MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Tajuddin and Ismail | Malaysian Journal of Fundamental and Applied Sciences, Vol. 19 (2023) 164-172

RESEARCH ARTICLE

A New One-Parameter Size-Biased Poisson Distribution for Modelling Underdispersed Count Data

Razik Ridzuan Mohd Tajuddin*, Noriszura Ismail

Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

Abstract This paper proposes a new one-parameter discrete distribution for positive count data, named underdispersed size-biased Poisson distribution, as an alternative to modeling underdispersed positive count data. Several properties and measures are presented, such as moments about origins, variance, skewness, kurtosis, index of dispersion, coefficient of variation, and recurrence relationship. Estimators are also developed based on two estimation techniques, i.e., maximum likelihood and moment method. It was found that both estimation techniques yield an identical estimator, which is unique, positively biased, consistent, and asymptotically normal. Finally, a dataset is fitted to the proposed distribution to verify the ability of the proposed distribution to explain the real dataset with a comparison to two known size-biased distributions.

Keywords: Double size-biased poisson, two-component mixture distribution, underdispersion.

Introduction

When modelling positive count data, the heterogeneity in the data must be taken into account by considering a mixed distribution with truncation or those with similar effects as truncation, such as weighted distributions. Several examples of weighted distributions with different weights, for which a special case is known as the probability proportional to the size that involves the weight to be proportional to the size of observations, were introduced [1]. This type of weighted distribution is known as the size-biased distribution [2]. Modelling data using size-biased distributions is a common practice because of the nature of the collected samples due to the following three reasons – non-observability of events, partial destruction of observations, and sampling with unequal chances of observations [1]. Generally, when the events are unobserved, the data will be truncated. Furthermore, some data, especially the ones produced by nature, may also be destroyed. Besides that, targeting a specific event and tracing back its actual observations in the population may not give an equal chance for the event to occur in the population [1]. These reasons cause deformity in the collected sample data. Therefore, it is reasonable to consider size-biased distributions for modelling this type of data. Several examples of size-biased distributions include the size-biased Poisson, the size-biased binomial, the size-biased negative binomial distributions [2], and the size-biased Poisson-Lindley distribution [3]. A size-biased distribution can be written as $h(x) = xf(x)/\mu$, where $f(x)$ is an unweighted distribution and μ is the mean of the unweighted distribution.

Another way of handling count data with heterogeneity is by employing finite mixture models, as they are very flexible [4]. The mixing proportion and the dispersion of the two components representing the two subpopulations affect the model fitting of overall data [5]. The finite mixture models are used in modelling multimodal data in the area of insurance claims [6] and weather spells [7-8]. A two-component mixture distribution can be written as $h(x) = pf_1(x) + (1-p)f_2(x)$, where p is the mixing distribution for the two $f_i(x)$ distributions for $i = 1, 2$.

An attempt has been made in this paper to develop and explore a new flexible underdispersed (variance less than mean) distribution by considering both weighted distribution and two-component mixture distribution approaches. By combining both approaches, the new distribution has increased flexibility,

***For correspondence:**
rrmt@ukm.edu.my

Received: 15 July 2022
Accepted: 15 Feb. 2023

10.11113/mjfas.v19n2.2678

yet with a simple form, which facilitates the fitting of the model. The new distribution, which is a one-parameter distribution based on the two-component of the size-biased Poisson distribution, is believed to provide an adequate fit for positive count data with underdispersion property.

The paper is organized as follows. Section 2 discusses the probability mass function of the proposed distribution with its moments and moment-related measures. In Section 3, the maximum likelihood estimator and the moment estimator for the parameter of the proposed distribution, as well as their properties, are discussed. In Section 4, the proposed distribution is fitted to a dataset with a comparison to two other size-biased distributions. Finally, section 5 concludes the study and gives recommendations for future studies.

The Proposed Distribution

Probability Mass Function

Let X be a random variable that follows the proposed distribution, which we denote as an underdispersed size-biased Poisson (USBP) distribution with parameter $\lambda > 0$. The probability mass function (pmf) of USBP distribution is given as:

$$\Pr(X = x) = \frac{x\lambda^{x-1}\exp(-\lambda)}{(\lambda+1)(x-1)!}, \quad (1)$$

for $x = 1,2,\ldots$. The cumulative function, $F(x)$ for X is given as

$$F(x) = \frac{1}{(x-1)!}\left[\Gamma(x,\lambda) - \frac{\lambda^x \exp(-\lambda)}{\lambda+1}\right],$$

where $\Gamma(x,\lambda)$ is the upper incomplete gamma function. Figure 1 below shows the pmf plot of USBP distribution for different values of λ. The pmf plots suggest that the USBP distribution may provide a good fit to the data with mode and mean greater than one.

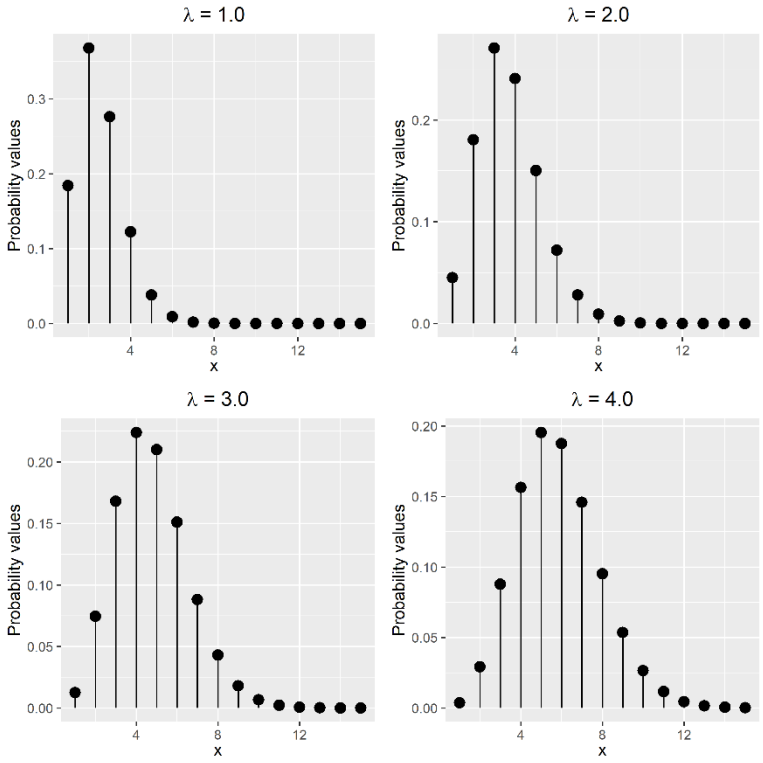

Figure 1. The pmf Plot of USBP Distribution for $\lambda = 1.0, 2.0, 3.0, 4.0$.

Note that USBP distribution can also be written as a mixture of two types of underdispersed size-biased Poisson distributions, which can be written as $\Pr(X = x) = \theta f(x) + (1-\theta)g(x)$, where:

$$f(x) = \frac{\lambda^{x-1}\exp(-\lambda)}{(x-1)!}, g(x) = \frac{\lambda^{x-2}\exp(-\lambda)}{(x-2)!}, \text{and } \theta = \frac{1}{\lambda+1}.$$

Contrary to the usual two-component mixture distributions, the two components of the USBP distribution have the same parameter λ, and the mixing proportion takes on a fixed form. Also note that the USBP distribution can also be obtained by considering size-biased of the size-biased Poisson distribution, which we can name as double size-biased Poisson distribution. The pmf of the size-biased Poisson (SBP) distribution can be written as:

$$f(x) = \frac{\lambda^{x-1}\exp(-\lambda)}{(x-1)!}; x = 1, 2, 3, ...,$$

with $\mu = \lambda + 1$, and thus, the USBP distribution can be derived as:

$$\Pr(X = x) = \frac{x}{\mu} f(x) = \frac{x}{\lambda+1}\left[\frac{\lambda^{x-1}\exp(-\lambda)}{(x-1)!}\right] = \frac{x\lambda^{x-1}\exp(-\lambda)}{(\lambda+1)(x-1)!}; x = 1, 2, 3, ...$$

Furthermore, the probability for each value of x can be obtained using the following recurrence relation

$$\Pr(X = x+1) = \lambda\left(1 + \frac{1}{x}\right)\Pr(X = x) \text{ with } \Pr(X = 1) = \frac{\exp(-\lambda)}{\lambda+1}.$$

To find the mode of the distribution, one can differentiate the log pmf of the USBP distribution and set it to zero. The differentiated log pmf, t is given by:

$$t = \frac{d\ln\Pr(X = x)}{dx} = \frac{1}{x} + \ln\lambda - \frac{\Gamma'(x)}{\Gamma(x)} = \frac{1}{x} + \ln\lambda - \psi(x),$$

where $\Gamma(x) = (x-1)!$ is the gamma function for x and $\psi(x)$ is the digamma function for x. The $\psi(x)$ function can be approximated with $\ln x - (2x)^{-1}$ [9]. The t function with the approximated $\psi(x)$ is set to zero, yielding the mode, x_m

$$x_m \approx \lambda\exp\left[W\left(\frac{3}{2\lambda}\right)\right],$$

where $W(\cdot)$ is the Lambert W function, and the integer value is close to the value of x_m is the mode of USBP distribution. Note that $x_m \geq 1$.

Moments and Some Related Measures

The r^{th} moments about the origin of USBP distribution can be obtained using:

$$\mu'_r = E(X^r) = \sum_{x=1}^{\infty} \frac{x^{r+1}\lambda^{x-1}\exp(-\lambda)}{(\lambda+1)(x-1)!}.$$

The first four moments about the origin are given respectively as:

$$\mu'_1 = \mu = \frac{\lambda^2 + 3\lambda + 1}{\lambda+1},$$

$$\mu'_2 = \frac{\lambda^3 + +\lambda^2 + 7\lambda + 1}{\lambda+1},$$

$$\mu'_3 = \frac{\lambda^4 + 10\lambda^3 + 25\lambda^2 + 15\lambda + 1}{\lambda+1},$$

$$\mu'_4 = \frac{\lambda^5 + 15\lambda^4 + 65\lambda^3 + 90\lambda^2 + 31\lambda + 1}{\lambda+1}.$$

With these four moments, some measures such as variance, skewness, kurtosis, index of dispersion (IOD), and coefficient of variation (CV) can be obtained. The formulae for the variance, the index of dispersion, and the coefficient of variation are respectively given as:

$$\sigma^2 = \frac{\lambda(\lambda^2 + 2\lambda + 2)}{(\lambda + 1)^2},$$

$$IOD = \frac{\sigma^2}{\mu} = \left[1 - \frac{1}{\lambda + 1}\right]\left[1 - \frac{\lambda - 1}{\lambda^2 + 3\lambda + 1}\right] < 1,$$

$$CV = \frac{\sigma}{\mu} = \frac{\sqrt{\lambda(\lambda^2 + 2\lambda + 2)}}{\lambda^2 + 3\lambda + 1}.$$

Based on the IOD formula, it can be observed that the USBP will always be underdispersed for $\lambda > 0$. Table 1 shows the trend of some measures as λ increases. From Table 1, it is clear that as λ increases, the mean, the variance, and the IOD increase. However, the IOD can only increase up to less than one, showing the underdispersion property of the USBP distribution. On the other hand, the skewness values are positive and decrease as λ increases, which suggests that the USBP is skewed to the right. Interestingly, the kurtosis values show a J-shaped curve, and for higher values of λ, this suggests that the USBP distribution has a heavy tail. Unlike kurtosis, the CV shows an inverse J-shaped curve, and for higher values of λ, this suggests that the USBP distribution is concentrated more around the mean.

Table 1. Moment-related measures of USBP distribution for different values of λ

Measures ↓ λ →	0.1	0.3	0.5	1.0	2.0	5.0	10.0
Mean	1.191	1.531	1.833	2.500	3.667	6.833	11.909
Variance	0.183	0.478	0.722	1.250	2.222	5.139	10.083
Skewness	2.147	1.199	0.935	0.716	0.581	0.421	0.310
Kurtosis	136.947	94.349	96.959	121.960	181.590	393.915	896.821
IOD	0.153	0.312	0.394	0.500	0.606	0.752	0.847
CV	0.359	0.451	0.464	0.447	0.407	0.332	0.267

Some Generating Functions

Several generating functions, such as moment generating function, probability generating function, and cumulant generating functions, are given respectively as:

$$M_X(t) = \frac{\exp(-\lambda)}{\lambda + 1}\exp(\lambda e^t + t)\,[\lambda \exp(t) + 1],$$

$$G_X(t) = M_X(\ln t) = \frac{\lambda t^2 \exp[-\lambda(1 - t)]}{\lambda + 1},$$

$$C_X(t) = \ln M_X(t) = \ln[\lambda \exp(t) + 1] - \ln(\lambda + 1) + \lambda \exp(t) - \lambda + t.$$

Estimation Methods

In this section, the parameter λ of the USBP distribution is estimated using the maximum likelihood and moment estimation techniques.

Maximum Likelihood

The maximum likelihood estimator (MLE) of λ is obtained by maximizing the likelihood function:

$$L(\lambda) = \prod_{x=1}^{\infty}[\Pr(X = x)]^{n_x},$$

where n_x is the frequency for x-valued data, or equivalently, by maximizing the log-likelihood function:

$$l = \ln L(\lambda) = \sum_{x=1}^{\infty} n_x \ln \Pr(X = x) \propto \sum_{x=1}^{\infty} n_x[(x - 1)\ln\lambda - \lambda - \ln(\lambda + 1)] = A \ln\lambda - n\lambda - n\ln(\lambda + 1),$$

10.11113/mjfas.v19n2.2678

where $A = \sum_{x=1}^{\infty} n_x(x-1) = n(\bar{x}-1), n = \sum_{x=1}^{\infty} n_x$ and $\bar{x}$ is the sample mean. Differentiating the log-likelihood function above and equating it to zero will result in a quadratic equation given as:

$$n\hat{\lambda}^2 + (2n - A)\hat{\lambda} - A = 0,$$

which can be solved using the quadratic formula given by:

$$\hat{\lambda} = \frac{-(2n-A) + \sqrt{4n^2 + A^2}}{2n}.$$

However, $-(2n - A) = -n(3 - \bar{x})$ and $4n^2 + A^2 = n^2[4 + (\bar{x} - 1)^2] = n^2(\bar{x}^2 - 2\bar{x} + 5)$. Therefore,

$$\hat{\lambda} = \frac{-(3-\bar{x}) + \sqrt{\bar{x}^2 - 2\bar{x} + 5}}{2}. \quad (2)$$

Moment Estimator

The moment estimator (ME) of λ can be obtained by equating the sample mean with the theoretical mean as given by:

$$\bar{x} = \frac{\tilde{\lambda}^2 + 3\tilde{\lambda} + 1}{\tilde{\lambda} + 1},$$

where $\bar{x}$ is the sample mean and $\tilde{\lambda}$ is the ME of λ. The above equation yields a quadratic equation, given as:

$$\tilde{\lambda}^2 + (3 - \bar{x})\tilde{\lambda} + (1 - \bar{x}) = 0,$$

which can be solved using a quadratic formula given by:

$$\tilde{\lambda} = \frac{-(3-\bar{x}) + \sqrt{\bar{x}^2 - 2\bar{x} + 5}}{2},$$

identical to $\hat{\lambda}$ in (2). Onwards, we use $\hat{\lambda}$ as the estimator of λ.

Some Properties of $\hat{\lambda}$

Theorem 1. *The estimate $\hat{\lambda}$ is a unique estimator for λ.*
Proof
The quadratic equation yields $\hat{\lambda}$ can produce two real roots since the discriminant $d = \bar{x}^2 - 2\bar{x} + 5 > 0$ with a minimum value at $d = 4$. However, $-(3 - \bar{x}) < \sqrt{\bar{x}^2 - 2\bar{x} + 5}$ and since $\lambda > 0$, therefore, there is only one acceptable solution to the quadratic equation.

Theorem 2. *The estimate $\hat{\lambda}$ is positively biased.*
Proof
Let $\hat{\lambda} = g(\bar{x})$ where $g(t) = \left[-(3-t) + \sqrt{t^2 - 2t + 5}\right]/2, t > 0$.
Then,

$$g''(t) = \frac{2}{(t^2 - 2t + 5)^{3/2}} > 0.$$

Therefore, $g(t)$ is strictly convex. By Jensen's Inequality, we know that $E[g(\bar{X})] > g(E[\bar{X}])$. Since

$$g(E[\bar{X}]) = g(\mu) = g\left(\frac{\lambda^2 + 3\lambda + 1}{\lambda + 1}\right) = \frac{\lambda(\lambda+2)^2}{2(\lambda+1)} > \lambda,$$

hence, $E[\hat{\lambda}] > \lambda$.

Theorem 3. *The estimate $\hat{\lambda}$ is consistent and asymptotically normal:*

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{d} N(0, I^{-1}(\lambda)),$$

10.11113/mjfas.v19n2.2678

where

$$I(\lambda) = \frac{\lambda^3 + 3\lambda^2 + 4\lambda + 1}{\lambda^2(\lambda+1)^2},$$

is the Fisher information about λ.

Proof

The USBP distribution satisfies the regularity conditions under which the estimator $\hat{\lambda}$ is consistent and asymptotically normal (see [10, chapter 6]). Therefore:

$$I(\lambda) = E\left(-\frac{d^2 \ln \Pr(X=x)}{d\lambda^2}\right) = E\left(\frac{x-1}{\lambda^2} - \frac{1}{(\lambda+1)^2}\right) = \frac{\lambda^3 + 3\lambda^2 + 4\lambda + 1}{\lambda^2(\lambda+1)}.$$

Theorem 3 implies that the asymptotic $100(1-\alpha)\%$ confidence interval for λ is $\hat{\lambda} \mp z_{\alpha/2} \frac{I^{-1/2}(\hat{\lambda})}{\sqrt{n}}$.

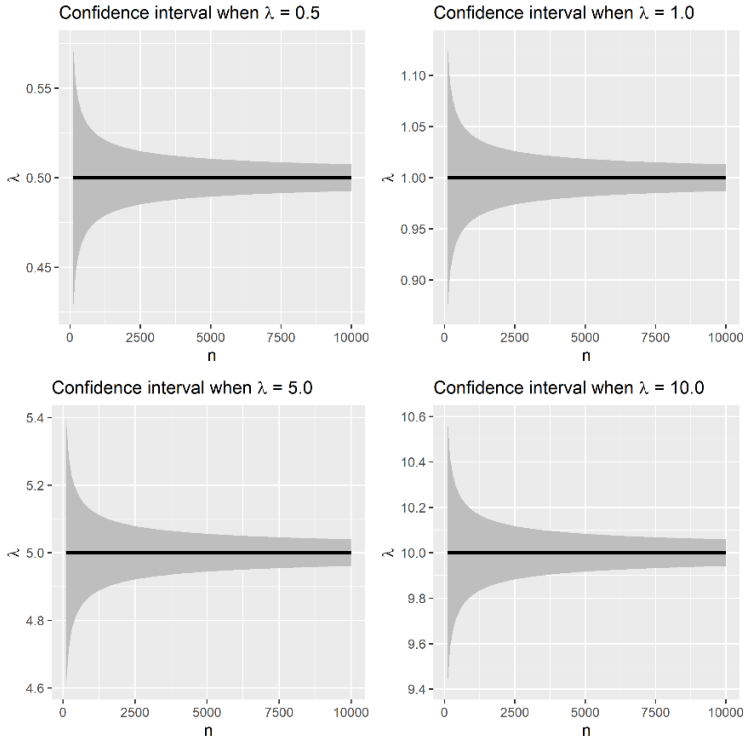

Figure 2. The 95% confidence interval of λ for $\lambda = 0.5, 1.0, 5.0, 10.0$.

For illustration purposes, the 95% confidence interval region based on $I^{-1/2}(\lambda)/\sqrt{n}$ for $\lambda = 0.5, 1.0, 5.0, \text{and } 10.0$ are shaded and given in Figure 2. It is clear from Figure 2 that as n increases, the confidence interval band becomes narrower. Also, as λ increases, the confidence interval band becomes wider.

Simulation Study

Since both moment and maximum likelihood estimators yield identical estimators, a simple simulation study is conducted to investigate the unbiasedness and consistent properties. For this simulation study, λ is set to 1, 2, 3, and 4, whereas the sample size, n is set to $200\ (200)\ 1000$. Each set of simulations is replicated 2000 times. To measure the unbiasedness and consistency properties of the estimator, the mean absolute deviation, MAD and the root-mean-squared error values, $RMSE$ are used. The formulae for MAD and $RMSE$ are respectively given as:

$$MAD = \frac{1}{2000}\sum_{i=1}^{2000}\left|\hat{\lambda}_i - \lambda\right|,$$

$$RMSE = \sqrt{\frac{1}{2000}\sum_{i=1}^{2000}\left(\hat{\lambda}_i - \lambda\right)^2},$$

where $\hat{\lambda}_i$ is the i^{th} estimated parameter of λ. Generally, a good estimator gives small MAD and $RMSE$ values. The results of the simulation study are given in Figure 3.

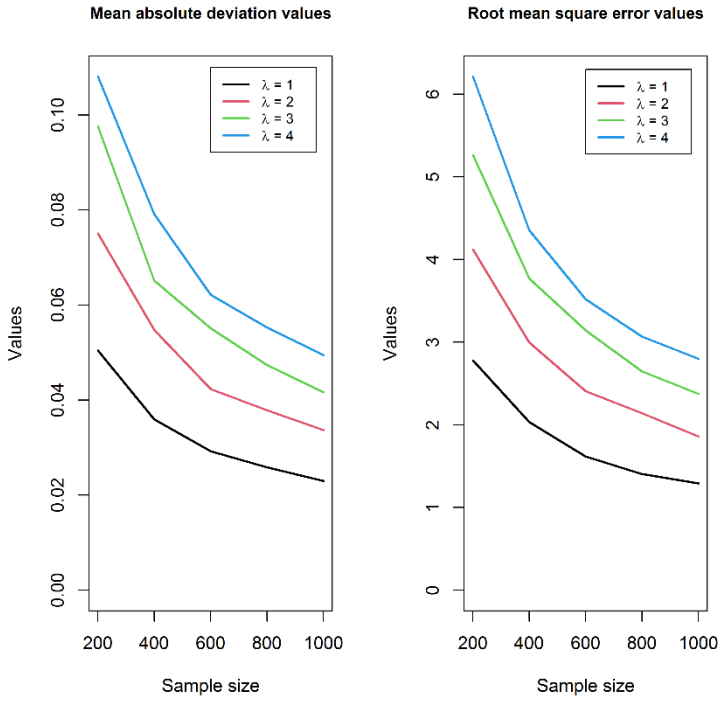

Figure 3. The MAD and $RMSE$ values for $\lambda = 1, 2, 3, 4$ with varying sample sizes

From Figure 3, it is clear that as n increases, both MAD and $RMSE$ values decrease, suggesting the estimator is asymptotically unbiased and consistent. Besides that, a larger sample size ($n \gg 1000$) is required to reduce the MAD and $RMSE$ values when dealing with large λ.

Applications

A dataset on bowel cancer data [11] is considered where $(n_1, n_2, n_3, n_4, n_5, n_6) = (8,12,16,21,12,31)$ for model fitting. The bowel cancer data is secondary data that was collected based on 122 patients with confirmed cancer status, with $n_0 = 22$ refers to deceased patients [11]. Therefore, only data based on the remaining 100 patients are used for modelling. The data is fitted to the USBP distribution and compared with the size-biased Poisson (SBP) distribution and the size-biased Poisson-Lindley (SBPL) distribution [3], and the best model is selected based on the chi-square goodness of fit test, mean squared error (MSE), as well as the root, mean squared error (RMSE) values, where

$$MSE = \sum_{x=1}^{k}\frac{(e_x)^2}{k} \text{ and } RMSE = \sqrt{MSE} = \sqrt{\sum_{x=1}^{k}\frac{(e_x)^2}{k}},$$

where e_x is the difference between observed and fitted x-valued data and $k = \max(x)$. Note that the MLE $\hat{\lambda}$ for SBP distribution is given as $\hat{\lambda} = \bar{x} - 1$ whereas the MLE $\hat{\theta}$ for SBPL distribution can be obtained by solving the maximum likelihood function given by [3] numerically (refer to page 305 in [3]). These two distributions have been selected for comparison because both have one parameter only as the proposed USBP distribution.

10.11113/mjfas.v19n2.2678

The summary of the model fittings is presented in Table 2. Based on Table 2, the significant p-value from the model fitting using the SBP distribution indicates that the SBP distribution does not fit the bowel cancer data adequately. On the other hand, both USBP and SBPL distributions provide adequate fitting to the data based on the non-significant p-value from the chi-square goodness of fit test. However, the MSE and the RMSE values from the model fitting based on the USBP distribution are the smallest. Hence, the USBP distribution is selected as the best model for describing the bowel cancer data.

Table 2. Summary of model fitting the bowel cancer data to the USBP, SBP, and SBPL distributions

Measures	Distributions		
	USBP	SBP	SBPL
Parameter estimate	2.3946	3.1000	0.8659
MSE	17.028	195.412	25.159
RMSE	4.127	13.979	5.016
χ^2	6.179	11.999	8.777
df	4	4	4
p-value	0.186	0.017	0.067

Conclusions

This study developed a new USBP distribution for count data and discussed the statistical properties as well as the estimation of parameters for the distribution. The USBP distribution is found to be underdispersed, skewed to the right, and has a heavy tail. The MLE and the ME of the parameter are found to be identical, and the estimator is unique, positively biased, consistent, and asymptotically normal. Simulation studies concluded that the estimator is asymptotically unbiased and consistent. Application study suggests that the USBP distribution provides an adequate and better fit than the SBP and the SBPL distributions.

The USBP distribution relies solely on one parameter, thus making it less flexible. However, it is believed that the USBP distribution can be further improved by 1) keeping the weight, θ, as a parameter and not as a function of λ thus providing flexible mixing proportion, and 2) considering two rate parameters i.e. λ_1 and λ_2 instead of a common λ, allowing the two types of size-biased distributions to be flexible. By doing such, the USBP distribution has the capacity to become more flexible. Developing new distributions based on the suggestions above will increase the number of parameters and ultimately decrease the degrees of freedom. The USBP distribution will then become a special case for the new distributions. With the recent advancement in a one-inflated model in the statistics literature, the USBP distribution can be further extended to cater to excess ones in the data.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgment

The authors gratefully acknowledge the financial support received in the form of research grants (FRGS/1/2019/STG06/UKM/01/5) from the Ministry of Education, Malaysia, and (GUP-2019-031) from Universiti Kebangsaan Malaysia.

References

[1] Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhya; The Indian Journal of Statistics, Series A*, *27*, 311-324.
[2] Patil, G. P., & Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, *34*, 179-189.
[3] Ghitany, M. E., & Al-Mutairi, D. K. (2008). Size-biased poisson-lindley distribution and its application. *Metron-International Journal of Statistics, 66,* 299-311.
[4] McLachlan, G., & Peel, D. (2002). General Introduction. In *Finite Mixture Models.* (Eds). Cressie, N. A. A.,

Fisher, I. N., Johnstone, I. M., Kadane, J. N., Scott, D. W., Silverman, B. W., Smith, A. F. M., Teugels, J. L., Barnett, V., Bradley, R. A., Hunter, J. S., & Kendall, D. G., pp. 1-39. New York: John Wiley & Sons.

[5] Tajuddin, R. R. M., Ismail, N., & Ibrahim, K. (2020). Several two-component mixture distributions for count data. *Communications in Statistics – Stimulations and Computation*, 1-12.

[6] Ismail, N., Mohd Ali, K. M., & Chiew, A. C. (2004). A Model for insurance claim count with single and finite mixture distribution. *Sains Malaysiana, 1185,* 48-61.

[7] Deni, S. M., Jemain, A. A., & Ibrahim, K. (2010). The Best probability models for dry and wet spells in peninsular malaysia during monsson seasons. *International Journal of Climatology, 30*(8), 1194-1205.

[8] Deni, S. M., & Jemain, A. A. (2009). Fitting the distribution of dry and wet spells with alternative probability models. *Meteorology and Atmospheric Physics*, 104, 13-27.

[9] Abramowitz, M., & Stegun, I. A. (Eds.). (1972). Gamma function and related functions. In *Handbook of Mathematical Functions: with Formula, Graphs and Mathematical Tables*, pp. 258-259. New York: Dover.

[10] Hogg, R. V., McKean, J. W., & Craig, A. T. (2005). Maximum likelihood estimation. In *Introduction to Mathematical Statistics*, (6th ed). pp. 313. New Jersey: Pearson Prentice Hall.

[11] Lloyd, C. J., & Frommer, D. (2004). estimating the false negative fraction for a multiple screening test for bowel cancer when negatives are not verified. *Australian and New Zealand Journal of Statistics, 46,* 531-542.