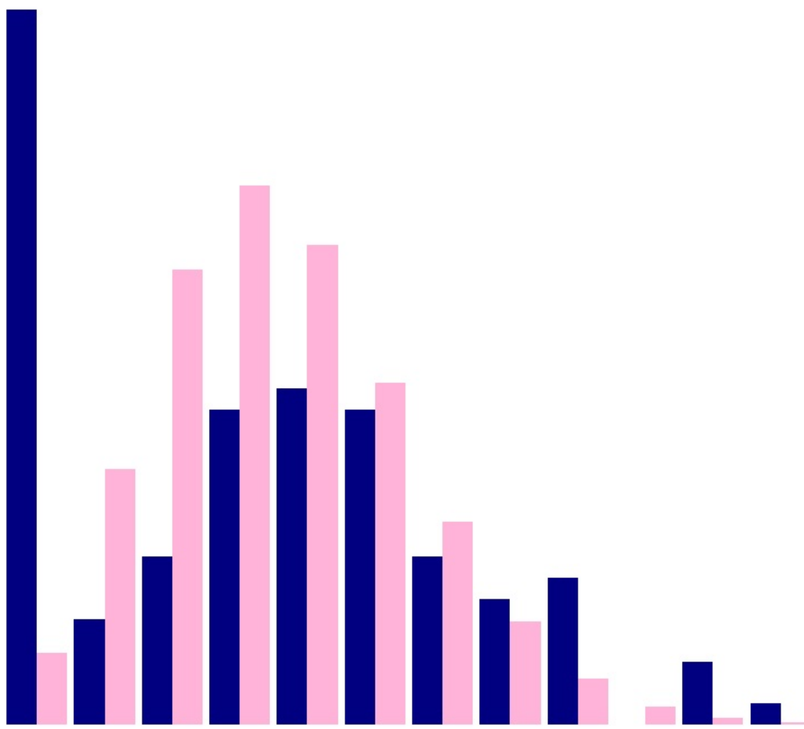
Zero- Inflated Poisson

A modification over Poisson Distribution



Shantanu Nayek
Roll no. : 452
Supervisor : Professor Surupa Chakraborty

I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.

Student's Signature : .... Date :..

Contents

1 Introduction

Given a count data in hand and if the count seems to be unbounded, statisticians are more inclined towards fitting a Poisson Model blindfold in most cases. But this may sometimes lead to erroneous conclusions regarding analysis or other statistical works based on the data. We know that for a Poisson Model, the mean and variance for a sample are very close in general, but for the data it may not be. There may be another problem like if there is a value (say 0 or 1 or 2 or any integer) which has a very high frequency in the sample, then an inflation may be present at the probability curve at that particular value.Based on the above situation, in this study, data on number of seminars/webinars attended by college students during a 2019-2021 was collected. Since it was an unbounded count data, initially we showed interest to fit a Poisson Model for analysis of the data. But it was observed some excess zeroes were there. This presence of excess zeroes, led to thinking of modification of the Poisson Model and we opt to work with the Zero Inflated Poisson distribution. Various measures were taken for justification of fitting a ZIP over a Poisson Model. Some of the justifications are, testing for the parameter π , standard errors for the two models, etc. On giving preference to the ZIP model, I intend to estimate the parameters of the model by Method of Maximum Likelihood and Method of Moments. Finally, there were interests in comparing the efficacy of the estimation procedures.

2 Organization of the paper

2.1 The Problem

Firstly, it is our natural tendency to opt for a Poisson Model for dealing with count data. But count data does not always satisfy a Poisson Model. In case, if we do such assumptions and do further analysis we will lead to some outcomes and observations.But those may not be appropriate. As a result of which the conclusions obtained will be vague and misleading in many situations.So, to increase the consciousness among the statisticians regaring choosing a model for such unbounded count data is a point of interest.
Secondly, in this paper I worked with a real life data on no. of webinars/seminars attended by college students in the span of 2019-2021. In this data fitting a Zero-Inflated Poisson Model over Poisson Model gives a better conclusions rather more appropriate conclusions. Some justifications were shown regarding the selected model over the another.
Thirdly, the parameters of the Zero Inflated Poisson Distribution were estimated by two estimation procedures namely, Method of moments and Method of maximum likelihood.Then a study for determining the better estimation procedure is done in terms of their standard errors.

2.2 Objectives of the paper

The paper has two major objectives. To increase consciousness in choosing a model for a count data correctly. The statisticians are blindfold in fitting a Poisson Model getting a count data in hand , but in practice it is not always perfect. Taking wrong assumptions of statistical model may lead to erroneous conclusions of statistical analyses. As a consequence of such difficulty, an illustrative real life sample is collected and with certain justifications it is shown that fitting of a Poisson Model is not a good job inspite of the data being an unbounded count data. For that a slight modification is done over the Poisson Distribution and fitted again. The modified distribution is Zero -Inflated Poisson Distribution. And we want to study whether there is any improvement while fitting this new model.
There are two parameters of Zero-Inflated Poisson Distribution. They are estimated by two well-known estimation methods , Method of Moments and Method of Maximum Likelihood. The two estimation procedures are compared in terms of their standard errors to conclude which procedure is more efficient for estimation purpose. There will be some difficulties in computing the standard errors, making some assumptions we proceed to derive approximate expressions and finally the exact value of the standard errors for the collected sample.

3 Collection of the data

Motive : In order to handle the misconception that statisticians have regarding fitting of count data , this data is collected . There were certain ideas that lead to the thought that the data may have a zero inflation.

Issue : Data on number of webinars/ seminars attended by college students between 2019-2021 was collected by questionnaire method using Google Forms.

Source : Some of the major sources are :

- St.Xavier's College(Autonomous), Kolkata
- R.K.M. Residential College,Narendrapur
- Haldia Government College
- Techno India University,Kolkata
- St.Stephens College,Delhi
- Calcutta Medical College
- Jadavpur University
- M.U.C. Women's College, Burdwan

Time of Collection : The data was collected in between 28/09/2021. -30/09/2021

Description of the data : No. of webinars/seminars attended by college students during the span of 2019-2021 was collected. The data also consists of the Gender, Academic Degree, Stream of Study, Current Year (as of September 2021).In this paper we are just interested in the count of no. webinars/seminars attended. The other information are collected for future studies(if essential).

No. of respondents : 120

Factors leading to Zero- Inflation : There were a number of factors that led to the thought that the data may be zero inflated -

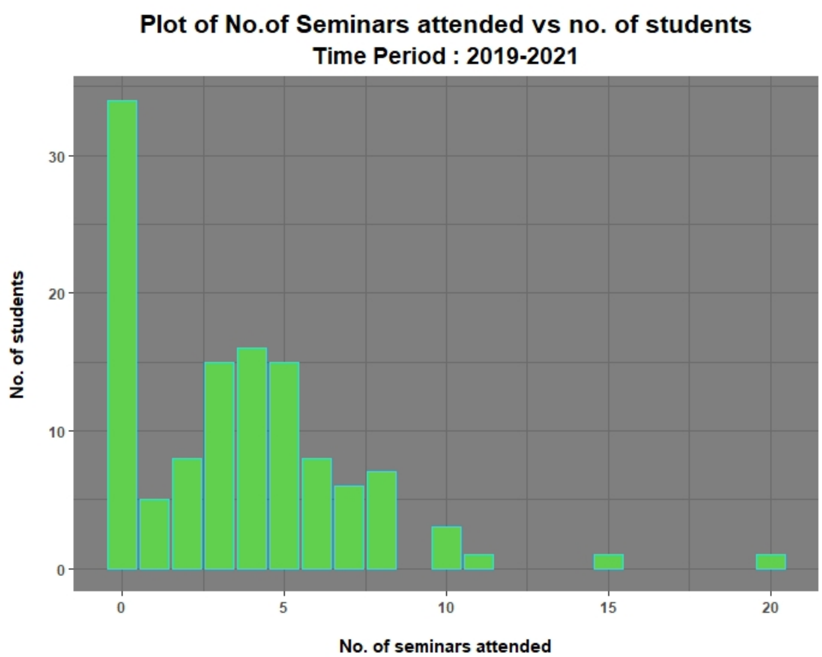
•There may certain colleges where seminars or webinars are not organised for the students that may be due to lack of infrastructure or other reasons.

•Students might not show intention to attend the seminars as he/she might feel to have well preparation on that topic.

•In many cases, students are not interested in filling long forms for attending seminars.

N.B. There may be some other reasons behind zero-inflation. The above mentioned are some of the significant reasons.

Visualisation of the Sample data :



4 Zero-Inflated Poisson Distribution

4.1 Description

Diane Lambert's Zero-inflated Poisson model concerns a random event containing excess zero-count data in unit time. The ZIP distribution with parameters π and λ, is denoted by ZIP (π, λ).

The pmf of zero-inflated Poisson Distribution is given by-

$$P(X=x)=\begin{cases}\pi+(1-\pi)e^{-\lambda}, & \text{if } x=0\\ \\ \frac{(1-\pi)e^{-\lambda}\lambda^x}{x!}, & \text{if } x=1,2,3,4,.....\end{cases}$$

The parameter π gives the extra probability thrust at the value 0; when it vanishes, ZIP (π, λ) reduces to Poisson(λ). The zero-inflated Poisson (ZIP) model mixes two zero generating processes. The first process generates zeros. The second process is governed by a Poisson distribution that generates counts, some of which may be zero.

4.2 Expectation and Variance

$$\begin{aligned}E(X) &= \sum_{x=0}^{\infty}[x.P(X=x)]\\ &= 0.[\pi+(1-\pi).e^{-\lambda}]+(1-\pi).\sum_{x=1}^{\infty}[\frac{x.e^{-\lambda}\lambda^x}{x!}]\\ &= (1-\pi)\lambda\end{aligned}$$

$$Var(X) = E(X^2) - E(X)^2$$

$$Now, E(X^2) = \sum_{x=0}^{\infty}[x^2.P(X=x)]$$

$$= 0.[\pi + (1-\pi).e^{-\lambda}] + (1-\pi).\sum_{x=1}^{\infty}[\frac{x^2.e^{-\lambda}\lambda^x}{x!}]$$

$$= (1-\pi)(\lambda + \lambda^2)$$

$$Var(X) = \lambda(1-\pi)(1+\lambda\pi)$$

4.3 Estimation of ZIP Parameters by Method of moments

Define , $\bar{X} = \frac{\sum_{i=1}^{n} Xi}{n}$, $\quad S^2 = \frac{\sum_{i=1}^{n}(Xi-\overline{X})^2}{n-1}$

Clearly, $(\widehat{1-\pi})\lambda$=$\bar{X}$, $\quad \lambda(\widehat{1-\pi)(1}+\lambda\pi)$= S^2 ....(*)

On solving (*) , we obtain

$\widehat{\lambda_{MME}}$=$\bar{X} + \frac{S^2}{\bar{X}} - 1$ $\quad \widehat{\pi_{MME}}$=$\frac{S^2-\bar{X}}{\bar{X}^2+S^2-\bar{X}}$

Now , the estimate of $\bar{X}$can be greater than S^2. So, $\widehat{\pi_{MME}}$ may take negative values. But , $0 < \pi < 1$.
A slight modification is done over the obtained estimators. The resultant corrected Method of Moments Estimators are :

$$\hat{\lambda} = \begin{cases} \bar{X}, & \text{if } \bar{X} \geq S^2 \\ \\ \widehat{\lambda_{MME}}, & \text{if } \bar{X} \leq S^2 \end{cases}$$

$$\hat{\pi} = \begin{cases} 0, & \text{if } \bar{X} \geq S^2 \\ \\ \widehat{\pi_{MME}}, & \text{if } \bar{X} \leq S^2 \end{cases}$$

4.4 Estimation of ZIP Parameters by Method of maximum likelihood

The likelihood function is given by -

L(π, λ) = $\prod_{i=1}^{n}$P(Xi=xi)

Let, Y be the number of Xi's taking value 0. The likelihood function can be re-written as :

L(π, λ)=$(\pi + (1-\pi)e^{-\lambda})^Y \prod_{i=1(1)n, Xi\neq 0} \frac{(1-\pi)e^{-\lambda}\lambda^x}{x!}$

The log likelihood function is given by :

log(L(π, λ))= $Y log(\pi + (1-\pi)e^{-\lambda}) + (n-Y)log(1-\pi) - (n-Y)\lambda$

$+ n\bar{X}log(\lambda) - log(\prod_{i=1}^{n} Xi)$ ....(**)

Differentiating (**)with respect to λ and equating it to 0 , we obtain :

$\frac{\delta log(L(\lambda,\pi))}{\delta\lambda} = 0$

or, $\frac{n\bar{X}}{\lambda} = (n-Y) + \frac{Y(1-\pi)e^{-\lambda}}{(\pi+(1-\pi)e^{-\lambda})}$ ....(i)

Differentiating (**)with respect to π and equating it to 0 , we obtain :

$\frac{\delta log(L(\lambda,\pi))}{\delta\pi} = 0$

or, $\frac{Y(1-\pi)(1-e^{-\lambda})}{(\pi+(1-\pi)e^{-\lambda})} = n - Y$ ....(ii)

On solving (i) and (ii) , the Maximum likelihood estimators , say $\widehat{\pi_{MLE}}$ and $\widehat{\lambda_{MLE}}$ are obtained by iterative method.

Note :

For our collected sample (n=120), the estimates are tabulated below :

Table 1: Estimates of π and λ by MME and MLE

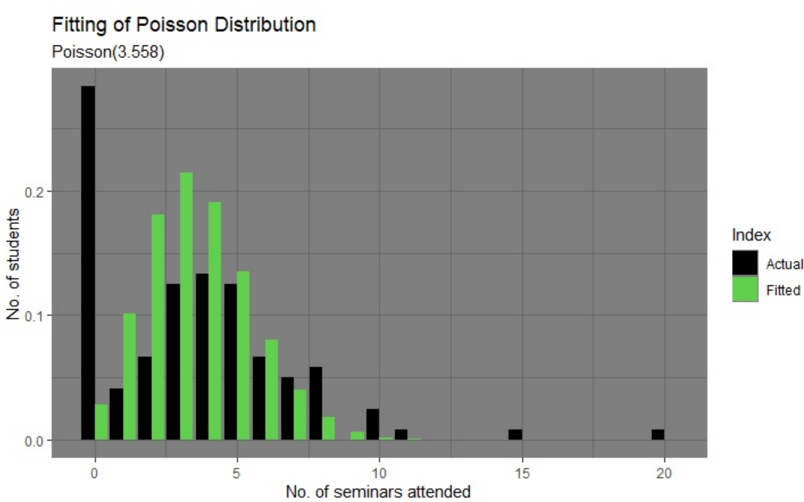
Parameter	Method	Estimate
π	MME	0.3832
π	MLE	0.2781
λ	MME	7.7691
λ	MLE	4.9292

5 Fitting the data

5.1 Fitting of Poisson Distribution

Since, it is an unbounded count data ,we showed our initial interest to fit Poisson distribution.

From the sample , we obtained the method of moments estimate of the parameter λ for fitting Poisson($\hat{\lambda}$). the obtained value of $\hat{\lambda}$ is 3.558.

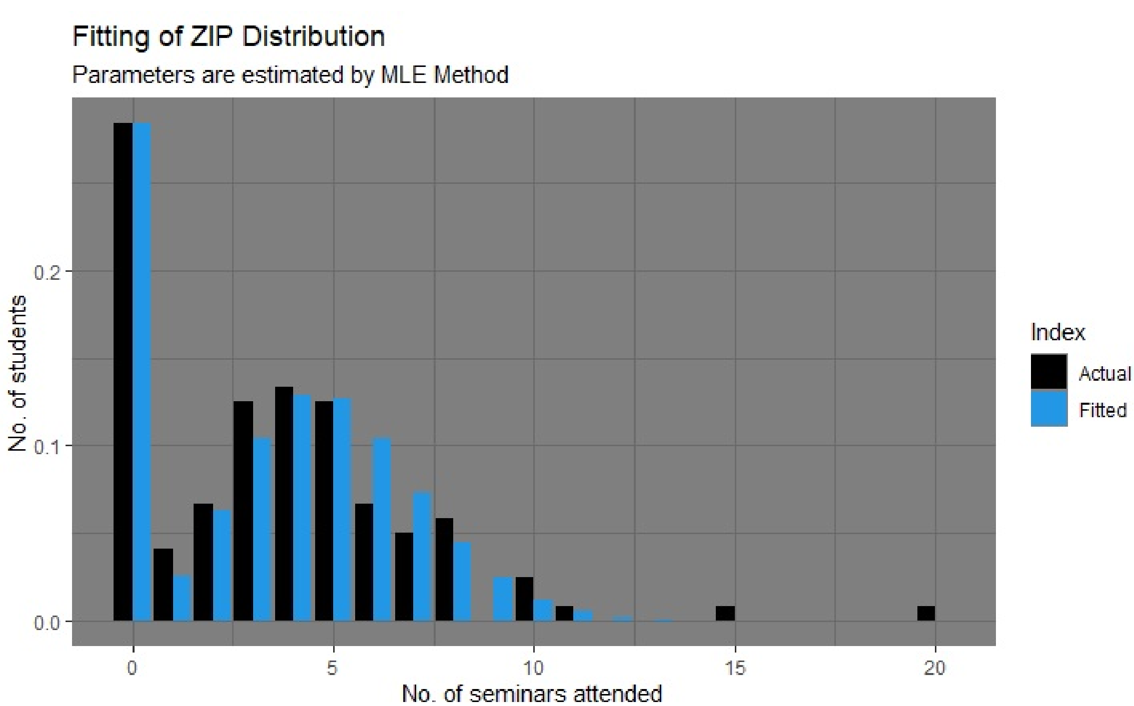


Observation : The fitted probability and the relative frequency at zero point deviates the most. The fit underestimates the probability at 0 and overestimates almost rest of the probabilities.So, we wanted to be concerned a bit about that zero point.

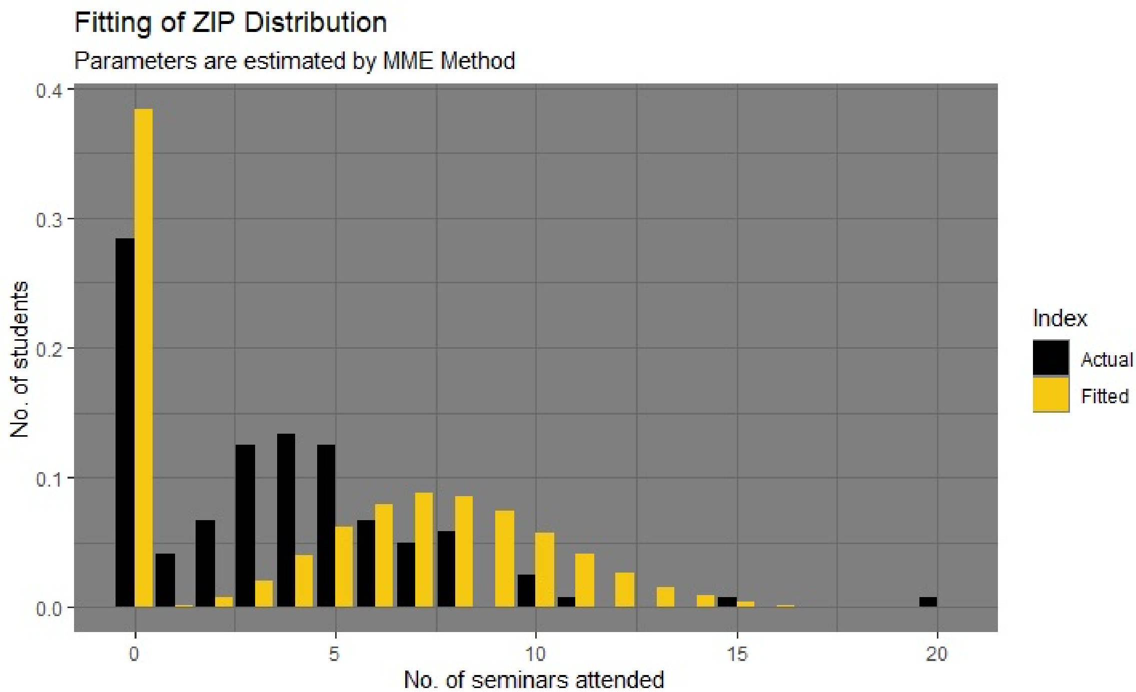
5.2 Fitting of ZIP Distribution

Since, the point zero was underestimated in the above fit, so we showed interest to fit Zero-Inflated Poisson Distribution. Zero Inflated Poisson Distribution is obtained by slight modification over Poisson Distribution , just providing some excess probability thrust to the point zero.

For fitting a ZIP distribution, the parameters are taken as the estimates obtained by Method of Maximum Likelihood (From Table 1).

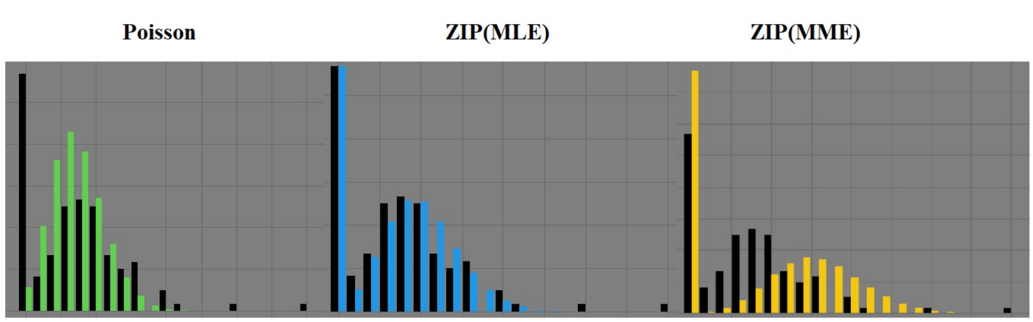


For fitting a ZIP distribution, the parameters are taken as the estimates obtained by Method of Moments(From Table 1).



5.3 Justification of preferring ZIP Distribution

5.3.1 Graphical representation using Column Diagram



Observations :

i) Fitting of ordinary Poisson Distribution underestimates probability at zero point and overestimates the probability in almost all other points.

ii)Fitting of ZIP Distribution , where parameters are estimated by Method of Moments overestimates the probability at zero and underestimates the probabilities in almost all other points.

iii) Apparently, fitting of ZIP Distribution, where the parameters are estimated by Method of Maximum Likelihood gives better results than the previous two cases.

5.3.2 Test of the parameter π

Let , π be the parameter of the Zero - Inflated Poisson Distribution which gives excess probability thurst to the zero point.

To test H_0 **:** π=0 **against** H_1 **:** $\pi \neq$0

Sample : The collected sample from the n(= 120) individuals is the chosen sample for the testing procedure.

Assumption : $\sqrt{n}(\hat{\pi} - \pi) \rightarrow A.N.(0, \sigma*^2)$ A.N. : Asymptotic Normal

Test Statistic : Z = $\frac{\sqrt{n}(\hat{\pi}-\pi)}{\sigma*}$

Under Ho , Z = $\frac{\sqrt{n}(\hat{\pi})}{\sigma*_o}$ where,

$\sigma*_o$ is unknown. So, we take the estimate of $\sigma*_o$ as the standard error of $\hat{\pi}$

obtained by Method of Maximum Likelihood.

Note :

- The estimate of $\hat{\pi}$ is obtained by Method of Maximum Likelihood.
- The Standard Error computation is explained in Section 6.1 in details with the help of Fisher Information Matrix.

Computation : For the sample ,

n=120, $\hat{\pi} = 0.2781128$, $\sigma*_o$ =0.04145

Z(observed)=6.71497 , α=0.05(assuming), $\tau_{0.025} = 1.96$

Decision : Since ,$|Z(observed)| > \tau_{0.025}$, we reject Ho at 5 percent level of significance.

Conclusion : In the light of the given data , it seems that fitting a Zero- Inflated Poisson Distribution is better than fitting a Poisson Distribution.

5.3.3 Comparision of Standard Errors

The standard error of the parameter λ under the assumption of Poisson Distribution is 0.1721. The standard error of the parameter λ under the assumption of Zero - Inflated Poisson Distribution is 0.2429.

It is to be noted that for Poisson Distribution the standard errors obtained by Method of Moments and Method of Maximum Likelihood are same. For , Zero - Inflated Poisson Distribution , we considered the standard error obtained bt Method of Maximum Likelihood.This is due to the fact that Maximum Likelihood Estimation procedure is more efficient than Moment Estimation procedure (discussed in Section 6).

Here, we observe that the standard error in case of Poisson Distribution is lesser than the standard error in case of Zero - Inflated Poisson Distribution. But in Section 5.3.1 and 5.3.2 we had extracted some evidence which gives a nod to Zero-Inflated Poisson Distribution over Poisson Distribution. But it is very interesting that standard error is higher for the Zero - Inflated Poisson Distribution. One of the major reasons of this anomaly is the number of parameters in both the distribution are not same. Hence, it is not well justified to compare standard errors in this situation.

6 Estimation of Standard Errors of the Estimators of ZIP parameters

6.1 Estimation of Standard Errors of Maximum likelihood estimators of π and λ using Fisher Information Matrix

The likelihood function is given by -

L$(\pi, \lambda) = \prod_{i=1}^{n}$P(Xi=xi)

Let, Y be the number of Xi's taking value 0. The likelihood function can be re-written as :

L(π, λ)=$(\pi + (1-\pi)e^{-\lambda})^Y \prod_{i=1(1)n, Xi\neq 0} \frac{(1-\pi)e^{-\lambda}\lambda^x}{x!}$

The log likelihood function is given by :

log(L(π, λ))= $Y log(\pi + (1-\pi)e^{-\lambda}) + (n-Y)log(1-\pi) - (n-Y)\lambda$

$+ n\bar{X}log(\lambda) - log(\prod_{i=1}^{n} Xi)$ ....(**)

Differentiating (**)with respect to λ , we obtain :

$\frac{\delta log(L(\lambda,\pi))}{\delta\lambda} = \frac{n\bar{X}}{\lambda} - (n-Y) - \frac{Y(1-\pi)e^{-\lambda}}{(\pi+(1-\pi)e^{-\lambda})}$ ....(i)

Again, differentiating (i) with respect to λ , we obtain :

$\frac{\delta^2 log(L(\lambda,\pi))}{\delta\lambda^2} = \frac{Y\pi(1-\pi)e^{-\lambda}}{(\pi+(1-\pi)e^{-\lambda})^2} - \frac{n\bar{X}}{\lambda^2}$ ....(ii)

Differentiating (i) with respect to π , we obtain :

$\frac{\delta^2 log(L(\lambda,\pi))}{\delta\lambda\delta\pi} = \frac{Ye^{-\lambda}}{(\pi+(1-\pi)e^{-\lambda})^2}$ ....(iii)

Differentiating (**)with respect to π , we obtain :

$\frac{\delta log(L(\lambda,\pi))}{\delta\pi} = \frac{Y(1-\pi)(1-e^{-\lambda})}{(\pi+(1-\pi)e^{-\lambda})} - \frac{(n-Y)}{(1-\pi)}$ ....(iv)

Again, differentiating (iv) with respect to π , we obtain :

$$\frac{\delta^2 log(L(\lambda,\pi))}{\delta\pi^2} = -\frac{Y(1-e^{-\lambda})^2}{(\pi+(1-\pi)e^{-\lambda})^2} - \frac{(n-Y)}{(1-\pi)^2} \qquad \text{.................................(v)}$$

We define ,

I_{11} = - E($\frac{\delta^2 log(L(\lambda,\pi))}{\delta\lambda^2}$)　　　　I_{22} = - E($\frac{\delta^2 log(L(\lambda,\pi))}{\delta\pi^2}$)

$I_{21} = I_{12}$= - E($\frac{\delta^2 log(L(\lambda,\pi))}{\delta\lambda\delta\pi}$)

The Fisher Information matrix $I(\lambda, \pi)$:

$$I(\lambda,\pi) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \text{, where } I_{11}, I_{12}, I_{21}, I_{22} \text{ are defined.}$$

The inverse of the Fisher Information Matrix $I(\lambda, \pi)$ is:

$$I^{-1}(\lambda,\pi) = \begin{bmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{bmatrix} \qquad \text{(let)}$$

Note : The explicit form of $I^{-1}(\lambda, \pi)$ is not shown here . Generally, for a given sample , we obtain $I(\lambda, \pi)$ explicitly. Then we obtain $I^{-1}(\lambda, \pi)$ by just taking inverse of the finite matrix $I(\lambda, \pi)$.

The standard errors are given by :

S.E.$(\widehat{\lambda_{MLE}}) = \sqrt{I^{11}}$

S.E.$(\widehat{\pi_{MLE}}) = \sqrt{I^{22}}$

6.2 Estimation of Standard Errors of Method of moment estimators of π and λ using Delta Method

The Method of Moment estimates of π and λ :

Let us define , $\bar{X} = \sum_{i=1}^{n} \frac{Xi}{n}$, $S^2 = \sum_{i=1}^{n} \frac{(Xi-\bar{X})^2}{(n-1)}$

Hence , $\widehat{\lambda_{MME}}=\bar{X} + \frac{S^2}{\bar{X}} - 1$ $\widehat{\pi_{MME}}=\frac{S^2-\bar{X}}{\bar{X}^2+S^2-\bar{X}}$ ....(*)

Note :

• The derivation of the above estimators are done in Section 4.3 .

• Since for our sample n=120 , the sample size is quite large. Hence, we approximate (n-1) in the expression of S^2 by n for simpler calculation.

Let us define ,

$\bar{X} = \sum_{i=1}^{n} \frac{Xi}{n} = m_1^{'}$ (let) ....(i)

$S^2 = \sum_{i=1}^{n} \frac{(Xi-\bar{X})^2}{(n-1)} \approx \sum_{i=1}^{n} \frac{Xi^2}{n}$ - $\bar{X}^2 = m_2^{'} - (m_1^{'})^2$ (let) ....(ii)

Combining equation (i) and (ii) with (*) , we obtain ,

$\widehat{\pi_{MME}}$ = 1 - $\frac{(m_1^{'})^2}{m_2^{'}-(m_1^{'})}$ (let) ....(iii)

$\widehat{\lambda_{MME}}$ = $\frac{(m_2^{'})}{(m_1^{'})}$ - 1 (let) ....(iv)

For simplicity, let

$m_1^{'} = T_1$ $m_2^{'} = T_2$.

Hence , we can write in general ,

$\widehat{\lambda_{MME}}$ = f($m_1^{'}, m_2^{'}$) $\widehat{\pi_{MME}}$ = g($m_1^{'}, m_2^{'}$)

i.e, $\widehat{\lambda_{MME}}$ = f(T_1, T_2) $\widehat{\pi_{MME}}$ = g(T_1, T_2)

We observe ,

Differentiating (iv) with respect to T1 ,

$$\frac{\delta f(T_1,T_2)}{\delta_{T1}} = \frac{-T_2}{T_1^2}$$

Differentiating (iii) with respect to T2 ,

$$\frac{\delta g(T_1,T_2)}{\delta_{T2}} = \frac{T_1^2-2T_1T_2}{(T_2-T_1)^2}$$

Differentiating (iv) with respect to T2 ,

$$\frac{\delta f(T_1,T_2)}{\delta_{T2}} = \frac{1}{T_1}$$

Differentiating (iii) with respect to T1 ,

$$\frac{\delta g(T_1,T_2)}{\delta_{T1}} = (\frac{T_1}{T_2-T_1})^2$$

Note:

Define , $\mu_r^{'} = \frac{\sum_{i=1}^{n} x_i^r}{n}$

$$\mathrm{E}(m_r^{'}) = \mu_r^{'}$$

$$\mathrm{Var}(m_r^{'}) = \frac{\mu_{2r}^{'}-(\mu_r^{'})^2}{n}$$

$$\mathrm{Cov}(m_r^{'}, m_q^{'}) = \frac{\mu_{r+q}^{'}-(\mu_r^{'}\mu_q^{'})}{n}$$

Using Delta Method ,

- $\text{Var}(f(T_1,T_2)) = \text{Var }(T_1)[\frac{\delta f(T_1,T_2)}{\delta_{T1}}]^2 + \text{Var }(T_2)[\frac{\delta f(T_1,T_2)}{\delta_{T2}}]^2 +$

$$2 \text{ Cov } (T_1,T_2) \, [\frac{\delta f(T_1,T_2)}{\delta_{T1}} \frac{\delta f(T_1,T_2)}{\delta_{T2}}]$$

$$= \frac{(\mu_2^{'}-(\mu_1^{'})^2)}{n}(\frac{-\mu_2^{'}}{(\mu_1^{'})^2})^2 + \frac{(\mu_4^{'}-(\mu_2^{'})^2)}{n}(\frac{1}{\mu_1^{'}})^2 + \frac{2(\mu_3^{'}-\mu_2^{'}\mu_1^{'})}{n}(\frac{-\mu_2^{'}}{(\mu_1^{'})^3})$$

- $\text{Var}(g(T_1,T_2)) = \text{Var }(T_1)[\frac{\delta g(T_1,T_2)}{\delta_{T1}}]^2 + \text{Var }(T_2)[\frac{\delta g(T_1,T_2)}{\delta_{T2}}]^2 +$

$$2 \text{ Cov } (T_1,T_2) \, [\frac{\delta g(T_1,T_2)}{\delta_{T1}} \frac{\delta g(T_1,T_2)}{\delta_{T2}}]$$

$$=\frac{(\mu_2^{'}-(\mu_1^{'})^2)}{n}(\frac{(\mu_1^{'})^2-2\mu_1^{'}\mu_2^{'}}{(\mu_2^{'}-\mu_1^{'})^2})^2 + \frac{(\mu_4^{'}-(\mu_2^{'})^2)}{n}(\frac{\mu_1^{'}}{\mu_2^{'}-\mu_1^{'}})^4 + \frac{2(\mu_3^{'}-\mu_2^{'}\mu_1^{'})}{n}\frac{\mu_1^2((\mu_1^{'})^2-2\mu_1^{'}\mu_2^{'})}{(\mu_2^{'}-\mu_1^{'})^2}$$

Hence , the estimate of the standard errors are :

$S.E.\widehat{(\lambda_{MME})} = \sqrt{Var(f(T_1,T_2))}$

$S.E.\widehat{(\pi_{MME})} = \sqrt{Var(g(T_1,T_2))}$

6.3 Tabular representation of outputs

Table 2: Summary of the estimators of π and λ

Parameter	Method	Estimate	S.E.	C.L.(lower)	C.L.(upper)
π	MME	0.3832	0.063	0.3832 - 1.96 * 0.063	0.3832 + 1.96 * 0.063
π	MLE	0.2781	0.041	0.2781 - 1.96 * 0.041	0.2781 + 1.96 * 0.041
λ	MME	7.7691	0.732	7.7691 - 1.96 * 0.732	7.7691 + 1.96 * 0.732
λ	MLE	4.9292	0.242	4.9292 - 1.96 * 0.242	4.9292 + 1.96 * 0.242

Note :

- S.E. : Standard Error.
- C.L. : Confidence Limits.
- The confidence limits are computed taking $\alpha = 0.05$.
- $\tau_{0.025} = 1.96$ [approximately]

6.4 Observation

- The standard error of $\lambda_{\widehat{MME}}$ is higher than of $\lambda_{\widehat{MLE}}$.

- The standard error of $\pi_{\widehat{MME}}$ is higher than of $\pi_{\widehat{MLE}}$.

7 Conclusion

From the above study , we can arrive at two major conclusions. Firstly , for a count data fitting of a Poisson Distribution may not always work good. So, getting a count data in hand we should always confirm that there is no unusual inflation at a point. There may be other factors also which may deviate from Poisson Distribution. More specifically , if there is a inflation at zero point assuming all other factors being usual, we may show our interest in fitting a Zero - Inflated Poisson Distribution over Poisson Distribution.

Secondly, considering Zero - Inflated Poisson Distribution , we computed the standard errors of the method of moments estimators and maximum likelihood estimators of the parameters. We observed that the standard errors of the Maximum Likelihood Estimators are lesser than standard errors of the Method of Moments Estimators for both situations. Moreover we observed that fitting of Zero - Inflated Poisson Distribution where estimate of the parameters are obtained by Method of Maximum

Likelihood is better than fitting of Zero - Inflated Poisson Distribution where estimate of the parameters are obtained by Method of Moments. Hence, as a whole among two estimation procedures , it seems that Method of Maximum Likelihood is more efficient than Method of Moments.

8 Acknowledgement

I would like to express my gratitude to my supervisor , respected Professor Surupa Chakraborty , who has given me this great opportunity to work on this topicand gather deeper knowledge and skills . Without her constant support and guidance the completion of this project work would not have been possible. I would also like to thank all my respected professors who have always guided and encouraged me throughout the course. I am grateful to all my friends who helped me a lot in gathering information and transforming my ideas into reality. Lastly , but not the least , I am in debted to my family for their immense love , support and guidance.

9 Appendix

Code 1: R Code for the Paper

```
#---------------------------------------------
#---------------------------------------------

            #READING THE EXCEL SHEET

setwd("X:/")
getwd()
data=read.csv("Data Dissertation.csv");data

#--------------------------------------------
#--------------------------------------------

             #Calling Libraries

library(ggplot2)
library(VGAM)
library(tidyr)

#--------------------------------------------
#--------------------------------------------

              #Data Visualisation

table(data$Count)
table(data$Stream)

#Defining theme for plots
theme_new=theme(

  plot.title =element_text(size=16,
                           hjust=0.5,face="bold"),

  plot.subtitle =element_text(size=14,
                              hjust=0.5,face="bold"),

  legend.title=element_text(size=10,hjust=0.5,
```

```
                              face="bold.italic"),

  axis.title=element_text(face="bold"),

  axis.text=element_text(face="bold")

)

z=data$Count
ggplot(NULL,aes(z))+
  geom_bar(col=5,fill=3)+theme_dark()+
  theme_new+
  labs(title="Plot of No.of Seminars attended
       vs no. of students",
       subtitle="Column Diagram",
       x="\nNo. of seminars attended",
       y="No. of students\n",
       col="Index")

x=0:20
y=c(34,5,8,15,16,15,8,6,7,0,3,1,0,0,0,1,
    0,0,0,0,1)
sum(y)
A=cbind(x,y);A

#---------------------------------------------------------
#---------------------------------------------------------
              #METHOD OF MOMENTS ESTIMATES

count=rep(x,times=y);count
table(count)
              #MEAN AND VARIANCE OF THE COLLECTED SAMPLE
sample_mean=mean(count);sample_mean

sample_variance=var(count);sample_variance

#---------------------------------------------------------
#---------------------------------------------------------

           #ESTIMATES  _ MME _ZIP _PIE ,LAMBDA

```

```
lambda_MME=sample_mean+(sample_variance/sample_mean)+1
lambda_MME

pie_MME=(sample_variance-sample_mean)/
  (sample_mean^2+
     (sample_variance-sample_mean));

#---------------------------------------------------------
#---------------------------------------------------------

             #ESTIMATES  _ MLE _ZIP _PIE ,LAMBDA

count
n=length(count)

y=length(count[count==0]);y   #NUMBER OF ZEROES IN THE SAMPLE

sample_no.zero=count[count!=0];sample_no.zero #THE
                       #SAMPLE EXCLUDING THE ZEROES
x=sample_no.zero
count

                 #DEFINING THE LIKELIHOOD FUNCTION

Likelihood=function(l,p)
{

  -sum(dzipois(count,l,p,log=T))
}

k=mle(Likelihood,start=list(l=lambda_MME,p=pie_MME))
summary(k)

#-------------------------------------------------------
#-------------------------------------------------------

         #MLE ESTIMATES USING POISSON DISTRIBUTION

Likelihood=function(l)
{

```

```
  -sum(dpois(count,l,log=T))
}

k=mle(Likelihood,start=9) #9 is a arbitrary choice
k

#-------------------------------------------------------
#-------------------------------------------------------

            #FITTING OF ZIP AND POISSON DISTRIBUTION

fit_POISSON_MLE=dpois(seq(0,20,1),lambda_POISSON_MLE);fit_
   POISSON_MLE

fit_ZIP_MLE=dzipois(seq(0,20,1),lambda_MLE,pie_MLE);fit_ZIP_
   MLE

fit_ZIP_MME=dzipois(seq(0,20,1),lambda_MME,pie_MME);fit_ZIP_
   MME

y=c(34,5,8,15,16,15,8,6,7,0,3,1,0,0,0,1,0,0,0,0,1)
ACTUAL_prob=y/sum(y);ACTUAL_prob             #ACTUAL DATA

                      #Graph of Poisson fitting

datat2=data.frame(0:20,Actual=ACTUAL_prob,Fitted=fit_POISSON_
   MLE)
data2=pivot_longer(data2,-1,"Index",values_to="v")
data2
ggplot(data2,aes(X0.20,v,fill=Index))+
  geom_bar(position="dodge",stat="identity")+
  labs(title="Fitting of Poisson Distribution",
       subtitle="Poisson(3.558)",x="No. of seminars attended",
       y="No. of students")+theme_dark()+
  scale_fill_manual(values=c(9,3))


               #Graph of ZIP Fitting by MLE Method


data3=data.frame(0:20,Actual=ACTUAL_prob,Fitted=fit_ZIP_MLE)
```

```
data3=pivot_longer(data3,-1,"Index",values_to="v")
data3
ggplot(data3,aes(X0.20,v,fill=Index))+
  geom_bar(position="dodge",stat="identity")+
  labs(title="Fitting of ZIP Distribution",
       subtitle="Parameters are estimated by MLE Method",
       x="No. of seminars attended",
       y="No. of students")+theme_dark()+
  scale_fill_manual(values=c(9,7))

               #Graph of ZIP Fitting by MME Method

data4=data.frame(0:20,Actual=ACTUAL_prob,Fitted=fit_ZIP_MME)
data4=pivot_longer(data4,-1,"Index",values_to="v")
data4
ggplot(data4,aes(X0.20,v,fill=Index))+
  geom_bar(position="dodge",stat="identity")+
  labs(title="Fitting of ZIP Distribution",
       subtitle="Parameters are estimated by MME Method",
       x="No. of seminars attended",
       y="No. of students")+theme_dark()+
  scale_fill_manual(values=c(9,7))

#-------------------------------------------------------
#-------------------------------------------------------

        #Computation of SE from Fisher Information Matrix
      #Let lambda=l ,pie=p ,Y= no. of Zeroes ,n=sample size

lambda_MLE=4.9292075;lambda_MLE
pie_MLE=0.2781128;pie_MLE

l=lambda_MLE
p=pie_MLE

n=120
Y=34
Xbar=(1-p)*l

I11=((Y*p*(1-p)*exp(-l))/(p+(1-p)*exp(-l))^2)-((n*Xbar)/l^2)
I12=(Y*exp(-l))/(p+(1-p)*exp(-l))^2
```

```
I22=-((Y*(1-exp(-l))^2)/(p+(1-p)*exp(-l))^2)-((n-Y)/(1-p)^2)

FISMAT=matrix(c(-I11,-I12,-I12,-I22),byrow=T,nrow=2)
A=solve(FISMAT)
A
SE_pie=sqrt(0.00171)
SE_lambda=sqrt(0.059)
SE_pie
SE_lambda

#-------------------------------------------------------
#-------------------------------------------------------

       #Computation of SE of MME using Delta Method

x=0:20
f=c(34,5,8,15,16,15,8,6,7,0,3,1,0,0,0,1,0,0,0,0,1)

n=sum(f)
u1=sum(x*f)/sum(f);u1
u2=sum((x^2)*f)/sum(f);u2
u3=sum((x^3)*f)/sum(f);u3
u4=sum((x^4)*f)/sum(f);u4

var_t1=(u2-u1^2)/n
var_t2=(u4-u2^2)/n
cov_t1t2=(u3-u1*u2)/n

dl.dt1=-(u2/u1^2)
dl.dt2=1/u1

dp.dt1=(u1^2-2*u1*u2)/(u2-u1)^2
dp.dt2=(u1/(u2-u1))^2

var_pie.hat=var_t1*(dp.dt1)^2+
  var_t2*(dp.dt2)^2+
  2*cov_t1t2*(dp.dt1)*(dp.dt2);var_pie.hat

SE_piehat=sqrt(var_pie.hat)
SE_piehat

```

```

var_lambda.hat=var_t1*(dl.dt1)^2+
  var_t2*(dl.dt2)^2+
  2*cov_t1t2*(dl.dt1)*(dl.dt2);var_lambda.hat

SE_lambdahat=sqrt(var_lambda.hat)
SE_lambdahat

#---------------------------------------------------------
#---------------------------------------------------------
```