A hyper-Poisson Model for Overdispersed and Underdispersed Count Data

V D Situmorang[1*], S Nurrohmah[1] and I Fithriani[1]

[1] Department of Mathematics, Universitas Indonesia, Depok, 16424, Indonesia

[*] Corresponding author: venda.damianus@sci.ui.ac.id

Abstract. The Poisson model is commonly used for modelling count data. However, it has a limitation, namely the equality between the mean and variance (equidispersion) of the data to be modeled. Unfortunately, overdispersion (variance greater than the mean) and underdispersion (variance smaller than the mean) are more often to be found in real cases. Therefore, different models need to be used to handle data with these cases. The hyper-Poisson model is one model that can be used to handle overdispersion or underdispersion cases flexibly. This paper describes the hyper-Poisson model and its application on overdispersed and underdispersed count data. Insurance policy claims data and red mites' appearance data are used in modelling overdispersed data. Meanwhile, miners' strikes data and pairs of shoes data are used in modelling underdispersed data. By modelling these data using hyper-Poisson and its comparison distribution, it shows that the hyper-Poisson distribution can model overdispersed or underdispersed data flexibly even though there is possibility that there are other distributions that can model it better.

1. Introduction

Count data is counting result data that describes the number of occurrences of an event in a given time period [1]. The value of count data is a non-negative integer because an event cannot occur in a negative number of integers. Count data modelling is widely used in various sciences such as actuarial science, health, demography, transportation, and others.

Counting distributions used to model count data. Counting distributions are discrete distributions with probabilities only on non-negative integers [2]. The Poisson model is a counting distribution which commonly used for modelling count data. However, it is constrained by its equidispersion assumption. In real data, often this assumption is not fulfilled, where the variance is greater than the mean (overdispersion) or the variance is smaller than the mean (underdispersion). Data with overdispersion or underdispersion is not suitable to be modelled with Poisson distribution. Ignoring these circumstances and still using Poisson distribution will cause some possible problems such as incorrect estimations of parameters and standard error, incorrect interpretation of the considered model, and more [3].

In 1964, Bardwell and Crow introduced an alternative that could model data with overdispersion or underdispersion cases flexibly. This model is called the hyper-Poisson model. They formed this distribution through the relationship of the derivative of the Poisson distribution to the differential-difference relation of the Poisson distribution [4]. Then, in 2018, Lesaris took a different approach in forming this distribution. The hyper-Poisson is formed through recursive properties of extended Lagrangian Katz family of distributions [5].

In this paper, the main properties of the hyper-Poisson model are summarized, and its parameters are estimated with method of moments. This method is used due to complexity of hyper-Poisson's parameters, so method of moments will estimate the parameters simpler and easier. In the final section, the model is applied to overdispersion and underdispersion count data. The first two data sets are overdispersed data from vehicle insurance policy claims and red mites' appearance. Meanwhile next two data sets are underdispersed data from frequency of coal miners strikes and number of pairs of shoes. All of data sets will also be modelled with other distribution as comparison to hyper-Poisson. Negative Binomial will be applied to overdispersed data and Binomial will be applied to underdispersed data.

2. Count Data Distribution

Count data distribution is a probability distribution that used in modelling discrete type random variable. Some examples of count data distribution are Poisson distribution, Binomial distribution, and Negative Binomial distribution.

2.1. Poisson Distribution

Suppose a random variable *X* has a probability density function:

$$f(x) = \begin{cases} \dfrac{\theta^x e^{-\theta}}{x!}, & x = 0,1,2,3 \dots \\ 0, & \text{elsewhere} \end{cases} \tag{1}$$

this random variable is said to have a Poisson distribution with parameter θ. Poisson distribution has same mean value and variance value (equidispersion) that is θ, which means that this distribution is suitable for modelling data that meets this equidispersion assumption [2].

2.2. Binomial Distribution

Suppose a random variable *X* has a probability density function:

$$f(x) = \begin{cases} \binom{n}{x} p^x q^{n-x}, & x = 0,1,2, \dots, n \\ 0, & \text{elsewhere} \end{cases} \tag{2}$$

where n is number of Bernoulli trials, x is desired number of successes out of n trials, p is probability of success in a trial, and $q = 1 - p$, then this random variable is said to have a Binomial distribution with parameters n and p. Binomial distribution has mean value np and variance value $np(1-p)$. These values show that Binomial distribution has greater mean value than variance value (underdispersion). So, this distribution is suitable for modelling data that meets this underdispersion assumption [2].

2.3. Negative Binomial Distribution

Suppose a random variable *X* has a probability density function:

$$f(x) = \begin{cases} \binom{k+x-1}{x} p^k q^x, & x = 0,1,2,3, \dots \\ 0, & \text{elsewhere} \end{cases} \tag{3}$$

where k is desired number of successes, x is number of failures before the k-th success, p is probability of success in a trial, and $q = 1 - p$, then this random variable is said to have a Negative Binomial distribution with parameters k and p. Negative Binomial distribution has mean value $\frac{kq}{p}$ and variance value $\frac{kq}{p^2}$. These values show that Negative Binomial distribution has greater variance value that mean value (overdispersion). So, this distribution is suitable for modelling data that meets this overdispersion assumption [6].

3. Neyman-Scott Test

The Neyman-Scott test is a statistical test used to check whether a population follows a Poisson distribution or not. The null hypothesis of this test is the population follows a Poisson distribution, so it can be modelled by Poisson distribution. Meanwhile, its alternative hypothesis is the population does not follow a Poisson distribution, so it has to be modelled by other suitable distribution. This test has test statistics given below:

$$T_{NS} = \sqrt{\frac{n-1}{2}}\left(\frac{S^2}{\bar{X}} - 1\right) \tag{4}$$

where n is number of observations, S^2 is the variance of the data, $\bar{X}$ is the mean of the data, and T_{NS} distribution is approaching normal standard. The null hypothesis of this test is rejected on α level of significance if $|T_{NS}| > \Phi^{-1}(1-\alpha)$ where $\Phi^{-1}(x)$ is notation of standard normal cumulative distribution function at x. For $\alpha = 0.05$, the null hypothesis of this test is rejected if the absolute value of T_{NS} is greater than 1.645.

This test also indicates whether the population is experiencing overdispersion or underdispersion. If $|T_{NS}| > \Phi^{-1}(1-\alpha)$ and $T_{NS} > 0$, then the population is experiencing overdispersion. Otherwise, if $|T_{NS}| > \Phi^{-1}(1-\alpha)$ and $T_{NS} < 0$, then the population is experiencing underdispersion [7].

4. The hyper-Poisson Distribution

The hyper-Poisson distribution has a probability mass function given below:

$$f(x) = \begin{cases} \dfrac{\Gamma(\gamma)\alpha^x}{\Gamma(\gamma+x)M(1;\gamma;\alpha)}, & x = 0,1,2,3\ldots; \alpha > 0; \gamma > 0 \\ 0, & \text{elsewhere} \end{cases} \tag{5}$$

where $\alpha > 0, \gamma > 0$, and

$$M(1;\gamma;\alpha) = \sum_{x=0}^{\infty} \frac{(1)_x}{(\gamma)_x x!}\alpha^x \tag{6}$$

is the confluent hypergeometric series and $(a)_x = a(a+1)\ldots(a+x-1)$ for $a > 0$ and x a positive integer [8]. Main characteristic in this distribution is its parameter γ that shows overdispersed, equidispersed, or underdispersed case. If $\gamma > 1$, it shows overdispersed case, equidispersed case if $\gamma = 1$, and underdispersed case if $0 < \gamma < 1$. That is because γ interprets dispersion or scale parameter, while α interprets shape parameter.

The cumulative distribution function is given by:

$$\frac{\Gamma(\gamma)}{M(1;\gamma;\alpha)} \sum_{w=0}^{x} \frac{\alpha^w}{\Gamma(\gamma+w)} \quad ; \quad w \geq 0, \gamma > 0, \alpha > 0 \tag{7}$$

Also, the probability generating function is given by:

$$G(z) = \frac{M(1;\gamma;\alpha z)}{M(1;\gamma;\alpha)} \tag{8}$$

In order to display dispersion and shape parameters, there will be presented some graphs of the probability density function of the hyper-Poisson distribution. The following graph is a hyper-Poisson distribution density function with $\alpha = 3$ and γ varied.

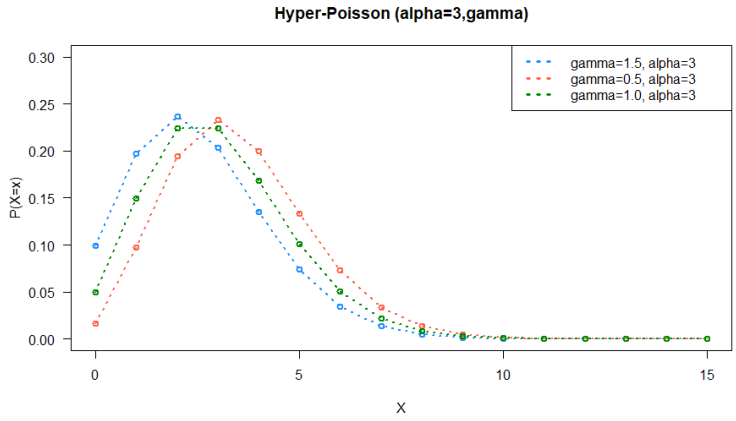

Figure 1. Probability Density Function of hyper-Poisson Distribution with γ Varied.

According to 0, the density function of hyper-Poisson distribution with $\alpha = 3$ has two maximum points for $\gamma = 1$, meanwhile for $\gamma = 0.5$ and $\gamma = 1.5$ only have one maximum point. Also, as the value of the parameter γ increases, the probability density function of the hyper-Poisson distribution will have lighter distribution tail. Next graph below is the hyper-Poisson graph with the values of $\gamma = 0.5$, 1, and 0.5 with α varied, as well as the Poisson distribution graph with parameter α.

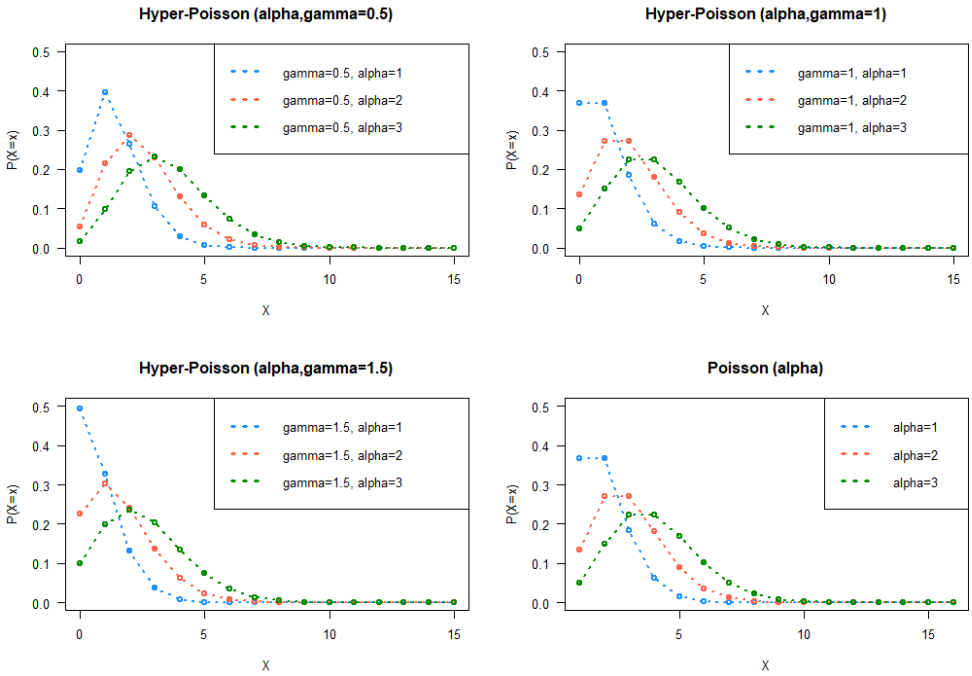

Figure 2. Probability Density Function of hyper-Poisson Distribution with α Varied.

According to 0, the density function of hyper-Poisson distribution with $\gamma = 0.5$ and 1.5 have a peak that decreases as the value of parameter α increases. It can also be seen that for $\gamma = 1$, the probability density function graph has two peaks and decreases as the value of parameter α increases. The graph of the probability density function of the hyper-Poisson distribution with $\gamma = 1$ also has the same graphical shape as the probability density function of the Poisson distribution with the corresponding α. This shows that the hyper-Poisson distribution will become a Poisson distribution with parameter α for $\gamma = 1$.

4.1. Mean and Variance by Probability Generating Function

Mean and variance can be obtained through the properties of the probability generating function, that is:

$$E(X) = G'(1)$$
$$Var(X) = G''(1) + G'(1) - [G'(1)]^2$$

First, the probability generating function will be derived twice to obtain $G'(z)$ and $G''(z)$:

$$G'(z) = \frac{\alpha}{\gamma}\frac{M(2;\gamma+1;az)}{M(1;\gamma;\alpha)}$$
$$G''(z) = \frac{2\alpha^2}{\gamma(\gamma+1)}\frac{M(3;\gamma+2;az)}{M(1;\gamma;\alpha)}$$

Second, by substituting $z = 1$ and supposing $\Lambda_k = \frac{M(1+k;\gamma+k;\alpha)}{M(1;\gamma;\alpha)}$, mean and variance will be obtained:

$$E(X) = \frac{\alpha}{\gamma}\Lambda_1 \tag{9}$$

$$Var(X) = \frac{\alpha}{\gamma}\Lambda_1 + \frac{\alpha^2}{\gamma}\left(\frac{2}{\gamma+1}\Lambda_2 - \frac{1}{\gamma}\Lambda_1^2\right) \tag{10}$$

4.2. Mean and Variance by Recurrence Relation

If the expression of mean and variance by probability generating function is considered complicated, there is another way to obtain mean and variance by recurrence relation of the hyper-Poisson distribution. It can be proved that equation 0 verifies the recurrence equation:

$$(\gamma + x)f(x+1) = \alpha f(x) \tag{11}$$

From this equation, multiplying both members by $(x+1)^k$ and adding on x will obtain moment equation.

$$(\gamma - 1)E\left(X^k\right) + E\left(X^{k+1}\right) = \alpha\sum_{i=0}^{k}\binom{k}{i}E\left(X^i\right) \tag{12}$$

If $k = 1$, there will be an equation that contains first moment (mean) and second moment:

$$E(X^2) = \alpha + \left(\alpha - (\gamma - 1)\right)E(X) \tag{13}$$

Adding equation 0 on x will obtain value of $E(X)$ (mean).

$$E(X) = \alpha - (\gamma - 1)\left(1 - f(0)\right) \tag{14}$$

Because value of $E(X)$ and $E(X^2)$ are already known, value of variance can be obtained:

$$Var(X) = \alpha + \left(\alpha - (\gamma - 1)\right)E(X) - E(X)^2 \tag{15}$$

5. Parameter Estimation of hyper-Poisson Distribution

Parameter estimation of the hyper-Poisson distribution is estimated by the method of moments because the hyper-Poisson distribution has a gamma function in it. With this gamma function, the *maximum likelihood* method will be very difficult to apply. Therefore, the method of moments is chosen in estimating the parameters. Method of moments is a method that equates population moment $E(X^k)$ and sample moment M_k starting from $k = 1$ until there are enough equations to find parameter estimates [9].

Moments of hyper-Poisson distribution obtained from equation 0. By substituting $k = 1$ and $k = 2$, second moment and third moment will be obtained while first moment has obtained from equation 0. By method of moments, the following equation is obtained:

$$M_1 = E(X)$$
$$M_2 = E(X^2)$$
$$M_3 = E(X^3)$$

where M_1, M_2, and M_3 are first, second, and third sample moment of a hyper-Poisson distributed random sample. By solving these equations, estimated parameter of hyper-Poisson distribution will be obtained as follows:

$$\hat{\alpha} = \frac{M_1 M_3 - M_2^2}{2M_1^2 + M_1 - M_2} \tag{16}$$

$$\hat{\gamma} = 1 + \frac{M_1 M_3 + M_3 - M_2^2 - M_2 - 2M_1 M_2}{2M_1^2 + M_1 - M_2} \tag{17}$$

6. Applications

In this section, the hyper-Poisson distribution is used to model overdispersed and underdispersed count data. Other distribution will also be used as comparison distribution.

6.1. Overdispersed Count Data

6.1.1. Vehicle Insurance Policy Claims Data. The data set that used in this subsection is vehicle insurance policy claims of a Turkish insurance company occurred between 2012 and 2014. The data contains six categories of claim frequency and 10814 observations (policyholders) [10].

Table 1. Vehicle Insurance Policy Claims

Claim Frequency	Observed Values
0	8544
1	1796
2	370
3	81
4	22
5	1

This data set has mean value 0.265582 and variance value 0.334681. Greater variance value shows that this is an overdispersed data set. This overdispersed claim can also be proven by Neyman-Scott test. The test statistic of Neyman-Scott T_{NS} for this data is 19.13072. This shows that this data is experiencing overdispersion. Thus, this data will be modelled with hyper-Poisson distribution and Negative Binomial distribution as comparison distribution. The method of moments was used to estimate each distribution's parameter.

Table 2. Result Comparison between hyper-Poisson and Negative Binomial

Claim Frequency	Observed Values	Expected Values	
		hyper-Poisson	Negative Binomial
0	8544	8540.594408	8540.205104
1	1796	1798.714248	1799.838904
2	370	376.240688	375.46208
3	81	78.163592	78.055452
≥ 4	23	19.443572	19.562526
Method of Moments Estimation		$\hat{\alpha} = 30.645328$ $\hat{\gamma} = 145.508893$	$\hat{k} = 1.020761$ $\hat{p} = 0.793537$
Degree of freedom		2	2
$\chi^2_{0.05;\mathrm{df}}$		5.991	5.991
χ^2		0.862403	0.804438

Mean	0.265583	0.265582
Variance	0.3346	0.334682

Through the chi-square value, it can be seen that the hyper-Poisson distribution has a chi-square value of 0.862403 and the Negative Binomial distribution has a chi-square value of 0.804438. This result also shows small difference of mean and variance values from model and data. From this comparison, it can be concluded that Negative Binomial is slightly better to model the data than hyper-Poisson, but both distributions are still suitable in modelling the data.

6.1.2. Count of Red Mites on Apple Leaves Data. The second data set that used for modelling overdispersion is observed red mites on apple leaves. The data contains eight categories of appearance frequency and 150 observations [11].

Table 3. Red Mites on Apple Leaves

Appearance Frequency	Observed Values
0	70
1	38
2	17
3	10
4	9
5	3
6	2
7	1

This set has mean value 1.146667 and variance value 2.258488. Greater variance value shows that this is an overdispersed data set. This overdispersed claim can also be proven by Neyman-Scott test. The test statistic of Neyman-Scott T_{NS} for this data is 8.369041. This shows that this data is experiencing overdispersion. Thus, this data will be modelled with hyper-Poisson distribution and Negative Binomial distribution as comparison distribution. The method of moments was used to estimate each distribution's parameter.

Table 4. Result Comparison between hyper-Poisson and Negative Binomial

		Expected Values	
Claim Frequency	Observed Values	hyper-Poisson	Negative Binomial
0	70	67.4697	67.2909
1	38	38.3436	39.17535
2	17	21.1521	21.0462
3	10	11.33595	10.9914
4	9	5.907	5.65785
≥ 5	6	5.1855	5.0913
Method of Moments Estimation		$\hat{\alpha} = 18.814195$ $\hat{\gamma} = 33.10564$	$\hat{k} = 1.182605$ $\hat{p} = 0.507715$
Degree of freedom		3	3
$\chi^2_{0.05;\mathrm{df}}$		7.815	7.815
χ^2		2.817941	3.148075
Mean		1.149612	1.146664
Variance		2.212575	2.25848

Through the chi-square value, it can be seen that the hyper-Poisson distribution has a chi-square value of 2.817941 and the Negative Binomial distribution has a chi-square value of 3.148075. This result

shows that hyper-Poisson is better to model the data than Negative Binomial, but both distributions are still suitable in modelling the data.

6.2. Underdispersed Count Data

6.2.1. Coal Miners' Strikes Data. The data set that used in this subsection is frequency of coal miners strikes over a 4-weeks period in UK. The data contains five categories of strike frequency and 156 observations [12].

Table 5. Coal Miner's Strike

Strike Frequency	Observed Values
0	46
1	76
2	24
3	9
4	1

This data set has mean value 0.99359 and variance value 0.737138. Greater mean value shows that this is an underdispersed data set. Just like overdispersed data before, this underdispersed claim can also be proven by Neyman-Scott test [7]. The test statistic of Neyman-Scott T_{NS} for this data is -2.27222. This shows that this data is experiencing underdispersion. Thus, this data will be modelled with hyper-Poisson distribution and Binomial distribution as comparison distribution. The method of moments was used to estimate each distribution's parameter.

Table 6. Result Comparison between hyper-Poisson and Binomial

Claim Frequency	Observed Values	Expected Values	
		hyper-Poisson	Binomial
0	46	48.281220	49.430940
1	76	71.148636	66.201096
2	24	28.642692	32.814600
≥ 3	10	7.773012	7.558356
Method of Moments Estimation		$\hat{\alpha} = 0.553889$ $\hat{\gamma} = 0.375867$	$\hat{n} = 3.849534$ $\hat{p} = 0.258107$
Degree of freedom		1	1
$\chi^2_{0.05;\mathrm{df}}$		3.841	3.841
χ^2		1.829153	4.845053
Mean		0.984856	Model does not fit
Variance		0.74413	Model does not fit

Through the chi-square value, it can be seen that the hyper-Poisson distribution has a chi-square value of 1.829153 and the Binomial distribution has a chi-square value of 4.845053. From this comparison of chi-square values, it can be concluded that hyper-Poisson has much better chi-square values than Binomial. With chi-square values that exceed 3.841, Binomial distribution is not fit to model this data.

6.2.2. Pairs of Shoes Data. The data set that used in this subsection is the number of pairs of running shoes owned by running club members. The data contains five categories of number of pairs and 60 observations [13].

Table 7. Number of Pairs of Running Shoes

Number of Pairs	Observed Values
1	18

2	18
3	12
4	7
5	5

This data set has mean value 2.383333 and variance value 1.569724. Greater mean value shows that this is an underdispersed data set. Just like overdispersed data before, this underdispersed claim can also be proven by Neyman-Scott test. The test statistic of Neyman-Scott T_{NS} for this data is -1.85414. This shows that this data is experiencing underdispersion. Thus, this data will be modelled with hyper-Poisson distribution and Binomial distribution as comparison distribution. The method of moments was used to estimate each distribution's parameter.

Table 8. Result Comparison between hyper-Poisson and Binomial

Claim Frequency	Observed Values	Expected Values	
		hyper-Poisson	Binomial
1	18	14.2512	11.7627
2	18	20.4459	18.23418
3	12	14.70414	15.69372
4	7	7.05594	8.09682
5	5	2.54046	2.50254
Method of Moments Estimation		$\hat{\alpha} = 1.442039$ $\hat{\gamma} = 0.005132$	$\hat{n} = 6.981578$ $\hat{p} = 0.341375$
Degree of freedom		2	2
$\chi^2_{0.05;\mathrm{df}}$		5.991	5.991
χ^2		4.157667	6.820738
Mean		2.436066	Model does not fit
Variance		1.444088	Model does not fit

Through the chi-square value, it can be seen that the hyper-Poisson distribution has a chi-square value of 4.157667 and the Binomial distribution has a chi-square value of 6.820738. From this comparison of chi-square values, it can be concluded that hyper-Poisson has much better chi-square values than Binomial. With chi-square values that exceed 5.991, Binomial distribution is not fit to model this data.

6.3. Conclusion

In the previous section, modelling overdispersed and underdispersed data with the hyper-Poisson distribution and with its comparison distributions has been done. In the case of overdispersion, it can be seen that both distributions are equally good at modelling data as seen from small difference in chi-square values. For the first data, it can be seen that Negative Binomial is slightly better to model the data than hyper-Poisson, but both distributions are still suitable in modelling the data. But, for the second data, it can be seen that hyper-Poisson is better to model the data. Meanwhile, in the case of underdispersion, the hyper-Poisson distribution is better at modelling both data compared to Binomial distribution. So, from these data illustrations, it can be concluded that the hyper-Poisson distribution can model overdispersed or underdispersed data even though there is possibility that there are other distributions that can model it better.

7. Discussion

The proposed model that is hyper-Poisson appears to be an enticing alternative to model overdispersed and underdispersed count data. It has shown that hyper-Poisson distribution is a flexible distribution to model data with different circumstances such as overdispersed, equidispersed, and underdispersed.

This paper also shows some of hyper-Poisson distribution properties such as probability generating function, mean, and variance. Parameter γ in this distribution interprets dispersion parameter, thus γ can

show whether the data is overdispersed, equidispersed, or underdispersed. It is also possible to show that hyper-Poisson with $\gamma = 1$ is a Poisson distribution with the corresponding α.

Finally, hyper-Poisson is applied to model overdispersed and underdispersed data. From the data illustrations, they show that hyper-Poisson can model overdispersed and underdispersed data well even though there could be better distribution to model the data. Nevertheless, hyper-Poisson provides an advantage to model the data flexibly with only one distribution.

Acknowledgments

The author is grateful for the support and constructive suggestions provided by the supervisors, which improved the paper.

References

[1] S. Coxe, S. G. West, and L. S. Aiken, "The analysis of count data: A gentle introduction to poisson regression and its alternatives," *Journal of Personality Assessment*, vol. 91, no. 2. Routledge, pp. 121–136, 2009. doi: 10.1080/00223890802634175.

[2] S. A. Klugman, H. H. Panjer, and G. E. Willmot, *Loss Models: from Data to Decisions*, 5th ed. John Wiley & Sons, Inc., 2019. [Online]. Available: www.wiley.com

[3] C. C. Kokonendji, "Over-and Underdispersion Models," 2014.

[4] G. E. Bardwell and E. L. Crow, "A Two-Parameter Family of Hyper-Poisson Distributions," 1964.

[5] S. Mareyan Lesaris, "DISTRIBUTIONS ARISING FROM BIRTH AND DEATH PROCESSES AT EQUILIBRIUM AND THEIR EXTENSIONS," *Research Report in Mathematics*, vol. 18, 2018.

[6] R. Fewster, "Stochastic Processes." 2014.

[7] L. D. Brown and L. H. Zhao, "A new test for the Poisson distribution," 2001.

[8] M. A. Chaudhry, A. Qadir, H. M. Srivastava, and R. B. Paris, "Extended hypergeometric and confluent hypergeometric functions," *Appl Math Comput*, vol. 159, no. 2, pp. 589–602, Dec. 2004, doi: 10.1016/j.amc.2003.09.017.

[9] R. V. Hogg and A. T. Craig, *Introduction to Mathematical Statistics*, 5th ed. Pearson Education, 1995.

[10] L. S. Sarul, "AN APPLICATION OF CLAIM FREQUENCY DATA USING ZERO INFLATED AND HURDLE MODELS IN GENERAL INSURANCE," *Pressacademia*, vol. 4, no. 4, pp. 732–732, Dec. 2015, doi: 10.17261/pressacademia.2015414539.

[11] C. S. Kumar and B. U. Nair, "A Three Parameter hyper-Poisson Distribution and Some of Its Properties," *Statistica*, vol. 2, 2014.

[12] M. S. Ridout and P. Besbeas, "An empirical model for underdispersed count data," *Stat Modelling*, vol. 4, pp. 77–89, 2004, [Online]. Available: http://stat.uibk.ac.at/SMIJ

[13] J. S. Simonoff, *Analyzing Categorical Data*. New York, NY: Springer New York, 2003. doi: 10.1007/978-0-387-21727-7.