011). Bovine tuberculosis is a disease caused by a bacterium which can affect the cattle (OIE, 2012) and also contaminate humans (Grange et al., 1996). Bovine tuberculosis cases show a high level of overdispersion as $s(X) = 4.94$. Overdispersion may be due to spatiotemporal dependencies, however this value is particularly high and must mainly be due to the outcome of grouped cases in some areas during some periods (local overdispersion). We also notice that null values are over-represented, extreme values (> 5) occur more frequently, and there is a lack of intermediate values, in particular 1 (Table 1).

We want to test the adequacy of several distributions with the repartition of the bovine tuberculosis data: 1) the Poisson distribution (which is the standard count distribution), 2) the

 ISSN: 2102-6238

TABLE 1. *Number of cases and their occurrence.*

Number of cases	1	2	3	4	5	6	7	8	9	11	16	18	20	22
Number of occurrences	4134	62	25	11	4	5	5	4	3	2	1	1	1	1

negative binomial distribution, 3) the Neyman distribution, 4) the Thomas distribution, 5) the ZIP distribution, 6) the Generalized Poisson distribution and 7) the Pólya-Aeppli distribution. We consider the empirical mean and variance, and we estimate the parameters which characterize the previously mentionned distributions using the moments method (Table 2). Some of these parameters make sense. For the Neyman distribution, the probability of the occurrence of a primary case is $\tilde{\lambda} = 0.040$ and the average number of secondary cases, for each cluster, is $\tilde{\theta} = 3.943$. The results are very similar for the Thomas distribution: $\tilde{\lambda} = 0.030$ and $\tilde{\theta} = 4.182$. For the Pólya-Aeppli distribution, the outcome of a primary case occurs with a probability $\tilde{\lambda} = 0.053$. The ZIP distribution is equivalent to a Poisson distribution whom parameter is $\tilde{\lambda} = 4.100$ and with an extra-proportion of null values $\tilde{\pi} = 0.962$.

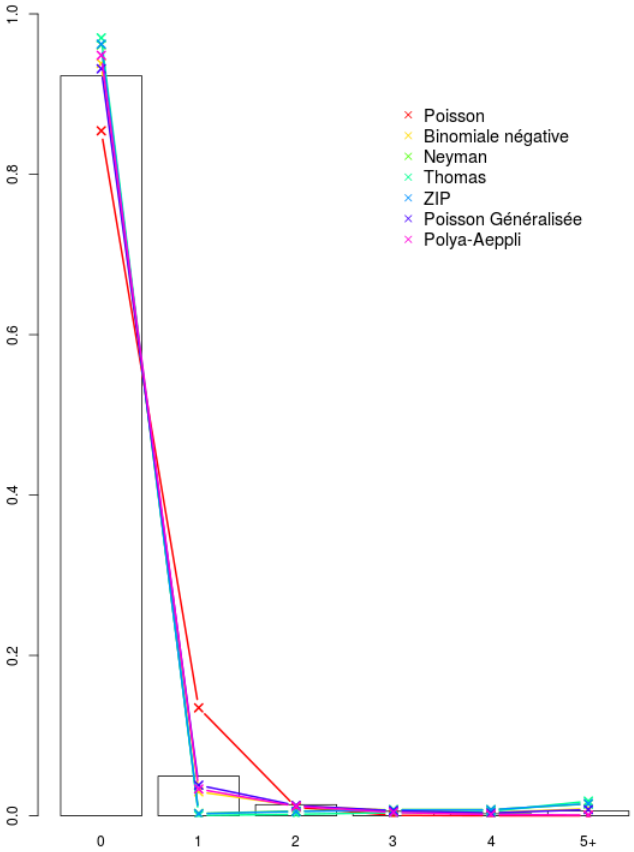

FIGURE 1. *Distribution of the cases and of the theoric values.*

 ISSN: 2102-6238

TABLE 2. *Estimated parameters, probability mass distribution and indicators for each distribution.*

Value	Data Frequency	Poisson distribution	Negative Binomial distribution	Neyman distribution	Thomas distribution	Zero-Inflated Poisson distribution	Generalized Poisson distribution	Polya-Aeppli distribution
Parameter(s)		$\lambda = 0.158$	$n = 0.040$ $p = 0.202$	$\lambda = 0.040$ $\theta = 3.943$	$\lambda = 0.030$ $\theta = 4.182$	$\pi = 0.962$ $\lambda = 4.100$	$\lambda = 0.158$ $s = 0.550$	$\lambda = 0.053$ $p = 0.663$
0	0.9228	0.8542	0.9381	0.9614	0.9700	0.9622	0.9316	0.9483
1	0.0493	0.1346	0.0299	0.0029	0.0005	0.0026	0.0381	0.0334
2	0.0138	0.0106	0.0124	0.0058	0.0019	0.0054	0.0129	0.0118
3	0.0056	0.0005	0.0067	0.0076	0.0039	0.0073	0.0063	0.0042
4	0.0025	2.10^{-5}	0.0041	0.0075	0.0055	0.0075	0.0036	0.0015
5	0.0009	7.10^{-7}	0.0026	0.0060	0.0058	0.0061	0.0022	0.0005
6	0.0011	2.10^{-8}	0.0018	0.0040	0.0048	0.0042	0.0015	0.0004
7	0.0011	4.10^{-10}	0.0012	0.0023	0.0034	0.0025	0.0010	7.10^{-5}
8	0.0009	8.10^{-12}	0.0009	0.0012	0.0020	0.0013	0.0007	2.10^{-5}
9	0.0007	1.10^{-13}	0.0006	0.0006	0.0011	0.0006	0.0005	8.10^{-6}
≥ 10	0.0013	2.10^{-15}	0.0017	0.0007	0.0011	0.0004	0.0017	4.10^{-6}
Mean Square Error		0.1662	0.0132	0.0993	0.1106	0.1040	0.0059	0.0117

For each distribution, using the estimated parameters, we compute the probability mass distributions (Table 2, Figure 1). We notice that the Poisson distribution fails to model the bovine tuberculosis data: it strongly under-estimates the null values and the extrem values, and it generates three times too much clusters which contain an only case. In fact this phenomenon (isolated case) is not so common for infectious pathologies. At the opposite, the Neyman, Thomas and ZIP distributions over-estimate the ratio of null values and under-estimate the proportion of intermediate values. They also slightly over-estimate the outcome of extreme values. The negative binomial, Generalized Poisson and Pólya-Aeppli distributions provide a relevant approximation of the real distribution of bovine tuberculosis cases, even if the Pólya-Aeppli distribution under-estimates the occurrence of extreme values.

To compare each distribution with the empirical repartition of the real data, we compute the Mean Square Errors (MSE, Figure 2). The top-ranked distributions are, in order, the Generalized Poisson, the negative binomial and the Pólya-Aeppli distributions. As expected the highest MSE value is obtained for the Poisson distribution. The three other distributions (Neyman, Thomas and ZIP distributions) also provide high MSE, thus they do not appear as very relevant to model the bovine tuberculosis data.

9. Conclusion

At the beginning of the twentieth century, very few discrete probability distributions had been described. The Poisson distribution was very common, as were the binomial and negative binomial distributions. Between 1920 and 1960, several "contagious" distributions were developed by researchers to describe phenomena in their fields of interest (e.g., bacteriology, ecology, entomology, microbiology) for which the classic distributions provided a poor fit because of the overdispersion

of the data. Ultimately this work led to a variety of distributions, including the Neyman, Thomas, and Pólya-Aeppli distributions and cast a different light on the negative binomial.

These distributions can generally be expressed as the sum of random variables for which the limit of the sum index is a random variable. As a consequence of this formulation, such distributions become "contagious", that is to say they describe a process wherein the random occurrence of a primary event engenders proximate secondary events. Contagious distributions have several practical advantages. Although they are rarely included in statistical software programs, they can nonetheless be used by exploiting a mixture of more common distributions. By their nature, they are overdispersed. They can be bimodal if, in a situation of contagion, the number of secondary cases associated with a given primary case is sufficiently stable. Finally, they are very flexible and able to cover a large range of discrete distributions.

Six of the overdispersed distributions presented in this paper have been tested on bovine tuberculosis data. All of them over-performed the Poisson distribution which fails to model overdispersed data. The Generalized Poisson, negative binomial and Pólya-Aeppli distributions in particular provided relevant estimations of the distribution of real values. However, even if they share common characteristics, the overdispersed distributions have different specificities and may suit different contexts. For instance, the ZIP distribution is dedicated to missing, under-declared or censored data. The Generalized Poisson distribution seems well-adapted to overdispersed data which are quite similar to Poisson distributions. Contrary to the other distributions, Neyman and Thomas distributions are multimodal thus they seem well-adapted to count data which contain more numerous extrem values than intermediate values. They would be relevant for instance in the context of very contagious phenomena, i.e. for which the outcome of a case strongly increases the probability of other occurrences. The negative binomial and the Pólya-Aeppli distributions appear as closer to the Generalized Poisson distribution. They can respectively be expressed as sums of logarithmic and geometric distributions, which are strictly decreasing, contrary to the Poisson distribution. Thus the negative binomial and the Pólya-Aeppli distributions are likely more adapted to data with a very high proportion of null values and less extreme values than the Neyman and Thomas distributions.

Despite their potential utility, following their initial heyday, these overdispersed distributions were subsequently little used or studied. Even the field of epidemiology failed to adopt these distributions; not even the idea of accounting for contagion in count distributions took root. Yet, the negative binomial is increasingly employed because of the need to deal with overdispersed data, the desire to make commonly exploited count distributions more flexible, and the growing diversity of calculation methods. Different types of Zero-Inflated distributions have become popular because they make it possible to account for missing or censored values or, more generally speaking, overdispersion. Indeed, contagious distributions could be useful in such areas as the detection of epidemics, the simulation of issues related to human health, risk mapping, and the building of regression models. Broadening the current suite of count distributions could vastly expand modeling possibilities and the ability to fit distributions to data with very diverse profiles. Contagious distributions have already been used to model such data and, more specifically, phenomena where the occurrence of one event leads to the occurrence of additional events. They can

 ISSN: 2102-6238

be used in any type of situation where contagion processes are thought to be playing a role; more generally, they can deal with aggregated or clustered data, which are common in such fields as ecology, epidemiology and microbiology.

References

Académie_Française (1932-1935). *Dictionnaire de l'Académie française*. huitième edition.

Adelstein, A. (1952). Accident proneness: a criticism of the concept based upon an analysis of shunters' accidents. *Journal of the Royal Statistical Society. Series A*, 115(3):354–410.

Aeppli, A. (1924). *Zur Theorie verketteter Wahrscheinlichkeiten*. PhD thesis, Zurich.

Agarwal, D. K., Gelfand, A., and Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, 9:341–355.

Anscombe, F. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, 37(3):358–382.

ANSES and DGAL (2011). Bulletin Epidemiologique - Sante animale-alimentation. Technical report.

Archibald, E. (1948). Plant populations I: New application of Neyman's contagious distribution. *Annals of botany*, 12(221).

Beall, G. and Rescia, R. R. (1953). A generalization of Neyman's contagious distributions. *Biometrics*, 9(3):354–386.

Bliss, C. and Fisher, R. (1953). Fitting the negative binomial distribution to biological data / Note on the efficient fitting of the negative binomial. *Biometrics*, 9(2):176–200.

Brent, R. (2002). *Algorithms for minimization without derivatives*. Dover Publications.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43:671–681.

Consul, P. and Jain, G. (1973). A generalization of the poisson distribution. *Technometrics*, 15(4):791–799.

Eggenberger, F. (1924). Die wahrscheinlichkeitsansteckung. (371).

Eggenberger, F. and Polya, G. (1923). Über die Statistik verketteter Vorgange. *Zeitschrift fur Angewandte Mathmatik und Mechanik*, 1:279–289.

Evans, D. (1953). Experimental evidence concerning contagious distributions in ecology. *Biometrika*, 40(1/2):186–211.

Fahrmeir, L., Osuna, L., and Echavarría, L. O. (2006). Structured additive regression for overdispersed and zero-inflated count data. *Applied Stochastic Models in Business and Industry*, 22:351–369.

Famoye, F. and Singh, K. P. (2003). On inflated Generalized Poisson regression models. *Advances and Applications in Statistics*, 3:145–158.

Feller, W. (1943). On a general class of "contagious" distributions. In *Annals of Mathematical Statistics*, pages 389–400.

Feller, W. (1957). *An introduction to probability theory and its applications*. New-York, wiley edition.

Fisher, R., Corbet, A. S., and Williams, C. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *British Ecological Society*, 12(1):42–58.

Gates, S. (2002). Econometric analysis of count data (book). *Journal Of Peace Research*, 39(1):132.

Grange, J. M., Yates, M. D., and de Kantor, I. N. (1996). Guidelines for speciation within the Mycobacterium tuberculosis complex. Technical report, World Health Organization.

Greenwood, M. and Yule, G. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of. *Journal of the Royal Statistical Society*, 83(2):255–279.

Grogger, J. and Carson, R. (1991). Models for truncated counts. *Journal of Applied Econometrics*.

Gurland, J. (1957). Some interrelations among compound and generalized distributions. *Biometrika*, 44(1–2):265–268.

Gurland, J. (1958). A generalized class of contagious distributions. *Biometrics*, 14(2):229–249.

Gurland, J. (1959). Some applications of the negative binomial and other contagious distributions. *American journal of public health and the nation's health*, 49(10):1388–99.

Gurland, J. (1963). Some families of compound and generalized distributions. pages 1–29.

Hsu, P. (1938). Contribution to the theory of Students's't-test as applied to the problem of two samples. *Statistical Research Memoirs*, 2:1–24.

 ISSN: 2102-6238

Joe, H. and Zhu, R. (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. *Biometrical journal. Biometrische Zeitschrift*, 47:219–229.
Jones, P., Mollison, J., and Quenouille, M. (1948). A technique for the quantitative estimation of soil micro-organisms. *J. General Microbiol.*, 2:54–69.
Karlis, D. and Xekalaki, E. (2005). Mixed Poisson distributions. *International Statistical Review*, 73(1):35–58.
Kendall, D. G. (1949). Stochastic processes and population growth. *J. R. Statist. Soc. Ser. B*, 11:230–264.
Kennedy, P. (1994). *A Guide to Econometrics*. MIT Press.
Kitagawa, T. and Huruya, S. (1941). The application of the limit theorems of the contagious stochastic processes to contagious diseases. *Mem. Faculty of Sc. KyushuImperial University A*, 1:195–207.
Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.
Le Cam, L. (1960). An approximation theorem for the Poisson binomial distribution. *Pacific J. Math*, 10(4):1181–1197.
Lundberg, O. (1940). *On random processes and their application to sickness and accident statistics*. PhD thesis, University of Stockholm.
Massé, J.-C. and Theodorescu, R. (2005). Neyman type A distribution revisited. *Statistica Neerlandica*, 59(2):206–213.
McKendrick, A. (1914). Studies on the theory of continuous probabilities, with special reference to its bearing on natural phenomena of a progressive nature. *Proceeding of the London Mathematical Society*, 13:401–416.
Neyman, J. (1939). On a new class of "contagious" distributions, applicable in entomology and bacteriology. *The Annals of Mathematical Statistics*, 10(1):35–57.
Nuel, G. (2008). Cumulative distribution function of a geometric poisson distribution. *Journal of Statistical Computation and Simulation*, 78(3):385–394.
OIE (2012). Tuberculose bovine. *Fiches d'information generale sur les maladies*, pages 1–6.
Özel, G. and Ceyhan, I. (2010). The probability function of a geometric poisson distribution. *Journal of Statistical Computation and Simulation*, 80(5):479–487.
Palm, C. (1937). Inhomogenous telephone traffic in full-availability groups. *Ericsson Technics*, 1:1–36.
Park, K. S. and Jun, C. H. (2000). Use of contagious distributions in the semiconductor yield models considering cluster effect. In *Communications in Statistics - Theory and Methods*, pages 1–17.
Park, K. S. and Jun, C. H. (2002). Semiconductor yield models using contagious distributions and their limiting forms. *Computers and Industrial Engineering*, 42(2–4):115–125.
Pearson, K. (1895). Contributions to the mathematical theory of evolution - II. Skew variation in homogeneous material. *Philosophical transactions of the royal society of London (A)*, 186:343–414.
Polya, G. (1930). Sur quelques points de la théorie des probabilités. *Annales de l'Institut Henri Poincaré*, 1(2):117–161.
Quenouille, M. (1949). A relation between the logarithmic, Poisson, and negative binomial series. *Biometrics*, 5(2):162–164.
Rémond de Montmort, P. (1713). *Essai d'analyse sur les jeux de hasard*. Paris.
Rosenblatt, A. (1940). *Sur le concept de contagion de M. G. Pólya dans le calcul des probabilités. Divers schèmas. Application à la peste bubonique au Pérou*. Academia de Ciencias Exactas, Fisicas y Naturales de Lima.
Rutherford, R. (1954). On a contagious distribution. Technical report, University of Sydney, Sydney.
Satterthwaite, F. (1942). Generalized Poisson distribution. *Annals of Mathematical Statistics*, 13:410–417.
Schlattmann, P. and Böhning, D. (1993). Mixture models and disease mapping. *Statistics in Medicine*, 12:1943–1950.
Sherbrooke, C. C. (1968). Discrete compound poisson processes and tables of the geometric poisson distribution. *Naval Research Logistics*, 15:189–203.
Skellam, J. (1952). Studies in statistical ecology 1, spacial patterns. *Biometrika*, 39:346–362.
Stevenson, A. and Lindberg, C. A., editors (2015). *New Oxford American Dictionary*. Oxford University Press.
Subrahmaniam, K. (1966). On a general class of contagious distributions: The pascal-poisson distribution. *Trabajos de estadistica y de investigacion operativa*, 17(2–3):109–128.
Teich, M. C. (1981). Role of the doubly stochastic Neyman type-A and Thomas counting distributions in photon detection. *Applied optics*, 20(14):2457–67.
Thomas, M. (1949). A generalization of Poisson's binomial limit for use in ecology. *Biometrika*, 36(1):18–25.
von Luders, R. (1934). Die Statistik der seltenen Ereignisse. *Biometrika*, 26(1–2):108–128.
Winkelmann, R. (2003). *Econometric Analysis of Count Data*. Springer.
Woodbury, M. A. (1949). On a probability distribution. In *The annals of mathematical statistics*, pages 311–313. University of Michigan.

 ISSN: 2102-6238