Volume 70, Issue 2, 2026

Journal of Scientific Research

of

The Banaras Hindu University

Understanding Monthly Sale of Black Color Luxury Cars Through Probability Models

Brijesh P. Singh [1] and Shilpa Singh [*2]

[1]Department of Statistics, Institute of Science, Banaras Hindu University, Varanasi, brijesh@bhu.ac.in
[*2]Department of Statistics, Institute of Science, Banaras Hindu University, Varanasi, singhshilpa.stat.2025@bhu.ac.in

Abstract: **The analysis of count data within a fixed time interval is an important area of applied statistical modeling. In the automobile industry particularly luxury cars, monthly sales data often produce irregular patterns with significant number of zero observations. In order to analyze this phenomenon, classical Poisson distribution and Zero-Inflated Poisson distribution (ZIP) are fitted to the observed data. Parameters are estimated using maximum likelihood estimation . The fitting of the models is evaluated through graphical analysis and statistical model comparison criteria including the log-likelihood function, Akaike Information Criterion and goodness of fit procedures. Graphical comparisons between observed and expected frequencies are also employed to assess the adequacy of the fitted models visually.**

Index Terms: **Akaike Information Criteria, Automobile sales, Count data modeling, Goodness-of-fit, Maximum Likelihood Estimation, Model selection and Zero-inflated distribution.**

I. Introduction

Sales forecasting is an essential component for business planning and resource allotment. Data driven approaches are popular for providing predictions of future sales and demand, which is very important for inventory management and business growth. Artificial Intelligence based approaches have been widely adopted nowadays, but current Artificial Intelligence based methods are weak at capturing long term predictions and lack interpretability. Classical Statistics continues to remain incomprehensible to many analysts, even as they are amazed by the incredible power of machine learning. As a result, many have become more mechanical in their application of statistical tools.

Analysts fail to fathom that machine learning is not the only thoroughfare for solving real world problems. In many situations, it does not help them disentangle business problems, even when the data is involved. Distribution Fitting is the preeminent inferential technique used in the statistical world. It is generally the first school of thought that a person entering the world of statistics encounters. This can be applied to trend analysis of sales and demand for different automobile models and colors to maximize profit. Sales data recorded as counts are non-negative integer valued variables suitable for count data analysis. Count data are commonly seen in several areas of scientific research, including manufacturing, sales, insurance, transportation, public health, and ecology and are modelled using discrete probability distributions. Evans (1953) showed that different ecological count datasets were better described by various discrete probability distributions. Poisson distribution is the most widely used statistical model for count data because of its mathematical tractability and well-established theoretical framework (Fisher et al., 1943 ; Singh et al., 2012).

Count data are often under dispersed or over dispersed relative to the Poisson distribution, though the latter is less common in practice. Most frequent overdispersion is an excess of zero counts relative to the Poisson distribution, making the standard Poisson model inadequate in some situations. To address this limitation, several inflated probability models have been proposed. Singh et al. (2021) developed composite and inflated Poisson-Exponential probability distribution models and showed their applicability to patterns in rural out-migration data, implicating the usefulness of count data models.

The Zero-Inflated Poisson model (ZIP) was introduced by Singh (1963) and has been widely applied in various practical problems involving zero-inflated count data. Shahsavari et al. (2024) applied ZIP and robust ZIP model to evaluate the effect of outliers in the Length of Stay data in hospitals having a high proportion of zeros. These studies show the usefulness of ZIP models across various domains and motivate their use in industrial count data exhibiting similar zero-inflation characteristics. Given the extensive use of count data models and the presence of excess zeros in luxury automobile sales, which represent non-negative integer counts, the present study explores the suitability of the Poisson and ZIP distributions for modeling such data. It compares them to determine which model best describes the data.

[*] Corresponding Author

II. Probability Models

A. *Poisson Distribution*

Poisson distribution is discrete probability distribution of the number of events that occur in a fixed time interval or space, given the average number of times the event occurs over that time period (Cameron & Trivedi, 1986). The Poisson distribution is applicable only when several conditions hold:

- An event can occur any number of times in a fixed time interval.
- Events occur independently.
- The rate of occurrence is constant.
- The probability of occurrence of an event is directly proportional to the length of the time interval.

Let the number of automobile sales follow a Poisson Distribution with parameter λ, denoted as P(λ). Then the probability of observing events over the time period is,

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0,1,2, \dots$$

Where, $\lambda>0$

Mean and Variance of P(λ) : E(X) = λ and Var(X) = λ

Thus, the Poisson model assumes that the mean and variance are equal.

1) *Graphical Illustration*

Increasing the values of λ in P(λ) shifts the probability mass toward higher sale counts. Smaller λ, concentrates probability at zero and low counts, indicating lower average sales, as shown in Figure 1, using a line graph.

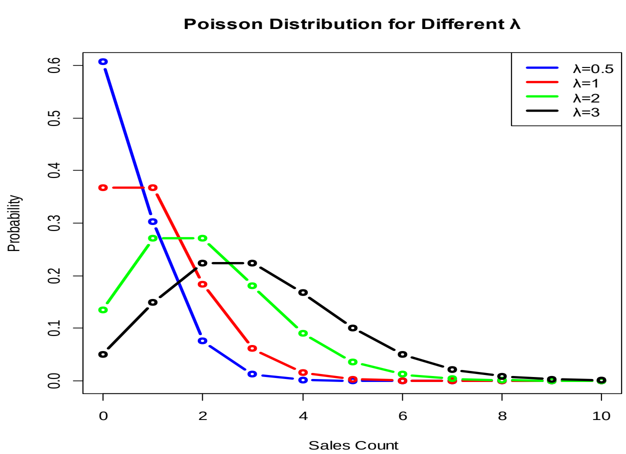

Fig. 1. Graph depicting effect of λ on P(λ).

B. *Zero-Inflated Poisson Distribution*

If the observed data have a large number of zero counts, then the zeros in count data may not be appropriately demonstrated by the standard Poisson distribution and the Zero-Inflated Poisson (ZIP) distribution is considered (Singh, 1963). The ZIP distribution with parameters α and λ, denoted by ZIP(α, λ), has the following probability mass function:

$$P(X = x) = \begin{cases} (1-\alpha) + \alpha e^{-\lambda} & ; \quad x = 0 \\ \alpha \dfrac{e^{-\lambda}\lambda^x}{x!} & ; \quad x \geq 1 \end{cases}$$

Where, $0 \leq \alpha \leq 1$, $\lambda > 0$

α: Probability of the months in the automobile is available in the show room for sale.

λ : Average number of automobiles sold per month when it is available for sale.

1) *Mean and Variance of ZIP(α, λ) :*

$$E[X] = \sum_{x=0}^{\infty} x . P(X = x)$$

$$= \alpha e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!}$$

$$= \alpha \, \lambda \, e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$$= \alpha \, \lambda \, e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$

$$= \alpha \, \lambda \, e^{-\lambda} \, e^{\lambda}$$

$$= \alpha \, \lambda$$

Similarly,

$$E[X^2] = \alpha(\lambda + \lambda^2)$$

Therefore,

$$\begin{aligned} Var(X) &= E(X^2) - (E(X))^2 \\ &= \alpha(\lambda + \lambda^2) - (\alpha \, \lambda)^2 \\ &= \alpha \, \lambda + \alpha(1-\alpha)\lambda^2 \end{aligned}$$

Since, $Var(X) > E(X)$, the ZIP distribution is capable of modeling overdispersion caused by excess zero observations.

2) *Graphical Illustration*

The ZIP distribution can be illustrated for different values of α ∈ [0,1] and λ > 0. To show how the probability of zero sales and the distribution of positive sales counts change with the parameters, the effect of each parameter is examined separately by fixing one parameter while varying the other.

a) *Effect of α when λ is fixed*

Smaller value of α increases the probability of zero sales. As α increases, the distribution gradually approaches the Poisson distribution, as shown in Figure 2, which depicts the effect of parameter α on ZIP(α, λ) when the other parameter λ is fixed at 1.

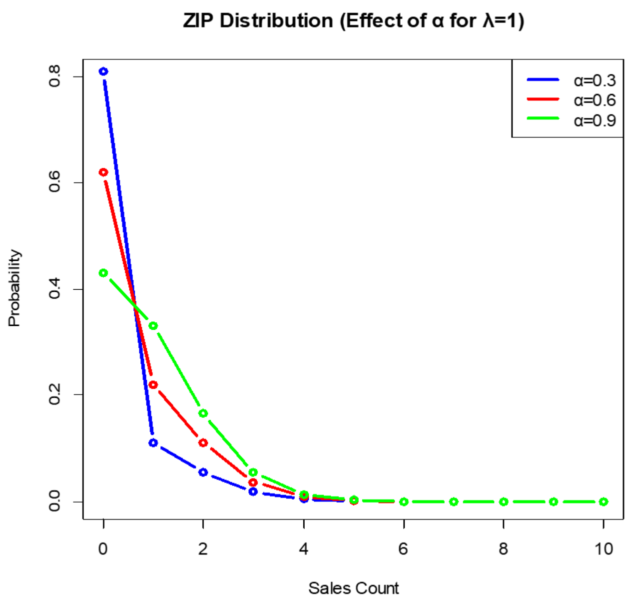

Fig.2.Graph depicting effect α on ZIP(α, λ) when λ is fixed at 1.

b) Effect of λ when α is fixed

Increasing λ shifts the probability toward higher sale counts. Smaller λ concentrates probability near zero and at low counts, as clearly shown in the line graph in Figure 3, which depicts the effect of parameter λ on ZIP(α, λ) when the other parameter α is fixed at 0.5.

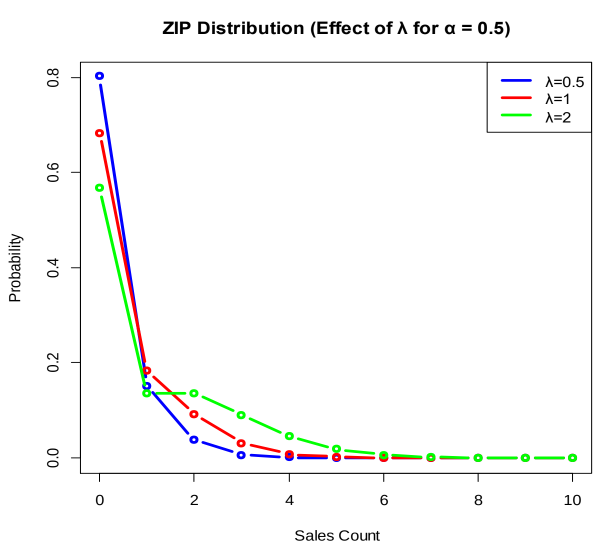

Fig.3. Graph depicting effect of λ on ZIP(α, λ) when α is fixed at 0.5.

III. Estimation of Parameters

A. Maximum Likelihood Estimate (MLE) for Poisson Distribution

Let $X_1, X_2, \dots, X_n$ be independent and identically distributed (i.i.d.) observations from a Poisson distribution with parameter λ. i.e. $X_i \sim P(\lambda)$, then the likelihood function of Poisson Distribution is,

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i\,!}$$

The log-likelihood function is,

$$\ell(\lambda) = -n\lambda + \left(\sum_{i=1}^{n} x_i\right)\log(\lambda) - \log\left(\prod_{i=1}^{n} x_i\,!\right)$$

$$= -n\lambda + \log(\lambda)\left(\sum_{i=1}^{n} x_i\right) - \sum_{i=1}^{n} \log\,(x_i!)$$

Now differentiating with respect to λ,

$$\frac{d\ell}{d\lambda} = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0$$

And equating it with zero we get the MLE of λ as,

$$\hat{\lambda} = \bar{X}$$

B. Maximum Likelihood Estimate (MLE) for Poisson Distribution

Let $X_1, X_2, \dots, X_n$ be an i.i.d. observations from a ZIP distribution with parameters α *and* λ. i.e.

$$X_i \sim ZIP(\alpha, \lambda).$$

Let

$n - n_0$ = number of zero observations

$n - n_0$ = number of positive observations

The likelihood function of $ZIP(\alpha, \lambda)$ is,

$$L(\alpha, \lambda) = \prod_{i=1}^{n} P(X_i = x_i)$$

$$= \left((1-\alpha) + \alpha e^{-\lambda}\right)^{n_0} \left(\alpha e^{-\lambda}\right)^{n-n_0} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

The log-likelihood function is,

$$\ell(\alpha,\lambda) = n_0 \log\left((1-\alpha)+\alpha e^{-\lambda}\right) + (n-n_0)\log(\alpha) - (n-n_0)\lambda + \left(\sum_{i=1}^{n} x_i\right)\log(\lambda) - \sum_{i=1}^{n}\log(x_i!)$$

Differentiating the log-likelihood function with respect to α and equating it with zero gives,

$$\frac{\partial \ell}{\partial \alpha} = \frac{n_0\left(e^{-\lambda}-1\right)}{(1-\alpha)+\alpha e^{-\lambda}} + \frac{n-n_0}{\alpha}$$

$$\frac{n_0\left(e^{-\lambda}-1\right)}{(1-\alpha)+\alpha e^{-\lambda}} + \frac{n-n_0}{\alpha} = 0 \qquad (1)$$

Similarly differentiating the log-likelihood function with respect to λ and equating it with zero gives,

$$\frac{\partial \ell}{\partial \lambda} = -\frac{n_0\,\alpha e^{-\lambda}}{(1-\alpha)+\alpha e^{-\lambda}} - (n-n_0) + \frac{\sum_{i=1}^{n} x_i}{\lambda}$$

$$\frac{n\bar{x}}{\lambda} = (n-n_0) + \frac{n_0\,\alpha e^{-\lambda}}{(1-\alpha)+\alpha e^{-\lambda}} \qquad (2)$$

The above two likelihood equations (1) and (2) are nonlinear, hence the maximum likelihood estimates of α and λ can be obtained numerically using the Newton-Raphson method. In the present study, these are solved numerically using the nlm optimization function in RStudio.

IV. Application of The Model

A. Data Description

The present study examines the monthly sales pattern of black cars priced above fifty lakhs using probability models. The dataset contains 69 monthly sales records for a black luxury car from September 2012 to May 2018 at a Mercedes-Benz showroom in Lucknow, Uttar Pradesh, India. Therefore, it contains only non-negative integer values. Each observation represents the number of cars sold in a given month, obtained from a real-world market dataset of a luxury car showroom. The monthly sales frequency over 69 months is shown in Table I. Prior to fitting the probability distributions, the observed sales data were explored using descriptive statistics and graphical methods to understand their underlying characteristics. These analyses showed that the largest proportion of months recorded zero sales, indicating overdispersion due to excess zeros.

Table I. Observed monthly sale of selected car in 69 months

Monthly Sale Count	0	1	2	3
Observed Frequency	59	7	2	1

B. Model Comparision

1) Akaike Information Criteria

For the observed sales count, the Poisson and ZIP models are compared using the Akaike information criteria (AIC) and log-likelihood measures in Table II.

Table II. Comparision of Poisson and ZIP using log-likelihood and AIC

Model	**-2(Log-Likelihood)**	**AIC**
Poisson	79.02	81.02
ZIP	67.36	71.36

The ZIP model shows lower -2(log-likelihood) and lower AIC compared to the Poisson model. Therefore, the ZIP model provides a better fit for the observed sales data.

2) Likelihood Ratio Test

It is used to compare the Poisson model with the ZIP model to determine whether the additional parameter in the ZIP model yields a significant improvement in model fitting. The null and alternative hypothesis are as follows:

H_0: *Poisson model adequately describes the data.*
H_1: *Zero-Inflated Poisson model provides a better fit.*

Test Statistic: $LR = -2(\ell_{Poisson} - \ell_{ZIP})$; $LR \sim \chi^2_{(2-1)}$

Calculated Value: $LR = 11.66$

Critical value at 5% level of significance,

$$\chi^2_{(0.05,1)} = 3.84$$

Since $LR > \chi^2_{0.05,1}$,therefore the null hypothesis H_0 is rejected at 5% level of significance. Thus, the Zero-Inflated Poisson model provides a significantly better fit than the Poisson model.

3) Grphical Comparision

The histogram in Figure 4 represents the observed sales frequency, while the curves show the expected frequencies under the Poisson and ZIP models. It clearly shows that the ZIP model better captures the high frequency of zero sales compared to the Poisson model.

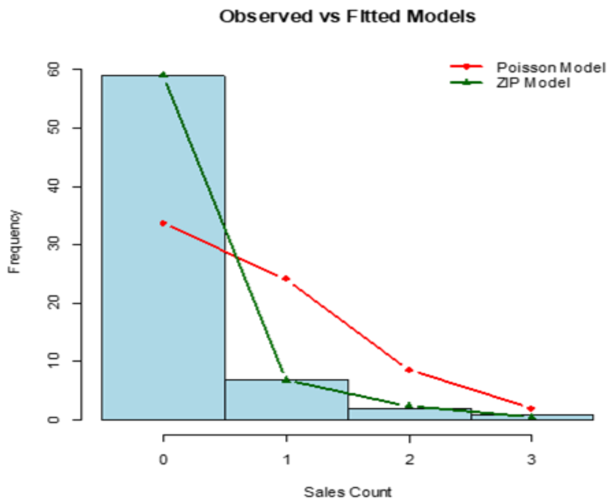

Fig.4. Comparision between the expected frequencies of Poisson and ZIP model.

C. Observed vs Expected frequencies of ZIP Model

The ZIP model better captures the high frequency of zero sales compared to the Poisson model and Table III presents the comparison between the observed frequencies of the given monthly sales data and the expected frequencies obtained from the fitted ZIP model. The expected frequencies under the ZIP model are close to the observed frequencies, indicating a good fitting between the model and the data.

Table III. Observed and Expected frequencies of ZIP model.

Monthly Sale Count	Observed Frequency	Expected Frequency(ZIP)
0	59	59.00
1	7	6.84
2	2	2.45
3	1	0.71
Total	69	69.00

D. Chi Square Goodness of Fit Test

The chi-square goodness-of-fit test is employed to formally assess whether the fitted ZIP model adequately represents the observed sales data by comparing the observed frequencies with the expected frequencies for the given model and evaluating using the chi-square statistic.

$$\chi^2_{(k-p-1)} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Where ,
k: Number of sales count categories.
p: Number of parameters in the ZIP model.

For Hypothesis:
H_0 : *ZIP model provides an adequate fit to the observed sales data.*
H_1 : *ZIP model does not provide an adequate fit.*

Calculated, $\chi^2_{(1)} = 0.3821$ *and* $p\text{-}value = 0.54$

Since $p - value > 0.05$, the null hypothesis is not rejected and there is no significant difference between observed and expected frequencies. Hence, the ZIP model provides an adequate fit to the observed sales data.

E. Interpretation of Estimated Parameters of ZIP(α, λ) :

$$\hat{\alpha} = 0.2836 \quad and \quad \hat{\lambda} = 0.7154$$

Where,
α: Probability of the months that the car is available for sale
$\approx$ 28% of months car was available.

$1 - \alpha$: Probability of non-availability of the car
$\approx$ 72% of months car was not available.
λ: Expected number of sales when the car is available

Frequency Interpretation Based on 69 Months Sale :

Car was available: $69 \times 0.2836 = 19.5684 \approx$ 20 months
Car was not available: $69 \times 0.7164 = 49.4316 \approx$ 49 months

Expected Sale Count: $E(X) = \hat{\alpha}\hat{\lambda} = 0.2836 \times 0.7154 \approx 0.20$

Average sales count $\approx$ 0.20 Cars per month
$\approx$ 2.4 $\approx$ 2 to 3 Cars per year

F. Interpretation of Sales Pattern:

- Sales are observed only in the months when the car is available in the showroom.
- A large proportion of zero sales per month arises due to non-availability of the car and low customer demand.

Conclusion

The empirical finding indicates that the ZIP model provides a more reliable representation of the observed monthly sales pattern of a selected Black color luxury car than the conventional Poisson model when the overdispersion arises from excess zeros. This improvement is primarily attributed to the ZIP model, which accounts for the presence of excess zero observations within the dataset after the comparison based on Maximum Likelihood Estimates, log-likelihood, AIC and Chi-square goodness-of-fit test. All computational work has been done in RStudio software (version 4.2.2). The findings emphasize that the characteristics of count data should be examined before selecting the probability distributions. The ZIP Distribution can provide better model fitting and more reliable predictions than the conventional Poisson Distribution for sales data with an excess number of zeros. This proposed modeling framework may also be extended to other similar overdispersed count real world data sets.

References

Cameron, A. C., & Trivedi, P. K. (1986). Econometric models based on count data: comparisons and applications of some estimators and tests. *Journal of applied econometrics*, *1*, *29–53*.

Evans, D. A. (1953). *Experimental evidence concerning contagious distributions in ecology*. Biometrika,40(1/2),186-211.https://academic.oup.com/biomet/article/40/1-2/186/338481

Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *Journal of Animal Ecology*, 12(1), https://www.jstor.org/stable/1411

Posit team (2025). RStudio: *Integrated Development Environment for R*. Posit Software, PBC, Boston, MA. URL http://www.posit.co/

Shahsavari, S., Moghimbeigi, A., Kalhor, R., Jafari, A. M., Bagherpour-Kalo, M., Yaseri, M., & Hosseini, M. (2024). Zero-Inflated Count Regression Models in Solving Challenges Posed by Outlier-Prone Data; an Application to Length of Hospital Stay. *Archives of Academic Emergency Medicine*, 12(1), 1–7.https://doi.org/10.22037/aaem.v12i1

Singh, Brijesh P., Maheshwari, S., & Singh, G. (2012). *A Study of Household Size in India through the Count Data Models*. In Singh & Upadhyay (Ed.), Emerging Applications of Bayesian Statistics and Stochastic Modelling (85–90). Mudranik Technologies.

Singh, Brijesh P., Singh, S., & Das, U. D. (2021). Rationalization of the Pattern of Rural Out Migration: An Application of a Composite and Inflated Probability Models. *Asian Journal of Probability and Statistics*, 14(4), 41–51.https://doi.org/10.9734/ajpas/2021/v14i430337

Singh, S.N. (1963): A note on inflated Poisson distribution. *Journal of Indian Statistical Association*, 1,140-144.

***