Characterization of the Pólya - Aeppli process

Stefanka Chukova
School of Mathematics, Statistics and Operations Research
Victoria University of Wellington, NZ,
e-mail: schukova@msor.vuw.ac.nz

Leda D. Minkova
Faculty of Mathematics and Informatics,
Sofia University "St. Kl.Ohridski",
e-mail: leda@fmi.uni-sofia.bg

Abstract

In this paper we study the Pólya - Aeppli process (PAP). We define PAP from three different points of view: as a compound Poisson process, as a delayed renewal process and as a pure birth process. We show that these definitions are equivalent. Also, using these definitions we identify several interesting characterizations of PAP.

Key words: Pólya - Aeppli distribution, compound distribution, delayed renewal process, pure birth process.

AMS subject classifications: 60G51; 60G55; 60G10.

1 Introduction

The standard, widely used model for count data is the Poisson process with intensity $\lambda > 0$ and probability mass function (PMF) given by

$$P(N_1(t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad k = 0, 1, \ldots. \tag{1}$$

One of the most important properties of the Poisson process is its equidispersion, i.e., the Poisson variance and mean are equal. Then the corresponding Fisher index, which is defined as the ratio of the variance to the mean, is equal to one. In many practical applications, the equidispersion property of the Poisson process is not observed in the count data at hand, so it motivates the search for more flexible models for this type of data. The most commonly used generalization of the Poisson process is the compound Poisson process. It is the basic counting process in risk models. The compound Poisson model is very useful for modeling batch arrivals in queueing systems, as well as cluster data. In Minkova [7] a compound Poisson process with geometric compounding distribution is defined and some of its possible applications in risk models, including a discussion on related ruin probability are given. The corresponding counting process is called a Pólya - Aeppli process. Here we use the acronym PAP for the Pólya - Aeppli process. For details on the Pólya-Aeppli distribution, see Johnson et al. [4], and also Minkova [5, 6].

The paper is organised as follows. In Section 2 we consider three equivalent definitions of PAP. The equivalence of the definitions is discussed in Section 3. Section 4 deals with some properties of PAP. Useful characterizations of PAP are derived in Section 5. Section 6 concludes this study.

2 Definitions of PAP

In what follows we view PAP from different viewpoints and offer several possible definitions of this process. In Section 3 we show that these definitions are equivalent.

2.1 PAP as a compound Poisson process

Now, consider $\{N(t), t \geq 0\}$ as a counting process, i.e., $N(t)$ is equal to the number of arrivals in the interval $(0, t]$. Let $X_1, X_2, \ldots$ be i.i.d. integer valued random variables. Suppose that $N(t)$ is given by the sum

$$N(t) = X_1 + \ldots + X_{N_1(t)}, \tag{2}$$

where $N_1(t)$ is a homogeneous Poisson process with intensity $\lambda > 0$, independent of $\{X_i\}_1^\infty$. Then, the process defined in (2) is called a compound Poisson process.

Suppose that $X_1, X_2, \ldots$ are geometrically distributed with parameter $1-\rho$, $\rho \in [0, 1)$ and support $\{1, 2, \ldots\}$, which is denoted by $X \sim Ge_1(1-\rho)$. Then the probability mass function (PMF) and the probability generating function (PGF) of X are given as follows:

$$P(X = i) = \rho^{i-1}(1-\rho), \quad i = 1, 2, \ldots \tag{3}$$

and

$$\psi_1(s) = Es^X = \frac{(1-\rho)s}{1-\rho s}. \tag{4}$$

Then, using (4), the PGF of $N(t)$ can be expressed as

$$\psi_{N(t)}(s) = e^{-\lambda t(1-\psi_1(s))}. \tag{5}$$

Based on the above construction, it is easy to see that $N(t)$, which is the number of arrivals in the interval $(0, t]$, is Pólya-Aeppli distributed with parameters λt and $\rho \in [0, 1)$, i.e., $N(t) \sim PA(\lambda t, \rho)$, i.e., the probability mass distribution function of $N(t)$ is given by:

$$P(N(t) = m) = \begin{cases} e^{-\lambda t}, & m = 0 \\ e^{-\lambda t} \sum_{i=1}^{m} \binom{m-1}{i-1} \frac{[\lambda t(1-\rho)]^i}{i!} \rho^{m-i}, & m = 1, 2, \dots. \end{cases} \tag{6}$$

Now, we are ready to give the first definition of the Pólya - Aeppli process.

Definition 2.1 *(Minkova [7]) A counting process $\{N(t), t \geq 0\}$ is said to be a Pólya - Aeppli process with parameters λ and ρ if*

(i) it starts at zero, $N(0) = 0$;

(ii) $N(t)$ is a process with independent increments;

(iii) for each $t > 0$, the number of arrivals $N(t)$ in any interval of length t is Pólya - Aeppli distributed with parameters λt and ρ.

To express the fact that a counting process $\{N(t), t \geq 0\}$ is a Pólya - Aeppli process with parameters λ and ρ, we use the notation $N(t) \sim PAP(\lambda, \rho)$. If $\rho = 0$, then $PAP(\lambda, 0)$ simplifies to the homogeneous Poisson process with intensity λ. The mean and variance of $PAP(\lambda, \rho)$ are given by

$$E(N(t)) = \frac{\lambda t}{1-\rho} \quad \text{and} \quad Var(N(t)) = \frac{\lambda t(1+\rho)}{(1-\rho)^2},$$

and related Fisher index is equal to

$$FI(N(t)) = \frac{Var(N(t))}{E(N(t))} = 1 + \frac{2\rho}{1-\rho} > 1.$$

Therefore, not only Poisson process is a particular case of Pólya-Aeppli process, but for $\rho \neq 0$ the Pólya-Aeppli process is over - dispersed, which provides a greater flexibility in modeling count data than the standard Poisson process.

2.2 PAP as a delayed renewal process

Our second PAP definition views the process as a delayed renewal process. Consider $PAP(\lambda, \rho)$ and denote by T_1 the time to the first arrival and by $T_2, T_3, \dots$ the consecutive interarrival times. Suppose that $T_1, T_2, \dots$ are mutually independent random variables. Moreover, assume that T_1 is exponentially distributed with parameter λ and $T_2, T_3, \dots$ are exponentially distributed with mass ρ at zero, which is denoted by $T_2 \sim \exp(\lambda, \rho)$, with corresponding distribution function

$$F_{T_2}(t) = 1 - (1-\rho)e^{-\lambda t}, \quad t \geq 0. \tag{7}$$

It is easy to verify that

$$F_{T_1}(t) = \frac{1}{ET_2} \int_0^t [1 - F_{T_2}(x)] dx. \tag{8}$$

Then, our second definition of $PAP(\lambda, \rho)$ is as follows:

Definition 2.2 *The delayed renewal process $\{T_1, T_2, \dots\}$ is called a Pólya - Aeppli process with parameters λ and ρ.*

2.3 PAP as a pure birth process

Next, noticing that the construction of $PAP(\lambda, \rho)$ allows the process to be defined as a pure birth process, our third PAP definition is as follows:

Definition 2.3 *A counting process $\{N(t), t \geq 0\}$ is said to be a Pólya - Aeppli process with parameters λ and ρ if*

(a) $N(0) = 0$;

(b) $N(t)$ is a process with independent and stationary increments;

(c) the state transition probabilities are defined as follows:

$$(9) \quad P(N(t+h) = n \mid N(t) = m) = \begin{cases} 1 - \lambda th + o(h), & n = m, \\ (1-\rho)\rho^{i-1}\lambda th + o(h), & n = m + i, i = 1, 2, \ldots, \end{cases}$$

for every $m = 0, 1, \ldots,$ where $o(h) \to 0$ as $h \to 0$.

3 Equivalence of PAP definitions

In what follows we show that the three PAP definitions given in Section 2 are equivalent.

Proposition 3.1 *The definition 2.1 and definition 2.2 of the Pólya - Aeppli process are equivalent.*

Proof. Let $\tau_n = T_1 + \ldots + T_n,\ n = 1, 2, \ldots$ be the waiting time until the n^{th} arrival. The well known relation

$$P(N(t) = n) = P(\tau_n \leq t) - P(\tau_{n+1} \leq t)$$

shows that the condition $N(t) \sim PA(\lambda t, \rho)$ and the assumptions related to $T_1, T_2, \ldots$ in Section 2.2 are equivalent, see Minkova [7]. Therefore, the Pólya - Aeppli process is a delayed renewal process. Due to condition (8), it is a stationary renewal process, see Serfozo [8], p. 145.

□

Proposition 3.2 *The definition 2.1 and definition 2.3 of the Pólya - Aeppli process are equivalent.*

Proof. The interpretation of condition (9) is that geometrically distributed clusters (or batches) arrive randomly with arrival rate λ. Let $P_m(t) = P(N(t) = m),\ m = 0, 1, 2, \ldots$. Then (9) yields the following Kolmogorov forward equations:

$$(10) \quad \begin{aligned} P_0'(t) &= -\lambda P_0(t), \\ P_m'(t) &= -\lambda P_m(t) + (1-\rho)\lambda \textstyle\sum_{j=1}^{m} \rho^{j-1} P_{m-j}(t), \quad m = 1, 2, \ldots, \end{aligned}$$

with initial conditions

$$(11) \quad P_0(0) = 1 \quad \text{and} \quad P_m(0) = 0, \quad m = 1, 2, \ldots.$$

Next, we show that the solution of equations (10), with initial condition (11), is given by (6).

Let us consider the PGF

$$h(s,t) = \sum_{m=0}^{\infty} s^m P_m(t)$$

of the process $N(t)$. Multiplying the mth equation of (9) by s^m and summing for all $m = 0, 1, 2, \ldots,$ we get the following differential equation

$$\frac{\partial h(s,t)}{\partial t} = -\lambda[1-\psi_1(s)]h(s,t). \tag{12}$$

The solution of (12), with the initial condition $P_0(0) = 1$, is

$$h(s,t) = e^{-\lambda t[1-\psi_1(s)]},$$

which is the PGF of the $PAP(\lambda, \rho)$, given by (5), which leads to (6).

□

Therefore all three PAP definitions introduced in Section 2 are equivalent and represent PAP from different viewpoints, using different mathematical apparatus with specifics that could be advantageous in practical applications related to count data.

4 Properties

Next, we study some interesting properties of PAP, with arrival times $0 = \tau_0 < \tau_1 \leq \ldots$.

Proposition 4.1 (The waiting time distribution) *The distribution function of the waiting time τ_n is given by*

$$F_{\tau_n}(t) = 1 - e^{-\lambda t} \sum_{i=0}^{n-1} \binom{n-1}{i} (1-\rho)^i \rho^{n-1-i} \sum_{j=0}^{i} \frac{(\lambda t)^j}{j!}, \quad n = 1, 2, \ldots. \tag{13}$$

Proof. To prove the statement we will use mathematical induction.

- For $n = 1$ we get the the distribution function of $\tau_1 = T_1$

  $$F_{\tau_1}(t) = 1 - P(N(t) = 0).$$

  Applying (6) for $m = 0$, it follows that

  $$F_{\tau_1}(t) = 1 - e^{-\lambda t} \quad t > 0.$$

  According to the basic properties of the counting processes we have the following relation

  $$F_{\tau_n}(t) = F_{\tau_{n-1}}(t) - P(N(t) = n-1), \quad n = 2, 3, \ldots. \tag{14}$$

- For $n = 2$, the distribution function of $\tau_2 = T_1 + T_2$ is

  $$F_{\tau_2}(t) = F_{\tau_1}(t) - P(N(t) = 1) = 1 - (1 + (1-\rho)\lambda t)e^{-\lambda t}.$$

- Suppose now that for $n \geq 2$, the distribution function of the waiting time is given by

$$F_{\tau_{n-1}}(t) = 1 - e^{-\lambda t} \sum_{i=0}^{n-2} \binom{n-2}{i} (1-\rho)^i \rho^{n-2-i} \sum_{j=0}^{i} \frac{(\lambda t)^j}{j!}. \tag{15}$$

Then, applying (6) for $m = n-1$ and after substituting (15) into (14) and simplifying, we obtain (13).

□

Remark 4.1 *The distribution function (13) is a generalization of the Erlang distribution. In the case $\rho = 0$ it simplifies to the distribution function of Erlang(λ, n).*

Proposition 4.2 (Martingale property) *For $N(t) \sim PA(\lambda t, \rho)$, the process $M(t) = N(t) - \frac{\lambda}{1-\rho} t$ is a martingale.*

Proof. Since $\frac{\lambda}{1-\rho}t$ is non - random, $E\left(N(t) - \frac{\lambda}{1-\rho}t\right) = 0$ and $M(t)$ has independent increments. Therefore, for $s \leq t$ and $\mathcal{F}_t = \sigma\{N(s), s \leq t\}$, we have

$$\begin{aligned} E\left[N(t) - \frac{\lambda}{1-\rho} t | \mathcal{F}_s\right] &= E\left[N(t) - N(s) - \frac{\lambda}{1-\rho}(t-s) | \mathcal{F}_s\right] + N(s) - \frac{\lambda}{1-\rho} s \\ &= N(s) - \frac{\lambda}{1-\rho} s, \end{aligned}$$

as required.

□

5 PAP Characterizations

Our next goal is to derive the characterization properties of the Pólya - Aeppli process, but before doing so, we need to study the characterization properties of the exponential distribution with mass at zero.

5.1 Characterization I

In Galambos and Kotz [3], p.12, the authors show that the following conditions

$$[1 - F(x)]' = -\lambda [1 - F(x)] \text{ and } \int_z^\infty [1 - F(x)]dx = \frac{1}{\lambda}[1 - F(z)], \ z \geq 0 \tag{16}$$

with

$$F(0+) = 0 \tag{17}$$

provide a characterization of the exponential distribution with parameter λ. The condition (17) is significant in their proof. Furthermore, if the condition (17) is omitted, they show, that the class of functions

$$F(x) = 1 - ce^{-\lambda x}, \quad c \in (0, 1], \quad x \geq 0 \tag{18}$$

satisfies the conditions (16), where the choice of the constant c characterizes the distribution. For example, $c = 1$ provides a characterization for the exponential $exp(\lambda)$ (condition (17) is satisfied). For $c \in (0, 1)$, it provides a characterization for the exponential distribution with mass $(1 - c)$ at zero. The corresponding distribution in (18) with $c = 1 - \rho < 1$, is $exp(\lambda, \rho)$, given in (7).

Let T_2 be a random variable with distribution function $F_{T_2}(x)$ and finite mean $ET_2 < \infty$. Define a random variable T_1 with distribution $F_{T_1}(z) = \frac{1}{ET_2}\int_0^z [1 - F_{T_2}(x)]dx$. Then the following characterization result holds.

Lemma 5.1 *The random variable T_1 is exponentially distributed with parameter λ if and only if $T_2 \sim \exp(\lambda)$ or $T_2 \sim \exp(\lambda, \rho)$.*

Proof.

- From the second part of (16), it follows that exponential distribution of T_2 implies exponential T_1. Also, it is easy to check that if $T_2 \sim \exp(\lambda, \rho)$, then the random variable with distribution function $F_{T_1}(z)$, is exponentially distributed with parameter λ.

- The converse result is also true. If $T_1 \sim \exp(\lambda)$, then from (8), it follows that $F_{T_2}(x) = 1 - ET_2 \lambda e^{-\lambda x}$, i.e., $F_{T_2}(x)$ belongs to the class (18), with $c = \lambda ET_2$.

  - If $ET_2 = \frac{1}{\lambda}$, then $c = 1$, and $T_2 \sim exp(\lambda)$.
  - If $ET_2 = \frac{1-\rho}{\lambda}$, then $c = 1 - \rho$, and $T_2 \sim exp(\lambda, \rho)$.

□

Theorem 5.1 *$T_1 \sim exp(\lambda)$ and $T_2 \sim exp(\lambda, \rho)$ iff $N(t)$ is a Pólya-Aeppli process.*

Proof. The proof follows from Lemma 5.1 and the equivalence of Definition 2.1 and Definition 2.2, given in Section 2.

□

5.2 Characterization II

Our second PAP characterization is based on Serfozo's result in [8], p.145, which gives characterization properties of the delayed renewal processes. We summarize this result in the following lemma:

Lemma 5.2 *The delayed renewal process $N(t)$ with intensity $\mu > 0$ is stationary if and only if*

C1: $E(N(t)) = \mu t$ and
C2: $F_{T_1}(t) = \mu \int_0^t [1 - F_{T_2}(x)]dx$.

The second PAP characterization is based on Lemma 5.2, and it is given by the following theorem:

Theorem 5.2 *The delayed renewal process with finite number of arrivals over a finite interval is stationary if and only if it is a Pólya - Aeppli process.*

Proof.

- Firstly we show that PAP is delayed stationary renewal process.

  According to the Definition 2.2, PAP can be viewed as a delayed renewal process, (see subsection 2.2), with first arrival time $T_1 \sim exp(\lambda)$ followed by i.i.d. interarrival times $T_2 \sim exp(\lambda, \rho)$. Using the fist definition of PAP as a compound Poisson process, we know that $EN(t) = \frac{\lambda}{1-\rho}t$. Next, using *C1* of Lemma 5.2, we set $\mu = \frac{\lambda}{1-\rho}$. Next we use (7) and the definitions of T_1 and T_2, with

  $$ET_2 = \frac{1-\rho}{\lambda},$$

  which leads to *C2* with the same value of μ. Therefore, PAP satisfies Lemma 5.2, hence PAP can be viewed as a delayed stationary renewal process.

- From Lemma 5.1 and the condition *C2* of Lemma 5.2 it follows immediately that the stationary delayed renewal process is PAP.

□

5.3 Characterization III

Let us consider the point process $0 = \tau_0 < \tau_1 \leq \dots$, such that $T_i = \tau_i - \tau_{i-1}, \ \ i = 1, 2, \dots$. Then $N(t)$ is the number of points in the interval $(0, t]$. The next theorem gives the characterization of the Pólya - Aeppli process as a point process.

Theorem 5.3 *A counting process $N(t)$ defined on the points $0 = \tau_0 < \tau_1 \leq \dots$, has stationary independent increments, a finite number of arrivals on finite time intervals, and satisfies the condition*

$$P(\tau_i = \tau_{i+1}) = \rho, \quad i = 1, 2, \dots \tag{19}$$

iff it is a Pólya - Aeppli process.

Proof.

- Firstly, we note that if a process is PAP, then it has stationary and independent increments. Moreover, due to definition 2.2, the distribution of the second onwards interarrival times is exponential with ρ mass at zero, i.e., $P(T_i = 0) = \rho$ for $i = 2, 3, \dots$, which is equivalent to (19).

- Now we aim to prove that if a point process with finite number of arrivals on a finite time interval has stationary and independent increments over the points $0 = \tau_0 < \tau_1 \leq \ldots$ and satisfies (19), then it is PAP.

  In Th.2.2.II, p.27 in Daley and Vere-Jones [2], the authors show that if a point process is with stationary and independent increments and has a finite number of arrivals over finite time interval, then it can be viewed as a compound Poisson process. The condition (19) means that for $i = 1, 2, \ldots$

  $$P(T_{i+1} = 0) = \rho. \tag{20}$$

  Therefore, the first interarrival time is exponential (due to Th. 2.2.II, p.27 in Daley and Vere-Jones [2]) followed by i.i.d. exponentially distributed with ρ mass at zero interarrival times (due to (20)), which according to definition 2.2, means that the considered point process is PAP.

$\square$

Remark 5.1 *If the counting process is defined on the points* $0 = \tau_0 < \tau_1 < \ldots$, *the statement of the Theorem 5.3 gives a characterization of the homogeneous Poisson process, see Cont and Tankov [1], Lemma 2.1 and Serfozo [8], p. 110, Remark 21. This is the case when* $\rho = 0$ *and the geometric compounding distribution degenerates, so its support becomes* $\{1\}$.

6 Conclusions

In this paper we have studied Pólya - Aeppli process, as an extension of the standard homogeneous Poisson process, aiming to address the equidispersion of the Poisson process. As mentioned earlier, the equidispersion property of the Poisson process could be unacceptable in some situations when dealing with modelling of count data. We have shown that Poisson process is a particular case of Pólya - Aeppli processes and that Pólya - Aeppli process could be over-dispersed, which provides greater flexibility in modeling count data. We have identified three possible definitions of the Pólya - Aeppli process in prove that these definitions are equivalent. These definitions allow for further study of this process using variety of mathematical tools. Some interesting and important properties of the Pólya - Aeppli process are also included in this study. Moreover, we formulated three characterization results for the Pólya - Aeppli process, which is of significant interest in the studies related to point processes. Next we will focus on further exploration of PAP and its possible applications. For example, inference related to PAP is an open interesting question that we aim to address in our future studies.

Acknowledgment

The second author was partially supported by the Bulgarian NSRF grant DDVU 0290/2010.

References

[1] Cont R. and Tankov P. (2004). *Financial Modelling With Jump Processes*, Chapmen & Hall. 5.1

[2] Daley D. J. and Vere-Jones D. (2003), An Introduction to the theory of Point Processes. Volume I: Elementary Theory and Methods, Second ed. Springer-Verlag. 5.3, 5.3

[3] Galambos J. and Kotz S. (1978). *Characterizations of Probability Distributions,* Springer - Verlag. 5.1

[4] Johnson N. L., Kemp A. W. and Kotz S. (2005). *Univariate Discrete Distributions*, Wiley Series in Probability and Mathematical Statistics, third edition. 1

[5] Minkova L.D. (2001). A Family of Compound Discrete Distributions, *Compt. Randue Bulg. Acad. Aci.,* **54**, 11 - 14. 1

[6] Minkova L.D. (2002). A Generalization of the Classical Discrete Distributions, *Commun. Statist. - Theory and Methods,* **31**, 871 - 888. 1

[7] Minkova L.D. (2004). The Pólya - Aeppli process and ruin problems, *J. Appl. Math. Stoch. Analysis,* **3**, 221 - 234. 1, 2.1, 3

[8] Serfozo R. (2009). *Applied Stochastic Processes,* Springer - Verlag Berlin Heidelberg. 3, 5.2, 5.1