

Overdispersed Count Data Using Negative Binomial-Quasi XGamma Distribution

Siriporn Samutwachirawong ^{1*}

Received: 29 October 2025; Revised: 02 April 2026; Accepted: 05 May 2026

Abstract

This paper introduces the negative binomial-quasi xgamma (NB-QX) distribution, a three-parameter model derived by mixing the negative binomial and quasi xgamma distributions. The NB-QX distribution offers significant flexibility in analyzing overdispersed count data and subsumes the negative binomial-gamma distribution as a special case. We derive its fundamental properties, including the probability mass function, factorial moments, mean, and variance. Parameter estimation is conducted via the maximum likelihood method. The model's performance is evaluated using three real-world count datasets, with goodness-of-fit assessed through the Kolmogorov–Smirnov (KS) test for discrete distributions, AIC, and BIC. The empirical results demonstrate that the NB-QX distribution outperforms Poisson, negative binomial, and negative binomial-gamma distributions, particularly in capturing overdispersion within a unified framework.

Keywords: negative binomial distribution, quasi-xgamma distribution, mixed distribution, count data, overdispersion

* E-mail Address (Corresponding Author): siriporn_sm@mju.ac.th

¹ Program in Statistics and Information Management, Faculty of Science, Maejo University.

1. Introduction

In the analysis of count data, particularly with the Poisson distribution, it is often assumed that the data are equidispersed, meaning the mean and variance are equal. However, this assumption frequently fails in practice when real-world data exhibit overdispersion, where the variance exceeds the mean. Such variance inflation violates the fundamental equidispersion assumption underlying the Poisson distribution, rendering standard Poisson regression models inadequate for analyzing overdispersed count data. This misspecification can result in several serious consequences: systematically biased parameter estimates, substantially underestimated standard errors, artificially narrow confidence intervals, and invalid statistical inferences. These issues ultimately lead to increased Type I error rates in hypothesis testing and unreliable conclusions regarding the significance of predictor variables. To address these limitations, researchers have developed numerous generalized discrete distributions that enhance model flexibility and better accommodate the variability observed in real-world count data. To address overdispersion, researchers have developed various extensions of the Poisson model, including compound distributions, hierarchical formulations, and mixture models that introduce additional parameters to capture excess variation. A particularly powerful technique is compounding, which constructs new distributions by mixing a base discrete distribution (e.g., Poisson or negative binomial) with a continuous mixing distribution. This hierarchical construction naturally induces overdispersion while preserving desirable properties of the base distribution. Mixed negative binomial distributions represent a prominent example of this approach, where the negative binomial distribution serves as the base and is combined with various continuous distributions. These compound models offer several advantages: they provide explicit and tractable mean-variance relationships, accommodate varying degrees of overdispersion through the mixing distribution parameters, maintain interpretability of regression coefficients, and deliver improved statistical inference with more accurate parameter estimates and predictions compared to the standard Poisson distribution.

In the literature, various alternative mixed negative binomial distribution provides better fit for overdispersed count data as compared to traditional statistical distributions. Notable examples include The Negative Binomial-Lindley distribution, introduced by Zamani and Ismail (2010) through compounding the negative binomial distribution with the Lindley distribution for insurance claims applications, demonstrated superior fit and improved predictive capability compared to standard Poisson and negative binomial distributions, particularly when analyzing highly overdispersed count data, owing to its enhanced flexibility in modeling variance structure. The negative binomial-beta exponential distribution, developed by Pudprommarat et al. (2011) is a mixed negative binomial distribution that provides improved flexibility and superior fit for highly overdispersed count data when compared to standard Poisson and negative binomial distributions. The negative binomial-gamma distribution, developed by Gençtürk and Yiğiter (2015), is a mixed negative binomial distribution that provides improved flexibility and better performance for count data when compared to the standard negative binomial distribution. The negative binomial mixed models were proposed by Zhang et al. (2017), effectively address the correlation and overdispersion inherent in clustered microbiome count data. The negative binomial-inverse gaussian regression model was proposed by Tzougas et al. (2019) as an alternative to mixed Poisson regression models for approximating the number of claims. The zero-inflated negative binomial mixed model was applied by García et al. (2023) to Spanish wildfire counts to effectively address the challenges of excess zeros and overdispersion resulting from Mediterranean seasonal patterns by simultaneously modeling fire occurrence and non-occurrence. The mixed negative binomial regression model was used by İyit (2024) to investigate the global impact of digestive system cancers on COVID-19 outcomes, as the model was chosen for its ability to effectively handle the overdispersion inherent in both cancer and pandemic count data across various countries and regions. The negative binomial-Lindley distribution with time-dependent parameters, proposed by Dzinyela et al. (2024), effectively addresses temporal variations and overdispersion in crash data. The model demonstrates a superior goodness-of-fit and reveals that time-dependent parameters are crucial for accurate predictive modeling.

As discussed above, lifetime distributions serve as mixing models for generating mixed negative binomial distributions. Sen and Chandra (2017) proposed the quasi xgamma distribution as a flexible lifetime model that generalizes both the gamma and xgamma distributions. They examined various properties, including survival characteristics, order statistics, and parameter estimation methods (maximum likelihood and method of moments). Comparative studies showed that the quasi xgamma distribution provides superior flexibility in modeling lifetime data relative to several competing distributions, including the gamma, lognormal, various Lindley-based distributions, and xgamma distributions. The quasi xgamma distribution is selected as the mixing distribution for several theoretically motivated reasons. First, it generalizes both the gamma and xgamma distributions as special cases, offering a strictly broader class of distributions with greater shape flexibility than either constituent family. Second, unlike Lindley and exponential distributions, which are characterized by a single parameter and limited flexibility in capturing heterogeneity, the quasi xgamma distribution accommodates a wider range of dispersion structures through its additional shape parameter. Third, compared to the Inverse Gaussian distribution, the quasi xgamma distribution yields more tractable closed-form expressions for the moments of the resulting compound distribution, including non-negative support and flexible shape parameters, all of which facilitate the derivation of

closed-form results and support accurate parameter estimation and statistical inference. These characteristics collectively make the quasi xgamma distribution a theoretically principled and practically superior choice for constructing a new mixed negative binomial model.

This paper proposes a new distribution, the negative binomial-quasi xgamma distribution, obtained by mixing the negative binomial and quasi xgamma distributions. The quasi xgamma distribution is selected as the mixing distribution due to its theoretical advantages over classical alternatives: it generalizes both the gamma and xgamma distributions, offers greater flexibility in capturing heterogeneity across varying dispersion structures, and yields closed-form expressions for key distributional properties. Unlike existing mixed negative binomial models that rely on Gamma, inverse Gaussian, or Lindley distributions as mixing components, the NB-QX distribution inherits the additional flexibility of the quasi xgamma family, enabling it to simultaneously accommodate both moderate and heavy-tailed overdispersion within a unified theoretical framework. The proposed distribution is unimodal and exhibits heavy tails. Basic properties, including the probability mass function (pmf), factorial moments, mean, and variance, are discussed. The parameters are estimated using the maximum likelihood method. The goodness-of-fit is thoroughly assessed using the Kolmogorov–Smirnov (KS) test for discrete distributions, the Akaike information criterion (AIC) (Akaike, 1974), and the Bayesian information criterion (BIC) (Schwarz, 1978). The proposed distribution provides a more flexible alternative for analyzing count data with overdispersion.

2. A New Mixed Distribution for Count Data with Overdispersion

Assume that a random variable X follows a negative binomial (NB) distribution with parameters r and $p = e^{-\lambda}$, where λ is a random variable giving quasi xgamma (QX) distribution with parameters θ and α . Then, random variable X has a negative binomial-quasi xgamma (NB-QX) distribution. The NB-QX distribution and its properties are presented in the following subsections.

2.1 The negative binomial (NB) distribution

If a random variable X has a NB distribution, which is denoted by $X \sim NB(r, p)$, its the probability mass function (pmf) is given by

$$P(X = x) = \binom{r+x-1}{x} p^r (1-p)^x, \quad (1)$$

with x denotes the number of failures before the r -th success occurs, where $x = 0, 1, 2, \dots$,
 r denotes the number of successes, where $r > 0$, and
 p denotes the probability of success on each trial, where $0 < p < 1$.

The first four moments about zero and the factorial moment of the NB distribution are, respectively, given by

$$\begin{aligned} E(X) &= \frac{r(1-p)}{p}, \\ E(X^2) &= \frac{r(1-p)[1+r(1-p)]}{p^2}, \\ E(X^3) &= \frac{r(1-p)}{p^3} [1 + (3r+1) + r^2(1-p)^2], \\ E(X^4) &= \frac{r(1-p)}{p^4} [1 + (7r+6)(1-p) + (6r^2 + 11r + 7)(1-p)^2 + r^3(1-p)^3], \\ \mu_{[k]}(X) &= E(X(X-1)\cdots(X-k+1)) = \frac{\Gamma(r+k)}{\Gamma(r)} \frac{(1-p)^k}{p^k}, \quad k = 1, 2, \dots \end{aligned} \quad (2)$$

where $\Gamma(s) = \int_0^\infty x^{s-1} e^{-x} dx$, $s > 0$, denotes the complete gamma function (Klugman et al., 2019).

2.2 The quasi xgamma (QX) distribution

If a random variable λ has a QX distribution, which is denoted by its probability density function (pdf) is given by

$$f(\lambda) = \frac{\theta}{(1+\alpha)} \left(\alpha + \frac{\theta^2}{2} \lambda^2 \right) e^{-\theta\lambda}, \quad (3)$$

λ denotes the continuous random variable, where $\lambda \in (0, \infty)$,

α denotes the shape parameter, where $\alpha > 0$, and

θ denotes the scale parameter, where $\theta > 0$.

Its moment generating function is given as

$$M_\lambda(t) = E(e^{t\lambda}) = \frac{1}{(1+\alpha)} \left[\alpha \left(1 - \frac{t}{\theta} \right)^{-1} + \left(1 - \frac{t}{\theta} \right)^{-3} \right]. \quad (4)$$

Three special cases of the QX distribution as follows:

(i) If $\alpha = 0$, then the QX distribution reduces to the gamma distribution with shape parameter 3 and scale parameter θ .

(ii) If $\alpha = 1$, then the QX distribution reduces to the new class of distribution with pdf

$$f(\lambda) = \frac{\theta}{2} \left(1 + \frac{\theta^2}{2} \lambda^2 \right) e^{-\theta\lambda}, \quad \lambda > 0 \text{ and } \theta > 0 \text{ (Sen and Chandra, 2017)}. \quad (5)$$

(iii) If $\alpha = \theta$, then the QX distribution reduces to xgamma distribution with pdf

$$f(\lambda) = \frac{\theta^2}{(1+\theta)} \left(1 + \frac{\theta}{2} \lambda^2 \right) e^{-\theta\lambda}, \quad \lambda > 0 \text{ and } \theta > 0 \text{ (Sen et al., 2016)}. \quad (6)$$

2.3 The negative binomial-quasi xgamma distribution

The theorem of the new mixed negative binomial distribution is provided, and its basic properties are presented.

Theorem 1 a random variable X has the NB distribution with parameters r and $p = e^{-\lambda}$, where λ is a random variable giving QX distribution with parameter θ and α with pdf given by Equation (3), then X is called a NB-QX distribution, which is denoted by $X \square \text{NB-QX}(r, \theta, \alpha)$. The pmf of a NB-QX distribution is given by

$$P(X = x) = \frac{1}{(1+\alpha)} \binom{r+x-1}{x} \sum_{j=0}^x \binom{x}{j} (-1)^j \left[\alpha \left(1 + \frac{(r+j)}{\theta} \right)^{-1} + \left(1 + \frac{(r+j)}{\theta} \right)^{-3} \right], \quad (7)$$

with $x = 0, 1, 2, \dots$ and $r, \theta, \alpha > 0$,

Proof. If $X|\lambda \square \text{NB}(r, p = e^{-\lambda})$, $\lambda \square \text{QX}(\theta, \alpha)$, then the pmf of X can be obtained by

$$\begin{aligned} P(X = x) &= \int_0^\infty p(x|\lambda) f(\lambda) d\lambda, \\ &= \int_0^\infty \binom{r+x-1}{x} e^{-\lambda r} (1 - e^{-\lambda})^x f(\lambda) d\lambda, \\ &= \binom{r+x-1}{x} \sum_{j=0}^x \binom{x}{j} (-1)^j M_\lambda[-(r+j)]. \end{aligned} \quad (8)$$

By replacing the $M_\lambda(t)$ of QX distribution in Equation (4) with $t = -(r+j)$ into Equation (8), then the pmf of X is

$$P(X = x) = \frac{1}{(1+\alpha)} \binom{r+x-1}{x} \sum_{j=0}^x \binom{x}{j} (-1)^j \left[\alpha \left(1 + \frac{(r+j)}{\theta} \right)^{-1} + \left(1 + \frac{(r+j)}{\theta} \right)^{-3} \right],$$

which proves the theorem.

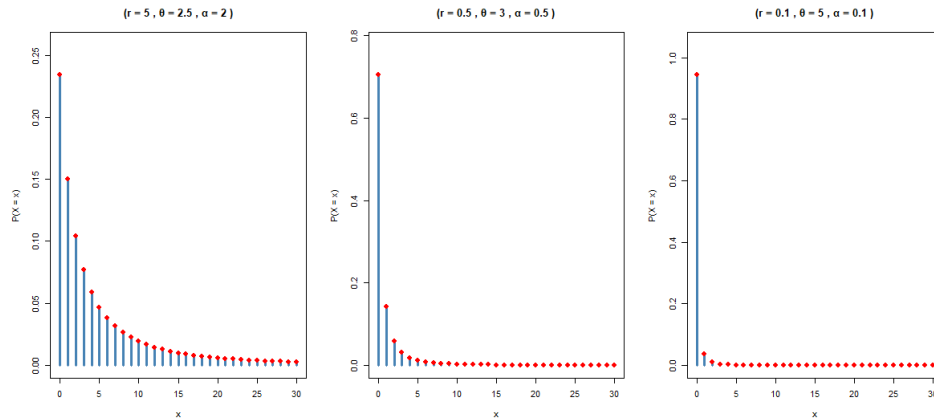


Figure 1 The pmf plot for different values of parameters.

Figure 1 illustrates pmf of the NB-QX distribution for different values of r, θ and α . The pmf plot indicate that all of them appear to be unimodal and positively skewed.

A function in Equation (7) is a pmf if and only if

1) $P(X = x) \geq 0$ for all x . It is clear that $P(X = x) \geq 0$.

2) $\sum_{x=0}^{\infty} P(X = x) = 1$, according to Theorem 1

$$\sum_{x=0}^{\infty} P(X = x) = \sum_{x=0}^{\infty} \binom{r+x-1}{x} e^{-\lambda r} (1 - e^{-\lambda})^x,$$

factor $e^{-\lambda r}$ and using the negative binomial series $\sum_{x=0}^{\infty} \binom{r+x-1}{x} y^x = (1-y)^{-r}$, where $y = -(e^{-\lambda} - 1) = 1 - e^{-\lambda}$.

Then,

$$\sum_{x=0}^{\infty} P(X = x) = e^{-\lambda r} (1 - (1 - e^{-\lambda}))^{-r} = e^{-\lambda r} e^{\lambda r} = 1.$$

It can be shown that the pmf in Equation (7) satisfies the required conditions.

Corollary 1 Let $X \square NB-QX(r, \theta, \alpha)$. If $\alpha = 0$, then the NB-QX distribution reduces to the negative binomial-gamma (NB-Gamma) distribution with shape parameter 3 and scale parameter θ . Its pmf is

$$P(X = x) = \binom{r+x-1}{x} \sum_{j=0}^x \binom{x}{j} (-1)^j (1 + \theta(r+j))^{-3}, \quad x = 0, 1, 2, \dots,$$

where $X|\lambda: NB(r, p = e^{-\lambda})$, has the NB distribution with parameters r and $p = e^{-\lambda}$, where λ has the gamma distribution with shape parameter 3 and scale parameter θ see Gençtürk and Yiğiter (2015).

Theorem 2 Let $X \square NB-QX(r, \theta, \alpha)$, then the factorial moment of order k of the NB-QX distribution is

$$\mu_{[k]}(X) = \frac{\Gamma(r+k)}{\Gamma(r)(1+\alpha)} \sum_{j=0}^k \binom{k}{j} (-1)^j \left[\alpha \left(1 - \frac{k-j}{\theta} \right)^{-1} + \left(1 - \frac{k-j}{\theta} \right)^{-3} \right], \quad (9)$$

with $k = 1, 2, \dots$ and $r, \theta, \alpha > 0$.

Proof. If $X| \lambda \sim NB(r, p = e^{-\lambda})$, $\lambda \sim QX(\theta, \alpha)$, then the factorial moment of order k can be found by using the factorial moment of order k of the NB distribution (Gómez-Déniz et al. 2008) as follows

$$\mu_{[k]}(X) = \frac{\Gamma(r+k)}{\Gamma(r)} E_{\lambda}(e^{\lambda} - 1)^k,$$

through the binomial expansion of $(e^{\lambda} - 1)^k = \sum_{j=0}^k \binom{k}{j} (-1)^j e^{\lambda(k-j)}$,

$$\begin{aligned} \mu_{[k]}(X) &= \frac{\Gamma(r+k)}{\Gamma(r)} E_{\lambda} \sum_{j=0}^k \binom{k}{j} (-1)^j e^{\lambda(k-j)}, \\ &= \frac{\Gamma(r+k)}{\Gamma(r)} M_{\lambda}(k-j), \end{aligned}$$

from the $M_{\lambda}(t)$ of QX distribution in Equation (4) with $t = k - j$, then the factorial moment of order k as

$$\mu_{[k]}(X) = \frac{\Gamma(r+k)}{\Gamma(r)(1+\alpha)} \sum_{j=0}^k \binom{k}{j} (-1)^j \left[\alpha \left(1 - \frac{k-j}{\theta}\right)^{-1} + \left(1 - \frac{k-j}{\theta}\right)^{-3} \right],$$

which proves the theorem.

The mean, second-order moment and variance can be obtained directly from Equation (9) which are given by

$$E(X) = r[M_{\lambda}(1) - 1], \quad (10)$$

$$E(X^2) = (r + r^2)M_{\lambda}(2) - (r + 2r^2)M_{\lambda}(1) + r^2,$$

$$Var(X) = (r + r^2)M_{\lambda}(2) - rM_{\lambda}(1)(1 + rM_{\lambda}(1)), \quad (11)$$

where $M_{\lambda}(\cdot)$ is defined in Equation (4).

Table 1 Mean and variance of the NB-QX distribution for different values of r, θ and α .

α	θ	$r = 0.1$		$r = 0.5$		$r = 1$		$r = 5$	
		mean	variance	mean	variance	mean	variance	mean	variance
0.05	2.5	0.349	12.471	1.744	82.183	3.489	213.936	17.443	3052.455
	3	0.229	2.408	1.143	15.051	2.286	37.633	11.429	489.388
	5	0.092	0.265	0.460	1.485	0.920	3.372	4.598	32.932
	10	0.036	0.057	0.180	0.298	0.359	0.632	1.797	4.598
1	2.5	0.215	6.736	1.074	44.698	2.148	116.941	10.741	1686.488
	3	0.144	1.347	0.719	8.546	1.438	21.621	7.188	289.277
	5	0.060	0.160	0.301	0.919	0.602	2.130	3.008	22.312
	10	0.024	0.037	0.121	0.195	0.241	0.421	1.207	3.311
1.5	2.5	0.185	5.463	0.926	36.291	1.852	95.015	9.259	1372.414
	3	0.125	1.110	0.625	7.059	1.250	17.888	6.250	240.188
	5	0.053	0.137	0.266	0.787	0.531	1.828	2.656	19.281
	10	0.022	0.032	0.108	0.171	0.215	0.370	1.077	2.933
3	2.5	0.141	3.551	0.704	23.597	1.407	61.797	7.037	893.073
	3	0.097	0.754	0.484	4.797	0.969	12.155	4.844	163.257
	5	0.043	0.102	0.213	0.584	0.426	1.356	2.129	14.272
	10	0.018	0.025	0.088	0.135	0.176	0.292	0.881	2.302

Table 1 presents the mean and variance of the NB-QX distribution for different values of the parameters r, θ and α , respectively. The mean and variance of the NB-QX distribution are increasing functions of parameter θ and α , and decreasing functions of parameter r .

3. Estimation

Maximum likelihood estimation (MLE), a popular estimation method, is used to estimate the parameters of the NB-QX distribution.

The likelihood function of $NB-QX(r, \theta, \alpha)$ is given by

$$L(r, \theta, \alpha) = \frac{1}{(1+\alpha)^n} \prod_{i=1}^n \binom{r+x_i-1}{x_i} \sum_{j=0}^x \binom{x_i}{j} (-1)^j \left[\alpha \left(1 + \frac{(r+j)}{\theta} \right)^{-1} + \left(1 + \frac{(r+j)}{\theta} \right)^{-3} \right]. \quad (12)$$

The log likelihood function of NB-QX distribution is given by

$$\begin{aligned} \log L(r, \theta, \alpha) = & \sum_{i=1}^n [\log \Gamma(r+x_i) - \log \Gamma(x_i+1) - \log \Gamma(r)] - n \log(1+\alpha) \\ & + \sum_{i=1}^n \log \left[\sum_{j=0}^x \binom{x_i}{j} (-1)^j \left[\alpha \left(1 + \frac{(r+j)}{\theta} \right)^{-1} + \left(1 + \frac{(r+j)}{\theta} \right)^{-3} \right] \right]. \end{aligned} \quad (13)$$

The ML estimators of $\hat{r}, \hat{\theta}$ and $\hat{\alpha}$ of r, θ and α , respectively, can be solving the following partial differential Equations

$$\frac{\partial \log L}{\partial r} = 0, \quad \frac{\partial \log L}{\partial \theta} = 0, \quad \text{and} \quad \frac{\partial \log L}{\partial \alpha} = 0.$$

The above Equations are solved using numerical optimization with the **optim** function in the *stats* package in R (R Core Team, 2025). To ensure numerical stability in the optimization of the log-likelihood function in Equation (13), the **optim** function in R was implemented using the L-BFGS-B method, which supports parameter constraints and improves convergence behavior. Multiple sets of initial values were used to verify that the algorithm converged consistently to the same solution, reducing the risk of convergence to a local maximum. Parameter constraints $r > 0$, $\theta > 0$, and $\alpha > 0$ were imposed throughout the optimization to ensure that all estimates remain within the valid parameter space.

4. Applications of NB-QX distribution

To illustrate the flexibility and practical applicability of the proposed distribution, we applied Poisson, NB, NB-Gamma (with shape parameter 3 and scale parameter θ), and NB-QX distributions to real data sets. The parameters for each distribution were estimated using MLE via the **optim** function in R. To compare the goodness-of-fit of these distributions, we computed three complementary criteria: the KS test statistic, which evaluates the maximum deviation between empirical and theoretical cumulative distribution functions; the AIC, which balances model fit and complexity; and the BIC, which applies a stronger penalty for model complexity and is particularly useful for comparing models with different numbers of parameters. Together, these criteria provide a comprehensive and robust evaluation of model performance. The KS test statistic was computed using the **ks.test** function from the *dgoF* package in R (Marsaglia et al., 2003; R Core Team, 2025). The distribution with the smallest KS statistic, highest log-likelihood, and lowest AIC and BIC values is considered to provide the best fit.

Data Set I: The proposed NB-QX distribution is applied to automobile insurance claim frequencies. The data, originally analyzed by Simmachan and Boonkrong (2025), represent the number of claims filed under 4,986 automobile insurance policies in Thailand during the calendar year 2017. Descriptive analysis reveals several important characteristics of this data set. The sample mean and variance are 0.48 and 0.68, respectively, yielding

a variance-to-mean ratio of 1.42, which provides clear evidence of overdispersion relative to the Poisson distribution's equidispersion assumption. Furthermore, a large proportion of policies (67.85%) had no claims, the distribution is positively skewed. These characteristics are typical of insurance claims data and require more flexible models than Poisson distribution.

Data Set II: The second data set, obtained from Shanker and Hagos (2015), consists of Haemocytometer yeast cell counts per square from a biological study. The data exhibits overdispersion with a sample mean of 0.46 and variance of 0.60, yielding a variance-to-mean ratio of 1.30. Additionally, 68.45% of observations are zero.

Data Set III: The third data set, obtained from Tin (2008), comprises hospitalization counts from 10,117 cases tracked over a 12-month survey period. The data contains an extremely high proportion of zeros, with 90.12% of cases recording no hospitalizations. Descriptive statistics reveal a sample mean of 0.14 and variance of 0.27, yielding a variance-to-mean ratio of 0.52, which indicates overdispersion.

Table 2 Parameter estimates, observed and expected frequencies, and goodness-of-fit statistics for each distribution fitted to data set I.

number of claims	Observed frequencies	Expected frequencies			
		Poisson	NB	NB-Gamma	NB-QX
0	3383	3093	3377	3366	3376
1	1065	1477	1088	1117	1085
2	369	353	352	339	356
3	116	56	114	107	115
4	41	7	37	36	37
5	8	1	12	13	12
6	3	0	4	5	4
7	0	0	1	2	1
8	1	0	0	1	0
parameter estimates		$\hat{\lambda} = 0.478$	$\hat{r} = 0.990$ $\hat{p} = 0.675$	$\hat{r} = 2.273$ $\hat{\theta} = 0.062$ $\alpha = 3$ (fixed)	$\hat{r} = 23.532$ $\hat{\theta} = 87.717$ $\hat{\alpha} = 1.676$
log likelihood		-4803.387	-4635.726	-4639.582	-4635.232
AIC		9608.774	9275.452	9283.164	9276.465
BIC		9615.288	9288.481	9296.193	9296.007
KS test		0.620	0.677	0.675	0.002

Table 3 Parameter estimates observed and expected frequencies, and goodness-of-fit statistics for each distribution fitted to data set II.

number of yeast cell	Observed frequencies	Expected frequencies			
		Poisson	NB	NB-Gamma	NB-QX
0	128	118	127	126	127
1	37	54	42	44	42
2	18	12	13	13	13
3	3	2	4	4	4
4	1	0	1	1	1
parameter estimates		$\hat{\lambda} = 0.460$	$\hat{\mu} = 1.288$ $\hat{p} = 0.556$	$\hat{\mu} = 3.429$ $\hat{\theta} = 0.041$ $\alpha = 3$ (fixed)	$\hat{\mu} = 22.059$ $\hat{\theta} = 100.00$ $\hat{\alpha} = 0.902$
log likelihood		-173.832	-170.024	-170.346	-167.490
AIC		349.663	344.048	344.692	340.980
BIC		352.895	350.510	351.154	350.673
KS test		0.631	0.678	0.673	0.020

Table 4 Parameter estimates observed and expected frequencies, and goodness-of-fit statistics for each distribution fitted to data set III.

number of hospital days	Observed frequencies	Expected frequencies			
		Poisson	NB	NB-Gamma	NB-QX
0	9118	8786	9134	9462	9135
1	725	1239	700	513	725
2	165	87	187	98	168
3	57	4	62	28	54
4	17	0	22	10	21
5	14	0	8	4	9
6	21	0	3	2	4
parameter estimates		$\hat{\lambda} = 0.141$	$\hat{\mu} = 0.168$ $\hat{p} = 0.544$	$\hat{\mu} = 0.254$ $\hat{\theta} = 0.137$ $\alpha = 3$ (fixed)	$\hat{\mu} = 0.330$ $\hat{\theta} = 7.1246$ $\hat{\alpha} = 0.553$
log likelihood		-4629.711	-4099.042	-4095.969	-4095.926
AIC		9261.421	8202.085	8195.939	8197.852
BIC		9268.644	8216.528	8210.382	8219.518
KS test		0.868	0.903	0.902	0.001

Table 2 (Automobile Insurance Claims, $n = 4,986$): The NB-QX distribution demonstrates superior performance, yielding the highest log-likelihood (-4,635.232), the lowest AIC (9,276.465), and the minimum KS statistic (0.002). While the BIC for NB-QX (9,296.007) is marginally higher than that of the NB model (9,288.481) due to the additional parameter penalty, the superior AIC and log-likelihood confirm that NB-QX provides the best overall balance between goodness-of-fit and model complexity.

Table 3 (Yeast Cell Counts, $n = 187$): The NB-QX distribution outperforms the Poisson, NB, and NB-Gamma models, achieving the highest log-likelihood (-167.490), the lowest AIC (340.980), and the smallest KS statistic (0.020).

Table 4 (Hospitalization Counts, $n = 10,117$): The NB-QX model provides the best fit, characterized by the highest log-likelihood (-4,095.926) and the smallest KS statistic (0.001). Like Table 2, despite a slightly higher BIC compared to the NB-Gamma model, the NB-QX distribution remains the preferred model based on AIC and log-likelihood values.

The estimated values of α (ranging from 0.553 to 1.676) across all datasets underscore the importance of the quasi-xgamma mixing component. Notably, the model effectively captures high proportions of zeros—reaching 90.12% in Dataset III—without requiring a zero-inflated formulation. This confirms that the NB-QX distribution offers a parsimonious yet flexible framework capable of accommodating both extreme zero-counts and positive skewness.

5. Conclusions

The results demonstrate a clear and consistent pattern favoring the NB-QX distribution across three diverse applications: automobile insurance claims from the Thai insurance industry (Table 2), haemocytometer yeast cell counts per square unit representing microbial abundance in biological samples (Table 3), and hospitalization admission frequencies from longitudinal healthcare surveillance data (Table 4). It achieves the smallest KS test in every case, indicating the closest correspondence between theoretical and empirical cumulative distribution functions. Similarly, it attains the highest log-likelihood values and lowest AIC values, reflecting both superior fit and appropriate model complexity. The consistency of these findings across diverse data types with different means, variances, and zero proportions provides robust empirical validation of the NB-QX distribution's flexibility and effectiveness in accommodating various forms of overdispersion in count data.

Future research directions include several promising extensions. First, zero-truncated and zero-hurdle versions of the NB-QX distribution could explicitly accommodate datasets with excess or deficient zeros. Second, regression frameworks incorporating univariate and multivariate NB-QX distributions would enable flexible covariate modeling for correlated count data. Third, comprehensive simulation studies evaluating the finite-sample properties of parameter estimators under various data-generating scenarios, including different levels of overdispersion and zero proportions, would provide valuable guidance for practitioners.

References

- Akaike, H. (1974). A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>.
- Bugallo, M., Esteban, M. D., Marey Pérez, M. F., & Morales, D. (2023). Zero inflated negative binomial mixed models for predicting number of wildfires in Spain. **Journal of Environmental Management**, 330, 116788. <https://doi.org/10.1016/j.jenvman.2022.116788>.
- Dzinyela, R., Shirazi, M., Das, S., & Lord, D. (2024). The negative binomial Lindley model with time dependent parameters: Accounting for temporal variations and excess zero observations in crash data. **Accident Analysis & Prevention**, 207, 107711. <https://doi.org/10.1016/j.aap.2024.107711>.
- Gençtürk, M., & Yiğiter, E. (2015). Negative binomial-gamma distribution for overdispersed count data. **Communications in Statistics–Theory and Methods**, 44(12), 2565–2576. <https://doi.org/10.1080/03610926.2013.791373>.
- İyit, N. (2024). The global effects of digestive system cancers worldwide on the COVID-19 pandemic by negative binomial (mixed) regression models. **Journal of Radiation Research and Applied Sciences**, 17(3), 100944. <https://doi.org/10.1016/j.jrras.2024.100944>.
- Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2019). **Loss models: From data to decisions** (5th ed.). Wiley.

- Marsaglia, G., Tsang, W. W., & Wang, J. (2003). Evaluating Kolmogorov's distribution. **Journal of Statistical Software**, 8(18), 1–4.
- Pudprommarat, C., Bodhisuwan, W., & Zeepongsekul, P. (2012). A new mixed negative binomial distribution. **Journal of Applied Sciences**, 12, 1853–1858. <https://doi.org/10.3923/jas.2012.1853.1858>.
- R Core Team. (2025). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Schwarz, G. E. (1978). Estimating the dimension of a model. **The Annals of Statistics**, 6(2), 461–464. <https://doi.org/10.1214/aos/1176344136>.
- Sen, P., & Chandra, P. (2017). The quasi xgamma distribution: A flexible lifetime model. **Journal of Statistical Theory and Practice**, 11(4), 755–774. <https://doi.org/10.1007/s41060-017-0056-5>.
- Sen, S., Maiti, S. S., & Chandra, N. (2016). The xgamma distribution: Statistical properties and application. **Journal of Modern Applied Statistical Methods**, 15(1), Article 38. <https://doi.org/10.22237/jmasm/1461851580>.
- Shanker, R., & Fesshayee, H. (2015). On Poisson Lindley distribution and its applications to biological sciences. **Biometrics & Biostatistics International Journal**, 2(4), 103–107. <https://doi.org/10.15406/bbij.2015.02.00036>.
- Simmachan, T., & Boonkrong, P. (2025). A comparison of count and zero-inflated regression models for predicting claim frequencies in Thai automobile insurance. **Lobachevskii Journal of Mathematics**, 45, 6400–6414. <https://doi.org/10.1134/S1995080224607604>.
- Tin, A. (2008). **Modeling zero-inflated count data with underdispersion and overdispersion**. Conference paper, Research Foundation for Mental Hygiene, New York, NY.
- Tzougas, G., Hoon, W. L., & Lim, J. M. (2019). The negative binomial–inverse Gaussian regression model with an application to insurance ratemaking. **LSE Research Online Documents on Economics**, 101728. <https://lse.ac.uk/lse-library>.
- Zamani, H., & Ismail, N. (2010). Negative binomial-Lindley distribution for insurance claims. **Journal of Mathematics and Statistics**, 6(4), 309–314. <https://thescipub.com/abstract/jmssp.2010.4.9>.
- Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., et al. (2017). Negative binomial mixed models for analyzing microbiome count data. **BMC Bioinformatics**, 18, 4. <https://doi.org/10.1186/s12859-016-1441-7>.