Thailand Statistician
July 2022; 20(3): 603-614
http://statassoc.or.th
Contributed paper

A Size Biased Polya-Aeppli Distribution and Its Applications

Anurag Gupta [a], Hira Lal Sharma* [b] and Sanjeeta Biswas [c]

[a] School of Studies in Statistics, Vikram University, Ujjain, M.P. India.

[b] Department of Mathematics and Statistics, JN Agricultural University, Jabalpur, M.P. India.

[c] Department of Agricultural Statistics, Bidhan Chandra Krishi Vishwa Vidyalaya, Kalyani, W.B. India.

*Corresponding author; e-mail: drhlsharma_jnkvv@rediffmail.com

Received: 22 November 2019
Revised: 3 August 2020
Accepted: 10 August 2020

Abstract

This paper is concerned with a new extension to Polya-Aeppli distribution called as size-biased Polya-Aeppli distribution (SBPAD). This SBPAD consists of two parameters and these are estimated by method of proportion of one'th cell (MPOC), method of moments (MM) including the estimating equations and maximum likelihood estimation (MLE). The expressions of its moment and probability generating functions have also been derived giving the relationship between mean and variance of SBPAD. At the end, the applications of the distribution to various four sets of data have also been added and found that the distribution describes the pattern of the data satisfactorily well.

Keywords: Method of moments, maximum likelihood, goodness of fit.

1. Introduction

Size-biased distributions are important distributions which provide a special approach for the problems where the gathered observations in the survey fall in the non-experimental, non- replicated, and nonrandom categories as well as the varying probability for the items to be included, always known as weighted distributions. It was first introduced by Fisher (1934) to model ascertainment bias, and later by Rao (1965) in a situation where distributions consider the observations from a sample recorded with unequal probability, such as from probability proportional to size (PPS) designs. During last few decades a number of papers have been derived and proposed during a period of time utilizing the concept of weighted and size-biased distributions and their applications in various fields. Recently, the various kinds of applications have been shown by different authors in different fields like engineering, medical, forestry, agriculture and social science (Adhikari and Srivastava (2014), Priti and Singh (2015), Ducey and Gove (2015), Shanker and Shukla (2017), Rather et al. (2018), Sium and Shanker (2018), Yadav and Kumar (2018), Beghriche and Zeghdoudi (2019).

The objective of the present paper is to derive a new extension to Polya-Aeppli distribution called as size biased Polya-Aeppli distribution (SBPAD). Section 2 provides the SBPAD and its method of estimation of parameters. Section 3 deals with derivation of moment generating and probability generating function. At the end, the applications of the distribution have been added.

2. Size Biased Polya-Aeppli Distribution

Polya-Aeppli distribution arises in a model formed by supposing that the objects (which are to be counted) occur in groups, the number of groups follows a Poisson distribution with parameter θ while the number of objects per group has the geometric distribution with parameter p.

This distribution is defined by

$$P[X=0]=e^{-\theta} \text{ for } k=0,$$

$$P[X=k]=e^{-\theta}p^{k}\sum_{j=1}^{k}\binom{k-1}{j-1}\left(\frac{\theta q}{p}\right)\Big/(j)! \text{ for } k\geq 1. \tag{1}$$

Let X be a random variable following a probability distribution with parameter θ such that $P[X=k,\theta], k=0,1,2,...,\theta>0$. Suppose the sample units are being taken from the distribution with probability proportional to size of unequal items. Then the corresponding size-biased distribution can be defined by its probability mass function $P_1[X=k]=w(k).[P(X=k,\theta)]/E[w(k)]$ where $w(k)$ is a non-negative weight function and $E[w(k)]$ is called the mean of the original.

Then, SBPAD is derived from original distribution as given in (1)

$$P[X=k]=e^{-\theta}p^{k}k\sum_{j=1}^{k}\binom{k-1}{j-1}\left[\left(\frac{\theta q}{p}\right)\Big/(j)!\right]\frac{q}{\theta} \text{ for } k=1,2,3,...,. \tag{2}$$

After re-arranging the terms and some algebraic mathematical manipulation in (2) which may be expressed as

$$P[X=k]=\frac{e^{-\theta}p^{k+1}}{\theta^{2}}\sum_{j=1}^{k}\binom{k}{j}\left[\left(\frac{\theta q}{p}\right)^{j+1}\frac{1}{(j-1)!}\right]. \tag{3}$$

Equation (3) shall be declared the probability mass function of SBPAD because $\sum_{k=1}^{\infty}p[X=k]=1,$ which may be shown as given below:

Substituting $k=1,2,3,...,$ in (3), successively, we have

$$P[X=1]=\frac{e^{-\theta}p^{2}}{\theta^{2}}\left[\binom{1}{1}\left(\frac{\theta q}{p}\right)^{2}\frac{1}{0!}\right],$$

$$P[X=2]=\frac{e^{-\theta}p^{3}}{\theta^{2}}\left[\binom{2}{1}\left(\frac{\theta q}{p}\right)^{2}\frac{1}{0!}+\binom{2}{2}\left(\frac{\theta q}{p}\right)^{3}\frac{1}{1!}\right],$$

$$P[X=3]=\frac{e^{-\theta}p^{4}}{\theta^{2}}\left[\binom{3}{1}\left(\frac{\theta q}{p}\right)^{2}\frac{1}{0!}+\binom{3}{2}\left(\frac{\theta q}{p}\right)^{3}\frac{1}{1!}+\binom{3}{3}\left(\frac{\theta q}{p}\right)^{4}\frac{1}{2!}\right] \text{ and so on.}$$

Whose sum $P[X=1]+P[X=2]+P[X=3]+...$ is equal to one. It indicates that (3) is a probability mass function of SBPAD.

2. Method of Estimation of SBPAD

2.1. Method of proportion of one'th cell

SBPAD consists of two parameters θ and p and these are estimated by equating the observed proportion of one'th cell $\left(P_1\right)$ of the model and mean value to its theoretical values, which are given below:

$$P\left[X=1\right]=P_1=\frac{N_1}{N}=e^{-\theta}q^2, \tag{4}$$

$$m_1=\frac{1+\theta+p}{q}, \tag{5}$$

where N_1, N and m_1 are first cell frequency, total frequency and mean value of observed frequency distribution.

From (4), $q^2=P_1e^{\theta}$ implies that $e^{\theta}=\frac{q^2}{P_1}$ implies that $\theta=\log\left[\frac{q^2}{P_1}\right]$ where log stands for logarithm to the base e. Substituting this value of θ in (5), we have the estimating equation of value q,

$$m_1=q^{-1}\left[1+\log\left(\frac{q^2}{P_1}\right)+p\right]. \tag{6}$$

Using method of iteration from (6), the estimate of p can be determined using R statistical package (R Core team 2018). Then, θ can be estimated from the relationship

$$\theta=\log\left[\frac{q^2}{P_1}\right],$$

provided that $\log\left[\frac{q^2}{P_1}\right]\geq 1.$

2.2. Method of moments

SBPAD consists of two parameters θ and p and these are estimated by equating the observed mean and variance of the model to their theoretical values, which are given below:

$$m_1=\frac{1+\theta+p}{q}, \tag{7}$$

$$m_2=\frac{2p+\theta+\theta p}{q^2}, \tag{8}$$

where m_1 and m_2 are observed mean and variance of the actual data which have been derived from either moment generating function or probability generating function. These derivations have been given in Section 3.

From (7), the value of p is $\frac{m_1-\left(1+\theta\right)}{1+m_1}$. Substituting this value of p in (8), we have the estimating equation of value θ in the following quadratic equation of θ as

$$A\theta^2+B\theta+C=0, \tag{9}$$

where $A=1+m_1+m_2$, $B=2+4m_2-2m_1^2$, $C=2+4m_2-2m_1^2$, here $B=C.$ From (9), the estimate of θ can be determined using R statistical package (R Core team 2018).

2.3. Method of maximum likelihood

Consider a sample consisting of n observations of the random variable X with probability mass function given in (3) above. The likelihood function can be written as

$$L \simeq \prod_{k=1}^{R}\left[\frac{e^{-\theta}p^{k+1}}{\theta^2}\sum_{j=1}^{k}\binom{k}{j}\left(\frac{\theta q}{p}\right)^{j+1}\frac{1}{(j-1)!}\right]^{n_k}, \tag{10}$$

where L is the likelihood function, R is the largest observed value of k, n_k is the sample frequency of and $\prod_{i=1}^{R}$ denote the product over n non-zero observations.

Taking the logarithm of L given in (10), differentiating with respect to θ and p in turn and setting the derivatives equal to zero gives the estimating equations. We have

$$\sum_{k=1}^{R}\frac{n_k}{p_k}\left[\frac{\{(k-\theta-1)P_k-pP_{k-1}\}}{\theta}\right]=0, \tag{11}$$

$$\sum_{k=1}^{R}\frac{n_k}{p_k}\left[\frac{\{P_{k-1}-(k+1)P_k\}}{q}\right]=0. \tag{12}$$

Maximum likelihood estimates when they exist may be found by solving the system of (11) and (12) by employing a trial and error technique by which covariance is greatly accelerated (Sampford 1955).

3. Derivation of Moment Generating and Probability Generating Function

3.1. Moment generating function of SBPAD

The moment generating function $M_k(t)$ is usually defined as $M_k(t)=E\left(e^{tk}\right)$, t being real numbers, then for SBPAD, it is defined as

$$M_k(t)=\frac{e^{-\theta}p^{k+1}}{\theta^2}e^{tk}\sum_{j=1}^{k}\binom{k}{j}\left(\frac{\theta q}{p}\right)^{j+1}\frac{1}{(j-1)!}. \tag{13}$$

The equation (13) reduces after some algebraic manipulation to the simplified form as

$$M_k(t)=\frac{q^2e^t}{\left(1-pe^t\right)^2}e^{-\frac{\theta\left(1-e^t\right)}{\left(1-pe^t\right)}}. \tag{14}$$

3.2. Probability generating function of SBPAD

The probability generating function $G(s)$ is usually defined as $G(s)=E\left(s^k\right), s$ being real numbers, then for SBPAD, it is defined as

$$G(s)=\frac{e^{-\theta}p^{k+1}}{\theta^2}s^k\sum_{j=1}^{k}\binom{k}{j}\left(\frac{\theta q}{p}\right)^{j+1}\frac{1}{(j-1)!}. \tag{15}$$

The equation (15) reduces after some algebraic manipulation to the simplified form as

$$G(s) = \frac{q^2 s}{(1-ps)^2} e^{-\theta \frac{(1-s)}{(1-ps)}}. \tag{16}$$

The mean and variance of SBPAD can be determined either from (14) or (16). Let us choose the Equation (14). Differentiating Equation (14) with respect to t and substituting $t = 0,$ $\frac{d}{dt}\left[M_k(t)\right]/t = 0,$ we have mean of SBPAD

$$m_1 = \frac{1+\theta+p}{q}. \tag{17}$$

Differentiating Equation (14) with respect to t two times and substituting $t = 0,$ and by the given equation, we have the variance of SBPAD

$$\frac{d^2}{dt^2}[M_k(t)/t=0] + \frac{d}{dt}\left[M_k(t)/t=0\right] - \left[\frac{d}{dt}\{M_k(t)/t=0\}\right]^2. \tag{18}$$

Equation (18) provides the variance of SBPAD in simplified form as

$$m_2 = \frac{2p+\theta+\theta p}{q^2}. \tag{19}$$

Let us choose the Equation (16). Differentiating Equation (16) with respect to s and substituting $s = 1,$ $\frac{d}{ds}\left[G(s)\right]/s = 1,$ we have mean of SBPAD

$$m_1 = \frac{1+\theta+p}{q}. \tag{20}$$

Differentiating Equation (16) with respect to S two times and substituting $s = 1,$ and by the given equation, we have the variance of SBPAD

$$\frac{d^2}{ds^2}[G(s)/s=1] + \frac{d}{ds}\left[G(s)/s=1\right] - \left[\frac{d}{ds}\{G(s)/s=1\}\right]^2.$$

Equation (20) provides the variance of SBPAD in simplified form as

$$m_2 = \frac{2p+\theta+\theta p}{q^2}. \tag{21}$$

From both moment and probability generating functions, the same expressions of mean and variance of SBPAD have been obtained.

3.3. Relationship between mean and variance of SBPAD

We derived that the mean and variance of SBPAD as given in (17) and (19) respectively. In fact, we want to know that mean is larger than the variance. If it is possible, let us assume that $m_2 > m_1$ i.e.,

$$\frac{2p+\theta+\theta p}{q^2} > \frac{1+\theta+p}{q}.$$

It implies that

$$2p+\theta+\theta p > q+\theta q+pq,$$

$$2p+\theta(1+p) > q+\theta q+pq.$$

Replacing 1 as $p+q,$ it indicates that

$$2p + 2\theta p > q + pq,$$
$$2p(1+\theta) > q(1+p),$$

i.e. $2p(1+\theta) + p^2 > 1,$ which is impossible. It indicates that mean in SBPAD is larger rather than the variance.

4. The Applications

In order to illustrate the practical applications of results obtained here, we consider four various sets of data taken from play groups-Eugene, spring playground D and playground A relating to the size distribution of freely-forming small groups at public places which were reported by James (1953) and Coleman and James (1961). The other sets of data were related to the number of counts of sites with particulars Immungold and number of snowshoe hares captured over 7 years reported by Matthews and Appleton (1993) and Keith and Meslow (1968).

Table 1 presents the observed and expected distribution of play groups-Eugene, Spring Playground D of 510 frequencies. It is important to understand that $(1+\theta+p)/q$ provides the average of the playgroups-Eugene, spring playground D while $\hat{\theta}$ gives the average number of clusters per groups-Eugene, spring playground D giving $\hat{q}$ as the average number of individuals per spring playground D. It indicates that the higher values of θ and lower values of q provide higher average of the Spring Playground D. The estimates of parameters p and θ in MPOC, MM and MLE methods of estimation are found to be 0.1000, 0.0800, 0.0850 and 0.2679, 0.3090, 0.3100, respectively. These estimates of the parameters involved in the distribution are almost found to be equal. Once the estimates of these parameters are obtained, the expected frequencies are easily calculated. The best fitting of the distribution has been given by MLE followed by MM and MPOC method of estimation as revealed through the χ^2 values at two degrees of freedom. It seems that the distribution provides the pattern of the data satisfactorily well.

Table 1 Distribution of observed and expected number of Play groups-Eugene, Spring Playground D of 510 frequencies

Group size	Observed frequency	Expected frequency		
		MPOC	MM	MLE
1	316	316.00	315.82	313.17
2	141	139.40	142.35	142.07
3	44	41.53	40.50	42.04
4	5	7.49	9.17	10.09
5	4	5.58	2.16	2.63
Total	510	510	510	510
		Estimates of parameters		
	p	0.1000	0.0800	0.0850
	θ	0.2679	0.3090	0.3100
	χ^2	1.4404	0.9690	0.9207
	d.f.	2	2	2

Table 2 reveals the observed and expected distribution of play groups-Eugene, spring playground A of 497 frequencies. It is to be noted that $(1+\theta+p)/q$ provides the average of the play groups-

Eugene, spring playground A while $\hat{\theta}$ gives the average number of clusters per groups-Eugene, spring playground A giving $\hat{q}$ as the average number of individuals per spring playground A. It indicates that the higher values of θ and lower values of q provide higher average of the spring playground A. The estimates of parameters p and θ in MPOC, MM and MLE methods of estimation are found to be 0.1000, 0.0800, 0.1030 and 0.2743, 0.3287, 0.2760, respectively. These estimates of the parameters involved in the distribution are approximately estimated to be equal. Once the estimates of these parameters are obtained, the expected frequencies are easily computed. For applying a χ^2 test of goodness of fit, some last cells are grouped together. The best fitting of the distribution has been given by MLE followed by MPOC and MM method of estimation as revealed through the χ^2 values at one degree of freedom. It seems that the distribution describes the pattern of the data satisfactorily well.

Table 2 Distribution of observed and expected number of Play groups-Eugene, Spring Playground A of 497 frequencies

Group	Observed	Expected frequency		
size	frequency	MPOC	MM	MLE
1	306	306.00	302.82	303.44
2	132	136.74	140.04	137.63
3	47	41.17	41.64	42.17
4	10	13.09	12.50	13.76
5	2			
Total	497	497	497	497
		Estimates of parameters		
	p	0.1000	0.0800	0.1030
	θ	0.2743	0.3287	0.2760
	χ^2	1.0806	1.2316	1.0302
	d.f.	1	1	1

Table 3 provides the observed and expected number of counts of sites with particulars Immungold data of 198 frequencies which was given by Matthews and Appleton (1993). It is important to advocate that $(1+\theta+p)/q$ provides the average of the counts of sites with particulars Immungold data while $\hat{\theta}$ gives the average number of clusters per counts of sites with particulars Immungold data giving $\hat{q}$ as the average number of individuals per counts of sites. It indicates that the higher values of θ and lower values of q provide higher average of the counts. The estimates of parameters p and θ in MPOC, MM and MLE methods of estimation are found to be 0.1500, 0.2200, 0.2210 and 0.1591, 0.0090, 0.0090, respectively. These estimates of the parameters involved in the distribution are somehow varying but taken to be approximately equal. Once the estimates of these parameters are obtained, the expected frequencies are easily found. For applying a χ^2 test of goodness of fit, some last cells are grouped together. The best fitting of the distribution has been given by MLE followed by MM and MPOC method of estimation as revealed through the χ^2 values at one degree of freedom. It seems that the distribution describes the pattern of the data satisfactorily well.

Table 3 Distribution of observed and expected number of counts of sites with particulars Immungold data

Group size	Observed frequency	Expected frequency		
		MPOC	MM	MLE
1	122	122.00	119.37	119.08
2	50	52.50	53.37	53.38
3	18	16.92	17.94	17.98
4	4	6.58	7.32	7.56
5	4			
Total	198	198	198	198
		Estimates of parameters		
	p	0.2000	0.2200	0.2210
	θ	0.0379	0.0091	0.0090
	χ^2	0.4944	0.3341	0.3112
	d.f.	1	1	1

Table 4 describes the distribution of observed and expected number of snowshoe hares captured over 7 years of 261 frequencies which was given by Keith and Meslow (1968). It is to be revealed that $(1+\theta+p)/q$ provides the average of the number of snowshoe hares captured over 7 years while $\hat{\theta}$ gives the average number of clusters per number of times hares caught giving $\hat{q}$ as the average number of individuals per hares caught. It indicates that the higher values of θ and lower values of q provide higher average of the average of the number of snowshoe hares captured over 7 years. The estimates of parameters p and θ in MPOC, MM and MLE method of estimation are found to be 0.2000, 0.2200, 0.2210 and 0.0379, 0.0091, 0.0090 respectively. These estimates of the parameters involved in the distribution are somehow varying but considered to be approximately equal. Once the estimates of these parameters are obtained, the expected frequencies are easily found for applying a χ^2 test of goodness of fit, some last cells are grouped together. The best fitting of the distribution has been given by MLE followed by MPOC and MM method of estimation as revealed through the χ^2 values at one degree of freedom. It seems that the distribution describes the pattern of the data satisfactorily well.

Table 4 Distribution of observed and expected number of snowshoe hares captured over 7 years

Group size	Observed frequency	Expected frequency		
		MPOC	MM	MLE
1	184	184.00	177.48	176.54
2	55	58.31	62.31	62.73
3	14	14.00	16.38	14.41
4	4	4.69	4.83	7.32
5	4			
Total	261	261	261	261
		Estimates of parameters		
	p	0.1500	0.1700	0.1720
	θ	0.0245	0.0130	0.0135
	χ^2	2.5239	3.5234	1.3426
	d.f.	1	1	1

5. Conclusions

This paper presents an analytical model for describing the inherent variation through a new extension of size biased Polya-Aeppli distribution (SBPAD) under some simplified assumptions applicable to various four sets of data. The pattern of four sets of data can be summarized by two parameters in the distribution p and θ assuming the distribution to be SBPAD. It is important to understand that $(1+\theta+p)/q$ provides the average of the play groups-Eugene, spring playground D, play groups-Eugene, spring playground A, counts of sites with particulars Immungold data and the number of snowshoe hares captured over 7 years while $\hat{\theta}$ gives the average number of clusters per groups-Eugene, spring playground D, play groups-Eugene, Spring Playground A, counts of sites with particulars Immungold data and the number of snowshoe hares captured over 7 years giving $\hat{q}$ as the average number of individuals per spring playground D, play groups-Eugene, spring playground A, counts of sites with particulars Immungold data and the number of snowshoe hares captured over 7 years. It indicates that the higher values of θ and lower values of q provide higher average of the spring playground D, Play groups-Eugene, spring playground A, counts of sites with particulars Immungold data and the number of snowshoe hares captured over 7 years. A satisfactory fit is achieved by MPOC, MM and MLE, the three methods of estimations of the parameters. Once the estimates of these parameters are obtained, the expected frequencies are easily obtained. The values of χ^2 do not approach significance at the respective degrees of freedom at 5% level of significance. This has been given by the graphical presentation between observed and expected frequencies in the four sets of data (Appendix). Thus, the observed and fitted distribution has almost the same shape. i.e, differences between the two curves are small and show no consistent features in the distribution. This is an acceptable fit for the derived SBPAD for four sets of data. It suggests that the distribution has successfully described the data. Thus, it may be useful in computing the various probabilities of Play group size, counts, snowshoe hares captured etc. of SBPAD.

Further, research is required to explore the possibility of SBPAD as inflated at the point one i.e., a mixture of some part is attached to the risk exposed to the counts/group size and some are not exposed. Thus, it may be declared as an inflated SBPAD (new distribution) applicable to other kinds of data where the first cell of the observed data might be in the form of inflation with certain probability. It might be also possible to derive the expressions of asymptotic variances and co-variances which might turn into the standard errors of the estimates of the parameters.

Acknowledgements

The authors are indebted to the editor and referees for giving the critical and valuable comments that have greatly helped in re-structuring the paper in the present form.

References

Adhikari TR, Srivastava RS. Size-biased discrete two parameter Poisson-Lindley distribution for modeling and waiting survival times data. Journal of Mathematics. 2014; 10(3): 39-45.

Beghriche A, Zeghdoudi H. A size biased gamma Lindley distribution. Thail Stat. 2019; 17(2): 179-189.

Coleman JS, James J. The equilibrium size distribution of freely forming groups. Sociometry. 1961; 24(1): 36-45.

Ducey MJ, Gove JH. Size-biased distributions in the generalized beta distribution family with applications to forestry. Forestry. 2015; 88(1): 143-151.

Fisher RA. The effect of methods of ascertainment upon the estimation of frequencies. Ann Hum Genet. 1934; 6(1): 13-25.

James J. The distribution of free-forming small group size. Am Socio Rev. 1953; 18(5): 569-570.

Keith LB, Meslow EC. Trap response by snowshoe hares. J Wildl Manag. 1968; 32(4): 795-801.

Matthews JNS, Appleton DR. An application of the truncated Poisson distribution to Immunogold assay. Biometrics. 1993; 49(2): 617-621.

Priti P, Singh BP. A size biased probability distribution for the number of male migrants. J Stat Appl Prob. 2015; 4(3): 411-415.

R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna: Austria; 2018 [cited 2019 Feb 18]. Available from: https://www.r-project.org.

Rao CR. On discrete distributions arising out of methods of ascertainment. Sankhya Ser A. 1965; 27(2/4): 311-324.

Rather AA, Subramanian C, Shafi S, et al. A new size biased distribution with application in Engineering and Medical Science. Int J Sci Res Math Stat Sci. 2018; 5(4): 66-76.

Sampford MR. The truncated negative binomial distribution. Biometrika, 1955; 42(1/2): 58-69.

Shanker R, Shukla KK. Size-biased Poisson-Garima distribution with applications. Biom Biostat Int J. 2017; 6(3): 335-340.

Sium S, Shanker R. Size-biased discrete-Lindley distribution and its applications to model distribution of freely-forming small group size. Biom Biostat Int J. 2018; 7(2): 131-136.

Yadav RK, Kumar U. Application of size-biased geometric distribution to migration data. Int J Stat Prob. 2018; 7(1): 85-90.

Appendices

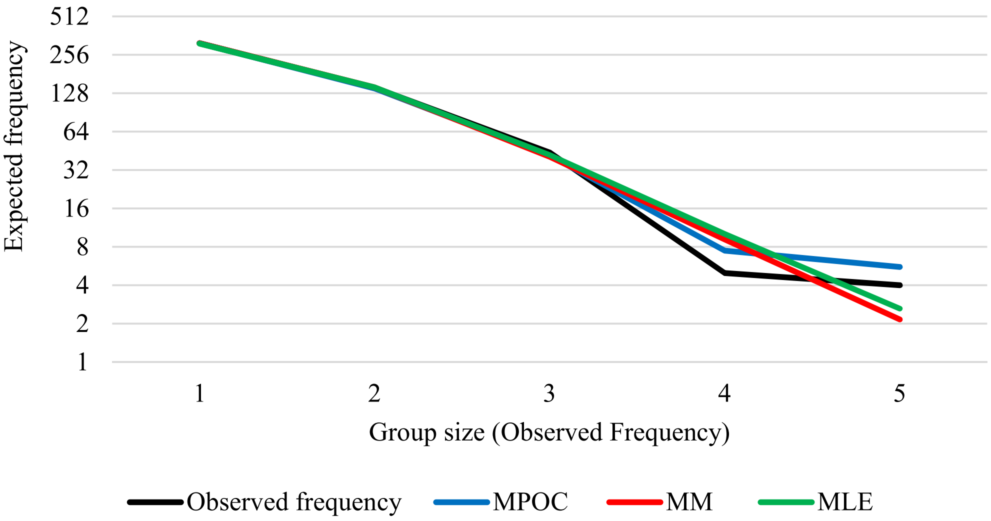

Figure 1 Distribution of observed and expected number of play groups-Eugene, spring playground D of 510 frequencies

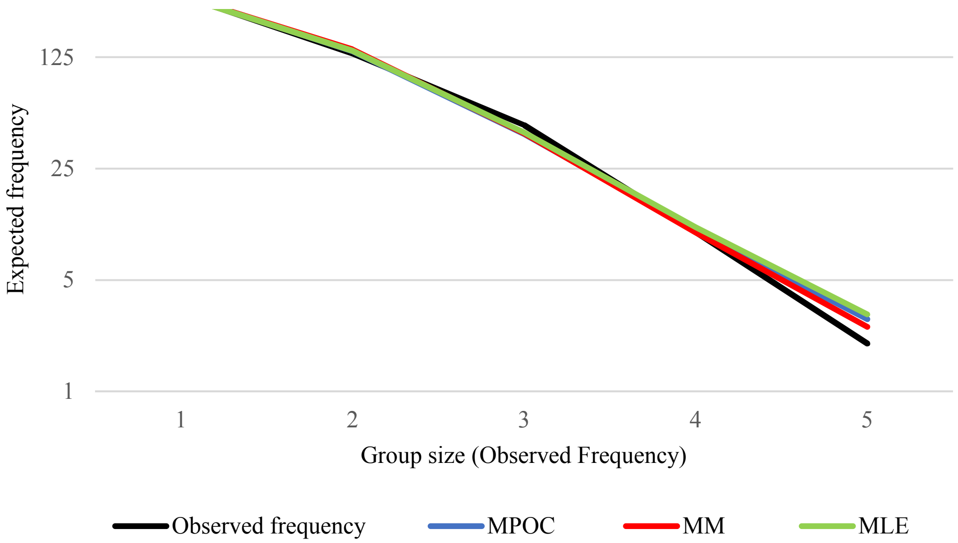

Figure 2 Distribution of observed and expected number of play groups-Eugene, spring playground A of 497 frequencies

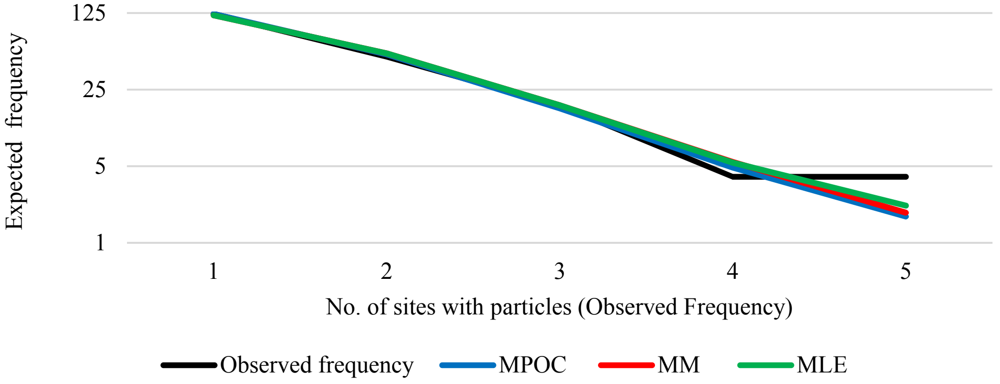

Figure 3 Distribution of observed and expected number of counts of sites with particles Immungold data of 198 frequencies

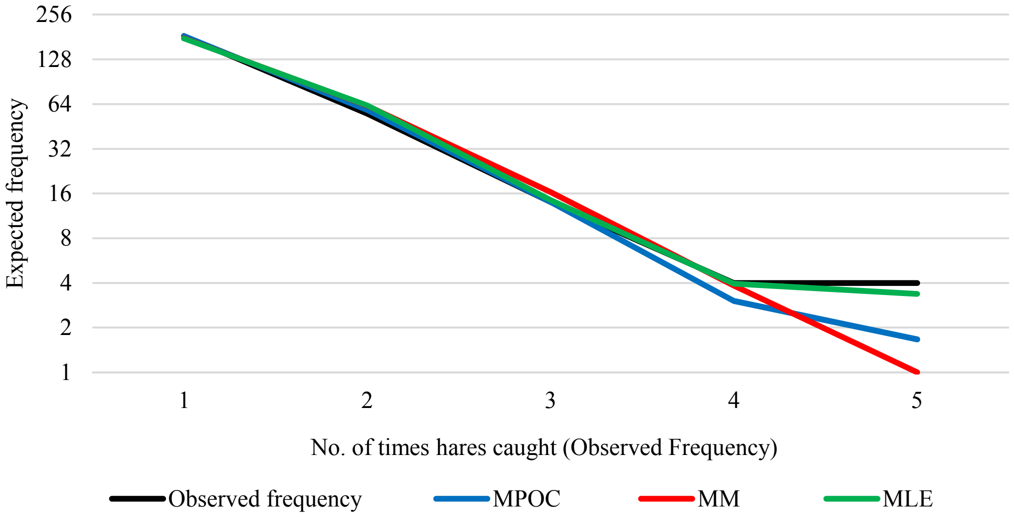

Figure 4 Distribution of observed and expected number of snowshoe hares captured over 7 years