ISSN 1995-0802, Lobachevskii Journal of Mathematics, 2021, Vol. 42, No. 13, pp. 3141–3149. © Pleiades Publishing, Ltd., 2021.

On the Partial-Geometric Distribution: Properties and Applications

Krisada Khruachalee[1*], Winai Bodhisuwan[#1], and Andrei Volodin[2***]**

(Submitted by A. M. Elizarov)

[1] *Department of Statistics, Kasetsart University, Bangkok, 10903 Thailand*

[2] *Department of Mathematics and Statistics, University of Regina, Regina, SK, S4S0A2 Canada*

Received August 17, 2021; revised August 31, 2021; accepted September 5, 2021

Abstract—In this article we introduce the new, two-parameter partial-geometric distribution (PG) that contains both geometric and first success distributions as a particular case. Some probability and statistical properties of the proposed distribution are discussed, including probability mass function, mean, variance, moment generating function, and probability generating function. We propose the method of maximum likelihood for estimating the model's parameters, and apply the PG distribution to two real datasets to illustrate the flexibility of the proposed distribution. We found the PG distribution is more dynamic than the geometric distribution in the sense that it can be applied to the under-dispersed data. The PG distribution also performs well with a goodness of fit test and some other model selection characteristics for model fitting of these two datasets. Thus, the PG distribution can be applied as an alternative model for the analysis of discrete data.

DOI: 10.1134/S1995080222010103

Keywords and phrases: *geometric distribution, first success distribution, maximum likelihood estimation method.*

1. INTRODUCTION

There is much interest in developing the most flexible probability distributions; many generalized classes of distributions have been developed and applied to describe various natural phenomena [1]. To provide an explanation of a natural phenomenon, researchers consider a construction of the new generalized class of distributions, and decide whether the underlying distribution should be regarded as discrete, continuous, or of a mixed type. The discrete distributions are very useful in many applications especially when the count phenomenon consists of non-negative integers. This happens when the number of times a discrete event occurrences are observed and examined in a specific area or period of time [2]. Examples include the number of trips per month that a person takes, the number of children a couple has, the number of Prussian soldier deaths during the Crimean war resulting from being kicked by a horse (a famous classical example related to the Poisson distribution); see [3–5], etc. In these situations, the continuous model is inappropriate to describe the count phenomenon. Accordingly, the discrete models are as significant as the continuous models.

Situations where a number of trials or experiments must occur before a predetermined number of successes, such as the number of bills that must be proposed to a legislature before 10 bills are passed, and recently the number of Thai citizens that must be tested for COVID-19 before 100 Thai citizens are confirmed to be infectious, are of the interest to many researchers keen to find suitable probability distributions that explain these natural phenomena.

[*] E-mail: krisada.khr@ku.th

[**] E-mail: fsciwnb@ku.ac.th

[***] E-mail: andrei@uregina.ca

[#] Corresponding author.

In addition, the complications of comparing the probabilities of success for the Bernoulli trials that mostly arise in medical and biological investigations, acceptance sampling in quality control, and modeling demand for a product are considered. These real-life phenomena can be described by the geometric distribution. However, there are criticisms that the geometric and first success distributions are sometimes considered to be the same, when they are actually different.

Because the confusion between the geometric and first success distributions plays a very important role in this study, we would like to introduce a new family of distributions called the partial-geometric (PG) distribution that contains both the geometric and first success distributions as a particular case. The idea to combine and consider the geometric and first success distributions as a member of one distribution family seems to be very natural, but we did not see it in literature. The number of the studies that propose modifications of the geometric distribution for various purposes is so large that we decided not to discuss them. The major advantage of the newly proposed PG distribution over the previously modified geometric distributions is the flexibility in applications to real-life data.

The remainder of the paper is organized as follows. In Section 2, we discuss the difference between geometric and first success distributions. The PG distribution and some of its mathematical properties are defined in Section 3. Then, the maximum likelihood estimates of the PG distribution parameters are discussed in Section 3.3. Finally, some practical applications of the proposed distribution are illustrated by a goodness of fit with two real datasets in Section 4.

2. MATERIALS AND METHODS

Based on the theoretical interpretation of the Bernoulli experiment, we note a confusion between two very simple and basic geometric and first success distributions. In some literature, these two distributions are considered the same. But, as mentioned in Gut (2009) [6], they are different and defined in the following way.

Let $0 < p < 1$ and $q = 1 - p$. A random variable X has a *Geometric* distribution with parameter p, denoted by $X \sim G(p)$, if its probability mass function (pmf) is

$$P(X = k) = pq^k, \quad k = 0, 1, 2, \ldots.$$

A random variable Y has a *First Success* distribution with parameter p, denoted by $Y \sim FS(p)$, if its pmf is

$$P(Y = k) = pq^{k-1}, \quad k = 1, 2, \ldots.$$

We can interpret the geometric distribution as the number of *failures* in Bernoulli experiments until we reach first successes, while the first success distribution is the number of *trials* in Bernoulli experiments need to reach the first success.

Referring to the properties of the probability generating function (pgf), the pgf of the geometric distribution $G(p)$ can be calculated in the following way:

$$g_{G(p)}(t) = Et^X = \sum_{k=0}^{\infty} t^k pq^k = p\sum_{k=0}^{\infty}(tq)^k = \frac{p}{1-qt},$$

where $|t| < 1/q$. In the same manner, the pgf of the first success distribution $FS(p)$ can be also calculated in accordance with the following procedure:

$$g_{FS(p)}(t) = Et^Y = \sum_{k=1}^{\infty} t^k pq^{k-1} = \frac{pt}{1-qt},$$

where again $|t| < 1/q$. According to the pgf of the geometric and first success distributions, we present Proposition 1 that can be used to illustrate the connection of these two distributions as follows.

Proposition 1. *If a random variable $X \sim G(p)$, then the random variable $Y = X + 1$ follows $FS(p)$ distribution. Similarly, if a random variable $Y \sim FS(p)$, then the random variable $X = Y - 1$ follows $G(p)$ distribution.*

Proof. If a random variable $X \sim G(p)$, then the pgf of the random variable $Y = X + 1$ can be written

$$g_Y(t) = Et^Y = Et^{X+1} = tEt^X = tg_{G(p)}(t) = \frac{pt}{1-qt} = g_{FS(p)}(t).$$

The second part of the Proposition can be shown in the similar way. □

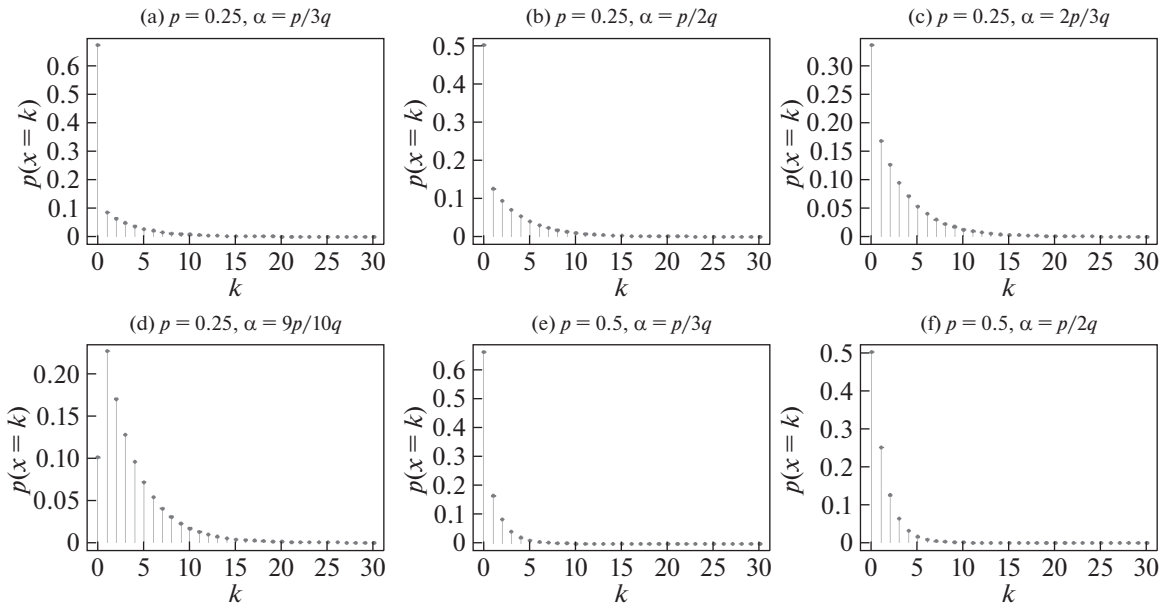

Fig. 1. The pmf plots of the partial-geometric distribution in various combinations of p (0.25 and 0.50) and α ($p/3q$, $p/2q$, $2p/3q$, and $9p/10q$).

3. PARTIAL-GEOMETRIC DISTRIBUTION AND ITS MATHEMATICAL PROPERTIES

3.1. Partial-Geometric (PG) Distribution

Changing the momentum by adding an extra parameter α, leads us to propose the PG distribution. A random variable X has a Partial-Geometric distribution with parameters $0 < p < 1$ and $0 \leq \alpha \leq \dfrac{p}{1-p}$, denoted by $PG(p, \alpha)$, if its pmf is

$$P(X = k) = \begin{cases} \alpha(1-p)^k & \text{if } k = 1, 2, \ldots \\ \beta & \text{if } k = 0, \end{cases}$$

where $\beta = 1 - \alpha(1-p)/p$.

In order to illustrate the appearance of the PG distribution, Figs. 1 and 2 show some pmf plots of the PG distribution with various values of the parameters p (0.25, 0.50 and 0.75) and α ($p/3q$, $p/2q$, $2p/3q$ and $9p/10q$) where $q = 1 - p$. We found that the scale of the PG distribution change due to the parameter p. On the other hand, the shape parameter of the PG distribution can be varied because of the parameter α. It is seen that the pmf rapidly decreases as parameter p increases. In addition, the PG distribution is clearly a unimodal curve when the α conversely increases to p/q. According to Figs. 1 and 2, we conclude that the PG distribution is right skewed and unimodal.

Measuring the dispersion of the partial-geometric (PG) distribution, the ID, the ratio between variance to mean, is applied under some specified values of parameters p and α from Figs. 1 and 2. The values of ID will indicate whether the distribution is over-dispersed ($ID > 1$) or under-dispersed ($ID < 1$) [7]. Table 1 illustrates that the partial-geometric (PG) distribution is more dynamic than the geometric distribution in the sense that it can be applied to the under-dispersed data as well where the geometric distribution is only suitable for over-dispersed data.

3.2. Probability Properties

Some probability properties of the PG distributions, especially the mean, variance, moment generating function (mgf), and pgf are provided in this section.

Theorem 1. *Let $X \sim PG(p, \alpha)$, then the mean of X is $E(X) = \frac{\alpha(1-p)}{p^2}$.*

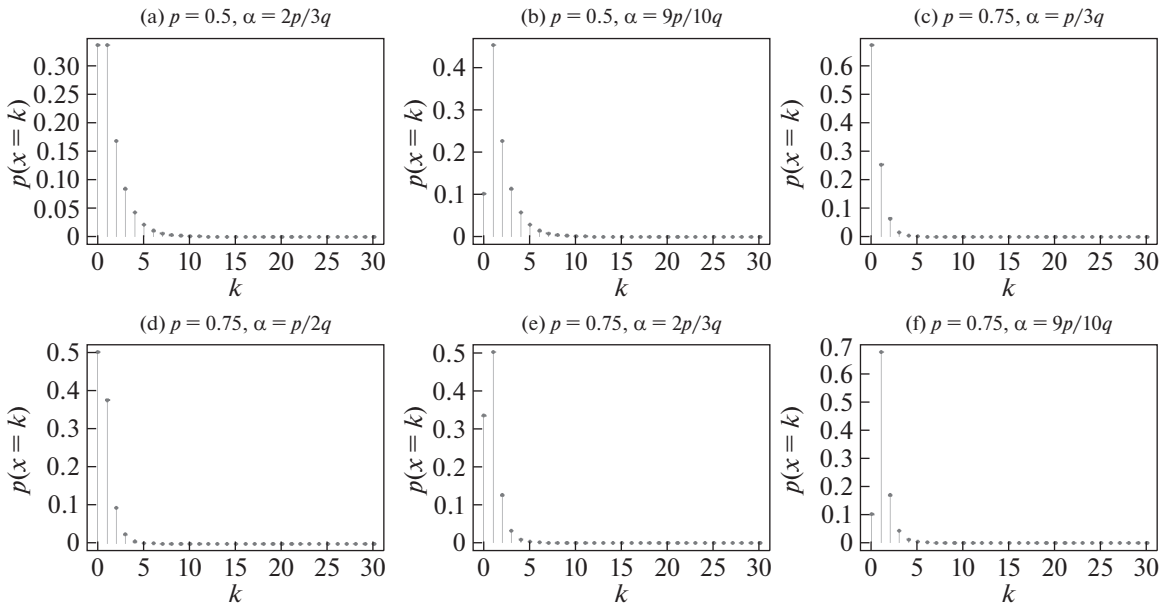

Fig. 2. The pmf plots of the partial-geometric distribution in various combinations of p (0.5 and 0.75) and α ($p/3q$, $p/2q$, $2p/3q$, and $9p/10q$).

Proof. The expectation of the PG distribution can be obtained from

$$E(X) = \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=1}^{\infty} k\alpha(1-p)^k = \alpha(1-p)\sum_{k=1}^{\infty} k(1-p)^{k-1}.$$

Since $\sum_{k=1}^{\infty} k(1-p)^{k-1} = \frac{1}{p^2}$ is the geometric series, the expectation will be $E(X) = \frac{\alpha(1-p)}{p^2}$. □

Table 1. The mean, variance and index of dispersion (ID) values of the partial-geometric (PG) distribution for different value of p and α

Figure	p	α	$E(X)$	$V(X)$	ID
1(a)	0.25	0.1111	1.3333	7.5556	5.6667
1(b)	0.25	0.1667	2.0000	10.0000	5.0000
1(c)	0.25	0.2222	2.6667	11.5556	4.3333
1(d)	0.25	0.3000	3.6000	12.2400	3.4000
1(e)	0.50	0.3333	0.6667	1.5556	2.3333
1(f)	0.50	0.5000	1.0000	2.0000	2.0000
2(a)	0.50	0.6667	1.3333	2.2222	1.6667
2(b)	0.50	0.9000	1.8000	2.1600	1.2000
2(c)	0.75	1.0000	0.4444	0.5432	1.2222
2(d)	0.75	1.5000	0.6667	0.6667	1.0000
2(e)	0.75	2.0000	0.8889	0.6914	0.7778
2(f)	0.75	2.7000	1.2000	0.5600	0.4667

Theorem 2. *Let $X \sim PG(p,\alpha)$, then the variance of X is*

$$\mathrm{Var}(X) = \frac{\alpha(1-p)\left(p(2-p)-\alpha(1-p)\right)}{p^4}.$$

Proof. The variance of the PG distribution can be obtained from

$$\mathrm{Var}(X=k) = E(X-E(X))^2 = E(X(X-1)) + E(X) - (E(X))^2.$$

With the expectation definition,

$$E(X(X-1)) = \sum_{k=0}^{\infty} k(k-1)P(X=k) = \sum_{k=2}^{\infty} k(k-1)\alpha(1-p)^k$$
$$= \alpha(1-p)^2 \sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2}.$$

Since $\sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2} = \frac{2}{p^3}$ is the geometric power series, then the $E(X(X-1))$ will be equal to $2\alpha(1-p)^2/p^3$. Therefore,

$$\mathrm{Var(X)} = \frac{2\alpha(1-p)^2}{p^3} + \frac{\alpha(1-p)}{p^2} - \left(\frac{\alpha(1-p)}{p^2}\right)^2 = \frac{\alpha(1-p)\left(p(2-p)-\alpha(1-p)\right)}{p^4}.$$

□

Theorem 3. *Let $X \sim PG(p,\alpha)$, then the mgf of X is $M_X(t) = \beta + \frac{\alpha(1-p)e^t}{1-(1-p)e^t}$,* where $|e^t| < 1/1-p$.

Proof. The mgf of the PG distribution can be achieved from

$$M_X(t) = \sum_{k=0}^{\infty} e^{tk}P(X=k) = e^{t\times 0}P(X=0) + \sum_{k=1}^{\infty} e^{tk}P(X=k)$$
$$= \beta + \sum_{k=1}^{\infty} e^{tk}\alpha(1-p)^k = \beta + \alpha\sum_{k=1}^{\infty}(e^t(1-p))^k.$$

Since $\sum_{k=1}^{\infty}(e^t(1-p))^k = \frac{(1-p)e^t}{1-(1-p)e^t}$ is the geometric series, then the mgf will be equals $\beta + \frac{\alpha(1-p)e^t}{1-(1-p)e^t}$, where $|e^t| < 1/1-p$. □

Theorem 4. *Let $X \sim PG(p,\alpha)$, then the pgf of X is $g_X(t) = \beta + \frac{\alpha(1-p)t}{1-(1-p)t}$, where $|t| < 1/(1-p)$.*

Proof. The pgf of the PG distribution can be acquired from

$$g_X(t) = \sum_{k=0}^{\infty} t^k P(X=k) = t^0 P(X=0) + \sum_{k=1}^{\infty} t^k P(X=k) = \beta + \alpha\sum_{k=1}^{\infty}(t(1-p))^k.$$

Since $\sum_{k=1}^{\infty}(t(1-p))^k = \frac{(1-p)t}{1-(1-p)t}$ is the geometric series, then the pgf will be equals $\beta + \frac{\alpha(1-p)t}{1-(1-p)t}$, where $|t| < 1/(1-p)$. □

3.3. Parameter Estimation

We consider the maximum likelihood estimation (MLE) that is the most commonly used method for parameter estimation. Let $X_1, X_2, \ldots, X_n$ be an independent and identically distributed random sample of size n from the partial-geometric distribution, $PG(p,\alpha)$, and $x_1, x_2, \ldots, x_n$ be the observed sample

values. For $k \geq 0$ denote the frequencies $f_k = \#\{i : x_i = k\}$, that is, f_k is the count of observations that are equal to k. Note that

$$\sum_{k=0}^{\infty} f_k = n \quad \text{and} \quad \sum_{k=0}^{\infty} k f_k = \sum_{k=1}^{\infty} k f_k = n\overline{X}.$$

The likelihood function can be written as

$$\mathcal{L}(p, \alpha | x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{k=0}^{\infty} [P(X = k)]^{f_k}$$

$$= [P(X = 0)]^{f_0} \prod_{k=1}^{\infty} [P(X = k)]^{f_k} = \left(1 - \frac{\alpha(1-p)}{p}\right)^{f_0} \prod_{k=1}^{\infty} \left(\alpha(1-p)^k\right)^{f_k}$$

$$= \left(1 - \frac{\alpha(1-p)}{p}\right)^{f_0} \alpha^{\sum_{k=1}^{\infty} f_k} (1-p)^{\sum_{k=1}^{\infty} k f_k} = \left(1 - \frac{\alpha(1-p)}{p}\right)^{f_0} \alpha^{n - f_0} (1-p)^{n\overline{X}}$$

$$= \left(1 + \frac{1}{\alpha} - \frac{1}{p}\right)^{f_0} \alpha^n (1-p)^{n\overline{X}}.$$

The log-likelihood function can be written as

$$\ell(p, \alpha | x_1, x_2, \ldots, x_n) = \log \mathcal{L}(p, \alpha | x_1, x_2, \ldots, x_n)$$

$$= f_0 \log\left(1 + \frac{1}{\alpha} - \frac{1}{p}\right) + n \log(\alpha) + n\overline{X} \log(1-p).$$

By taking partial derivatives by the parameters, we obtain

$$\frac{\partial \ell}{\partial p} = \frac{f_0}{1 + \frac{1}{\alpha} - \frac{1}{p}} \frac{1}{p^2} - \frac{n\overline{X}}{1-p}, \quad \frac{\partial \ell}{\partial \alpha} = -\frac{f_0}{1 + \frac{1}{\alpha} - \frac{1}{p}} \frac{1}{\alpha^2} + \frac{n}{\alpha}.$$

The method of maximum likelihood estimators are found by equating the partial derivatives to zero; that is

$$\frac{f_0}{1 + \frac{1}{\alpha} - \frac{1}{p}} \frac{1}{p^2} = \frac{n\overline{X}}{1-p}, \quad \frac{f_0}{1 + \frac{1}{\alpha} - \frac{1}{p}} \frac{1}{\alpha^2} = \frac{n}{\alpha}.$$

By rewriting the second equation and substituting to the first equation, we obtain the estimated maximum likelihood parameters $\hat{\alpha}$ and $\hat{p}$ of the PG distribution as

$$\hat{\alpha} = \frac{\left(1 - \frac{f_0}{n}\right)^2}{\overline{X} + \frac{f_0}{n} - 1} \quad \text{and} \quad \hat{p} = \frac{\sqrt{\hat{\alpha}^2 + 4\overline{X}\hat{\alpha}} - \hat{\alpha}}{2\overline{X}}.$$

4. APPLICATIONS TO REAL LIFE DATA

In order to evaluate the performance of the PG distribution, we consider two real datasets to fit with the two competing geometric and Poisson distributions. We do not consider the first success distribution as a competing because the data contain zeros.

The first dataset is accident data that provides the total number of the claims for 9,461 automobile insurance policies [10]. The second dataset is the number of hospitals stays of persons age 66 and over, for which there are 4,406 observations. These data were acquired from the national medical expenditure survey of how Americans use and pay for health services conducted in 1987 and 1988 to reveal a comprehensive picture of medical expenditure [11].

Table 2. Estimated parameters for the number of cliams of the automobile insurance policies

Number of claims	Observed frequency	Expected value by fitting distribution		
		Geometric	Partial-Geometric	Poisson
0	7840	7790.542	7840.123	7636.354
1	1317	1375.518	1295.538	1636.148
2	239	242.865	260.038	175.279
3	42	42.881	52.194	12.518
4	14	7.571	10.476	0.671
5	4	1.337	2.103	0.029
6	4	0.236	0.422	0.001
7	1	0.042	0.085	0.000
Estimates		$\hat{p} = 0.823$	$\hat{p} = 0.823, \hat{\alpha} = 0.682$	$\hat{\lambda} = 0.214$
LL		-5354.681	-5349.860	-5490.781
AIC		10711.36	10703.72	10983.56
BIC		10718.52	10718.03	10990.72
AD test		1.013	0.160	23.362
p-value		0.177	0.706	< 0.001

The appropriate distribution for fitting these datasets is evaluated with the Anderson-Darling (AD) goodness of fit test for discrete data [12]. In addition, the discrete AD test is obtained by applying the *dgof* package [13] in the R language. Moreover, there are also other model selection criteria used to determine the best fit model: the minus log-likelihood (-LL), the Akaike information criterion (AIC), and the Bayesian information criterion (BIC). The results of fitting different distributions to these datasets are recorded in Tables 2 and 3.

The fitted distributions for the number of claims and the number of hospital stays presented in Tables 2 and 3 illustrate that the p-value based on the discrete AD test statistic of the PG distribution provides a good fit to the data where it provides the largest p-value among others. Moreover, the PG distribution provides the lowest values of -LL, AIC and BIC. Obviously, the PG distribution provides the nearest expected value to the observed frequency. Therefore, the most appropriate fit distribution among these three distributions for the number of claims and the number of hospital stays is the PG distribution followed by the geometric and Poisson distributions respectively.

Figure 3 illustrates the plots of fitted frequency of the geometric, partial-geometric, and Poisson distributions with the observed datasets for the total number of claims of the automobile insurance policies and the number of hospital stays. It firmly shows that the partial-geometric (PG) distribution provides the most fitted performance to these two datasets among the three distributions. Thus, we consistently conclude that the PG distribution is more flexible than the geometric and Poisson distributions.

5. DISCUSSION AND CONCLUSION

Confusion between geometric and first success distributions led to the idea of developing a new family of distributions. By adding an extra parameter to an existing distribution for capturing more variation of the natural phenomena, the partial-geometric distribution that contains both geometric and first success distributions as a particular case is proposed. We found that the PG distribution is right skewed and unimodal. Moreover, it can also be applied to model under-dispersed data. We also derived some essential probability properties, for instance, probability mass function, mean, variance, moment generating function, and probability generating function. The maximum likelihood estimation method is

Table 3. Estimated parameters for the number of hospital stays

Number of hospital stays	Observed frequency	Expected value by fitting distribution		
		Geometric	Partial-Geometric	Poisson
0	3541	3399.590	3541.091	3277.140
1	599	776.528	573.702	970.021
2	176	177.373	193.160	143.561
3	48	40.515	65.035	14.165
4	20	9.254	21.897	1.048
5	12	2.114	7.372	0.062
6	5	0.483	2.482	0.003
7	1	0.110	0.836	0.000
8	4	0.025	0.281	0.000
Estimates		$\hat{p} = 0.772$	$\hat{p} = 0.772, \hat{\alpha} = 0.387$	$\hat{\lambda} = 0.296$
LL		-3067.988	-3015.114	-3304.509
AIC		6137.976	6034.228	6611.019
BIC		6144.366	6047.009	6617.409
AD test		14.182	0.303	59.875
p-value		< 0.001	0.540	< 0.001

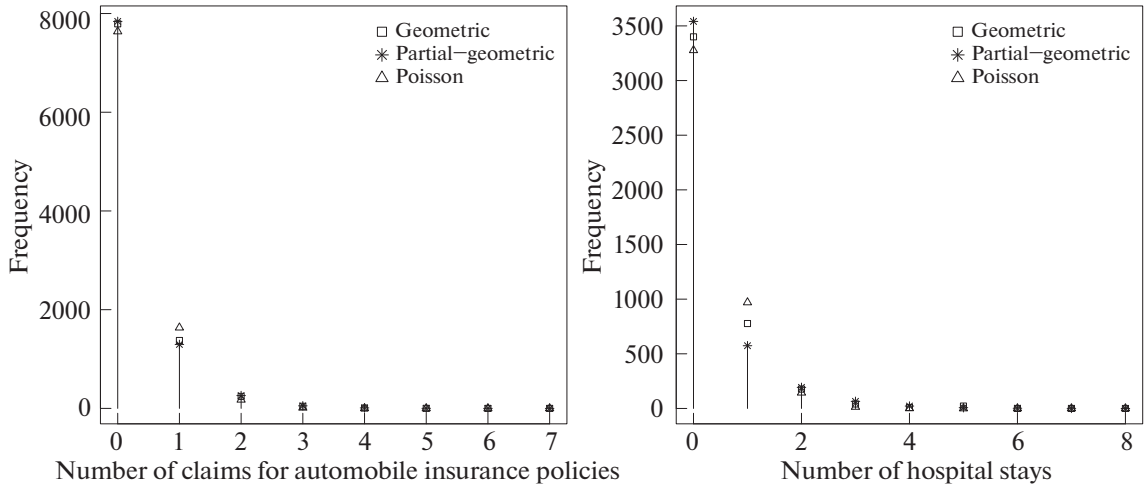

Fig. 3. The fitted frequency of the geometric, partial-geometric and Poisson distributions to real datasets.

employed to estimate the parameters of the PG distribution. Due to the practical applications with two real datasets, the PG distribution provides the highest p-value for the discrete AD test and also provides the lowest values of -LL, AIC and BIC as well. Therefore, the PG distribution is useful as an alternative to other distribution for the analysis of discrete data.

ACKNOWLEDGMENTS

The authors would like to thank the Department of Statistics, Faculty of Science, Kasetsart University.

FUNDING

The research of the author listed last was partially supported by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, project 1.13556.2019/13.1.

REFERENCES

1. A. Alzaatreh, C. Lee, and F. Famoye, "A new method for generating families of continuous distributions," METRON **71**, 63–79 (2013).
2. N. L. Johnson and S. Kotz, *Distribution in Statistics: Discrete Distributions* (Houghton Mifflin, New York, 1969).
3. L. V. Bortkiewicz, *Das Gesetz der kleinen Zahlen* (G. Teubner, Leipzig, 1898).
4. C. P. Winsor, "Quotations: Das Gesetz der kleinen Zahlen," Human Biology **19**, 154–161 (1947).
5. W. Weaver, *Lady Luck: The Theory of Probability* (Heinemann, London, 1964).
6. A. Gut, *An Intermediate Course in Probability* (Springer Science, New York, 2009).
7. S. Chakraborty and D. Chakravarty, "Discrete gamma distributions: Properties and parameter estimations," Commun. Stat.—Theory Method **41**, 3301–3324 (2012).
8. J. C. Nash, "On best practice optimization methods in R," J. Stat. Software **60** (2), 1–14 (2014).
9. R Core Team, *R: A Language and Environment for Statistical Computing* (2020).
10. S. A. Klugman, H. H. Panjer, and G. E. Willmot, *Loss Models: From Data to Decisions* (Wiley, Hoboken, NJ, 2012).
11. P. Deb and P. K. Trivedi, "Demand for medical care by the elderly: A finite mixture approach," J. Appl. Econometr. **12**, 313–336 (1997).
12. V. Choulakian, R. A. Lockhart, and M. A. Stephens, "Cramer-von Mises statistics for discrete distributions," Canad. J. Stat. **22**, 125–137 (1994).
13. T. B. Arnold and J. W. Emerson, "Nonparametric goodness-of-fit tests for discrete null distributions," R J. **3** (2), 34–39 (2011).