Journal of Statistical and Econometric Methods, vol.5, no.3, 2016, 61-70
ISSN: 1792-6602 (print), 1792-6939 (online)
Scienpress Ltd, 2016

On The Use of Truncated Zero Inflated Binomial (ZIB) Control Chart for Monitoring Tuberculosis Disease

Edokpa Idemudia Waziri[1] and Odunayo Joseph Braimah[2]

Abstract

The design of various statistical methods for monitoring rare health events shows significance of the issue in health sectors. Rare health events, as attribute quality characteristics cannot be monitored by ordinary Shewhart np charts since overdispersion occurs. A good approach to this problem is the use of control charts based on zero inflation in a binomial (ZIB) distribution. In this distribution, it is assumed that random shocks occur with some probability, and upon the occurrence of such random shocks, health event failures can be found, such that the number of failures in each sampling subgroup follows a binomial distribution. This study develops a truncated ZIB control chart applying probability limits in Shewhart based control limits for monitoring ZIB distributed observations. As the most widespread criteria, Average Run Length approach is used to evaluate the performance of this chart. The use of truncated zero inflation in a binomial (TZIB)

[1] Department of Mathematics, Ambrose Alli University, Ekpoma.
E-mail: jalowaziri@gmail.com
[2] Department of Mathematics, Ambrose Alli University, Ekpoma.
E-mail : ojbraimah2012@gmail.com

Article Info: *Received* : July 14, 2016. *Revised* : August 18, 2016.
Published online : September 1, 2016.

control chart is also investigated by a real case study, using the number of patients who undergone Tuberculosis treatment and later resulted to Drug Resistant Tuberculosis (DRTB) in General Hospital, Igarra, Akoko-Edo Local Government of Edo State. Results are compared with the traditional number of proportion defective (np) chart.

Keywords: Rare Health Events; Truncated Distribution; Tuberculosis; Upper Control Limit; Lower Control Limit; Center line.

1 Introduction

The usual control charts are often used in health engineering for evaluating hospitals performances and improvements. In addition, a number of special statistical methods have been developed for monitoring health services solely; see [1], [2]. Some quality characteristics which the researchers are recently interested in monitoring are infection rates, rates of patient falls, number of congenital malformations in a society, various sorts of waiting times as given by [3].

A good number of the statistical methods have been designed based on this type of interested quality characteristic observations (attribute or variable). For example, in order to monitor the incidence rate of a rare health event, like congenital malformation, different methods have been proposed. Among these are the $g-type$ CUSUM control charts, as in [4] and the Bernoulli CUSUM given by [5]. However, [5] believes that the performance of Bernoulli CUSUM and g-type CUSUM are better methods.

When the traditional np-chart is applied to aggregated data, the number of failures in each subgroup, based on a binomial distribution would be monitored. A situation that is becoming more and more common in the field of high quality industries and also health engineering services is the occurrence of a large number

of zero failures. In industrial terminology, this condition is named "high yields processes" al [6], while in health engineering services, it is called "rare health events"[2], [5]. It can be observed that when there are large numbers of zero data in the attribute quality characteristic, [2] asserted that an overdispersion occurs and the related distribution does not fit a binomial distribution any more. So, some alternative models should be developed. Such situation can occur for Poisson distribution, which for the first time the suitable related model was developed by [7] named "Zero Inflated Poisson" (ZIP) model. However that model was used as a response regression model. In fact, np-chart often underestimates the observed dispersion, resulting in calculation of improper narrow (tighter) control limits; subsequently leading to a higher false alarm rate in detecting out-of-control signals. Hence, modifying the basic binomial distribution to one, which could be used to model larger dispersion, can be developed. This model can be based on zero Inflation in binomial (ZIB) distribution.

2 Materials and Methods

2.1 Zero Inflated Binomial Distribution

Binomial distribution as an attribute data generating distribution is widely used in monitoring quality characteristics, both in industry and healthcare. When there are large number of zero data for an attribute quality characteristic (like the number of infected patients of a sampling subgroup in a hospital), the distribution does not fit any binomial distribution and a combination of zero-inflation and binomial distribution should be developed. The suited probability distribution function for such situation is as given in equation 1. In this equation, it is assumed that random shocks (for example, special season, incidence of an epidemic disease, Changing surgery procedures e.t.c) occur with probability P, and upon the occurrence of such random shocks, failures can be found and X number of failures

(like the number of infected patients of a n member sampling subgroup in a hospital) in each subgroup follows a binomial distribution with parameter p.

$$f_X(x;\theta,P) = f(x) = \binom{n}{x} P^x\,(1-\theta)^{n-x}\,, x = 1,2,\dots,n \qquad (1)$$

Since for all X values, $f_X\,(x_i\,;\,\theta,p)$ and $\sum_{X=1}^{n} f_X\,(x_i\,;\,\theta,p) = 1$, this function is exactly a probability distribution function.

The parameters $p\ and\ \theta$ moment method estimates (MME) for the distribution can be obtained as equation 4.

$$\begin{cases} E(X) = \frac{\sum_{t=1}^{m} X_t}{m} = \mu \\ E(X^2) = \frac{\sum_{t=1}^{m} X_t^2}{m} = \sigma^2 - \mu^2 \end{cases} \qquad (2)$$

$$\Rightarrow \begin{cases} np\theta = E(X) = \frac{\sum_{t=1}^{m} X_t}{m} \\ \theta.n(n-1)p^2 + np\theta = \frac{\sum_{t=1}^{m} X_t^2}{m} \end{cases} \qquad (3)$$

$$\Rightarrow \begin{cases} \hat{p} = \frac{\sum_{t=1}^{m} X_t^2 - \sum_{t=1}^{m} X_t}{(n-1)\sum_{t=1}^{m} X_t} \\ \hat{\theta} = \frac{(n-1)(\sum_{t=1}^{m} X_t)^2}{nm(\sum_{t=1}^{m} X_t^2 - \sum_{t=1}^{m} X_t)} \end{cases} \qquad (4)$$

Using an statistical test with a suitable significance level for the null hypothesis $H_0: \rho_{\hat{P},\hat{\theta}} = 0$ Versus $H_1: \rho_{\hat{P},\hat{\theta}} \neq 0$ cannot be rejected.

Therefore, we can have sensitivity analysis by shifting p and θ separately or simultaneously to calculate the Average Run lengths (ARLs) and evaluate the performance of the control chart.

2.2 Zero Inflated Binomial Control Charts

On the basis of general Shewhart control chart principles, if w is a statistic that measures the quality characteristic, and if mean and variance of w are given as $\mu_w$and σ_w^2 respectively, then the general model for the Shewhart control chart is given by [8] as:

$$Upper\ Control\ Limit\ = UCL = \hat{\mu}_w + L\hat{\sigma}_w \qquad (5)$$

$$Center\ Line = \hat{\mu}_w \tag{6}$$

$$Lower\ Control\ Limit\ = UCL = \hat{\mu}_w - L\hat{\sigma}_w \tag{7}$$

where L is the distance of the control limits from the center line, in multiples of the standard deviation of w. Considering w as the number of failures distributed as (1), we obtain trial control limits as:

$Upper\ Control\ Limit\ = UCL_{ZIB} =$

$$= n\hat{P}\hat{\theta} + L[n(n-1)\hat{P}^2\hat{\theta} + n\hat{P}\hat{\theta} - n^2\hat{P}^2\hat{\theta}^2]^{0.5} \tag{8}$$

$$Centre\ Line = CL_{ZIB} = n\hat{P}\hat{\theta} \tag{9}$$

$$Lower\ Control\ Limit\ = LCL_{ZIB} = n\hat{P}\hat{\theta} + L[n(n-1)\hat{P}^2\hat{\theta} + n\hat{P}\hat{\theta} - n^2\hat{P}^2\hat{\theta}^2]^{0.5} \tag{10}$$

For different P and $\hat{\theta}$ values, LCL_{ZIB} would never get any positive values. In addition, even increasing sample size n not only does not shift LCL_{ZIB} to positive values, but also shifts it to smaller negative ones.

In a high quality process, especially rare health events, where there are large numbers of zero data as failures, we can consider that random shocks occur with probability θ, and upon the occurrence of such shocks, failures can be found, and X number of failures in each sampled subgroup follows a binomial distribution with parameter p . In order to monitor and control such a process, for phase I, we can take m subsequent subgroups including n samples in each, computing $\hat{P}and\ \hat{\theta}$ from Equation 4, as the estimates of parameters p and θ.

As a strategy to define sample size n, in real health engineering cases, in order not to lose any data, usually all events in a predefined time interval is considered as a subgroup sample, so sample size is defined inherently, not selecting only some observations from the process.

2.3 Control chart performance

ARL is the average number of points that must be plotted until a point indicates an out-of-control condition. If the process observations are uncorrelated, then for any Shewhart control chart, the ARL can be calculated easily from

$$ARL = \frac{1}{\text{Probability (One point plots out of control)}} \tag{11}$$

If the observations plotted on the control chart are independent, then the number of points that must be plotted until the first point exceeds a control limit is a geometric random variable with parameter p . The mean of this geometric distribution is simply $\frac{1}{p}$, named the average run length (ARL).

Since zero values for X (as the number of failures) are excessive and also desirable and the positive values for X are undesirable, to calculate ARL we can concentrate only on X positive values (not zero ones) as a truncated distribution shown in Equation 12.

$$ARL^{*} = \frac{\text{Number of positive values, equal to or greater than } \text{UCL}_{\text{ZIB}}}{\text{Number of positive X}} \tag{12}$$

The data analysis was carried out using statistical software MINITAB 13 and R.

3 Result and Data Analysis

Martone and Gaynes[9] and Benneyan [10] reported that infection rates of patients in hospitals have been one of the most important quality attribute characteristics. As a motivating case, in the hospital, this study considered those patients who complained of cough and undergone a tuberculosis test and were diagnosed positive. Therefore, for X and the two parameters n and m, we have:
$n =$ Total number of patients who undergone Tuberculosis treatment and later resulted to drug resistant tuberculosis (DR TB) every day (sample size).

$m =$ Total number of consecutive sampling days in estimating ZIB parameters P and θ.

$X_t =$: Every day number of TB patients who undergone treatment and later resulted to Drug resistant TB.

Record of data gathered for 100 consecutive days is presented in Table 1. The numbers of patients who undergone treatment and later resulted to drug resistant TB on daily basis is different from 43rd to 53rd sample point. Based on [8], since the samples have small variation in size, we can use average numbers of sample sizes for our case calculations. So, the sample size is considered equal to 48 in the above case study.

Table 1: Daily number of patients diagnosed of Tuberculosis and later resulted to Drug Resistance Tuberculosis (DR TB) during treatment for 100 consecutive days

t	X_t	t	X_t	t	X_t	t	X_t	t	X_t	t	X_t	t	X_t	t	X_t	t	X_t	t	X_t
1	1	11	1	21	1	31	1	41	0	51	0	61	0	71	0	81	1	91	0
2	0	12	0	22	0	32	0	42	1	52	0	62	0	72	2	82	0	92	0
3	0	3	0	23	0	33	0	43	2	53	2	63	1	73	0	83	0	93	0
4	1	14	1	24	0	34	0	44	0	54	0	64	0	74	1	84	0	94	0
5	0	15	0	25	0	35	0	45	0	55	0	65	0	75	0	85	0	95	0
6	0	16	1	26	0	36	0	46	0	56	0	66	0	76	0	86	1	96	2
7	0	17	0	27	0	37	0	47	0	57	0	67	0	77	0	87	0	97	0
8	0	18	0	28	0	38	0	48	0	58	0	68	0	78	0	88	0	98	0
9	0	19	0	29	0	39	0	49	0	59	0	69	1	79	0	89	0	99	0
10	0	20	0	30	0	40	0	50	0	60	0	70	1	80	0	90	0	100	0

From equation 4, and the presented data in Table 1 above, we obtained the values of $\hat{P} = 0.0114$ and $\hat{\theta} = 0.4375$ respectively. At $\alpha = 0.05$ level of significant, using equation 8, the Upper Control Limit (UCL_{ZIB}) was obtained to be 2.

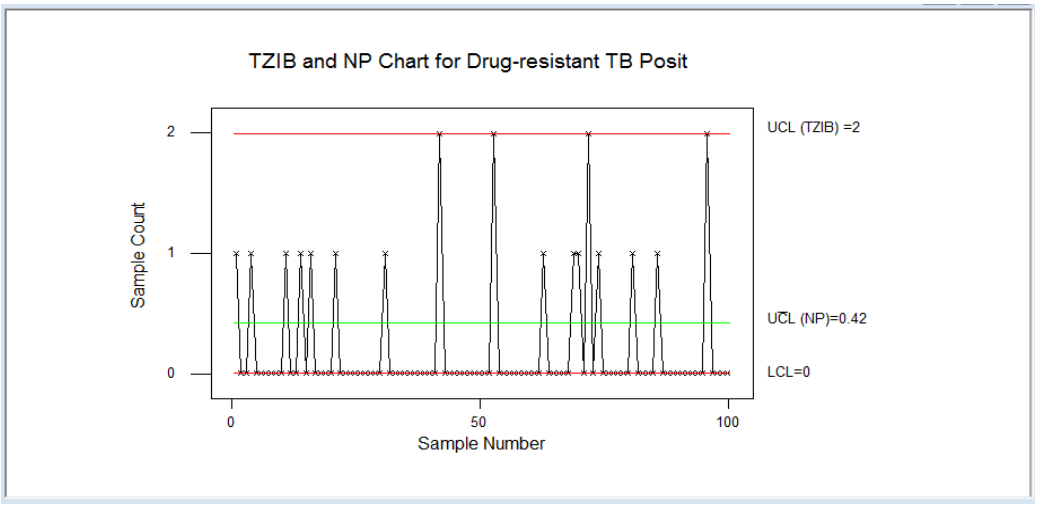

Figure. 1: Comparative plot of TZIB and NP Control chart

The Average Runs Length (ARL) curve related to the aforementioned case study is computed by simulation and depicted in Figure 2 below.

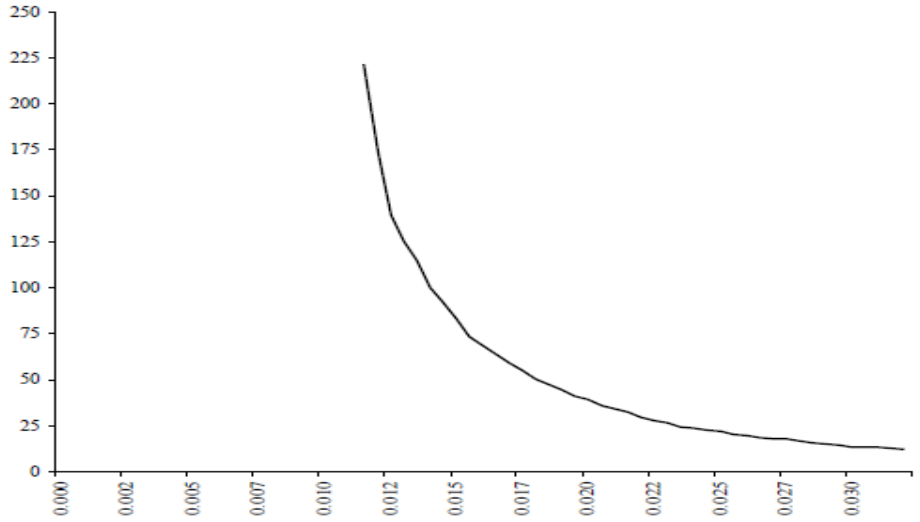

Figure 2: Average Run Length (ARL) curve for TZIB chart with in-control parameters $\widehat{P} = 0.0114$ and $\hat{\theta} = 0.4375$, and average sample size 48

4 Results

The upper control limit for both truncated zero inflated binomial (TZIB) and number proportion (np) control chart is plotted and displayed in figure 1. As we can see in figure 1 above, in comparing np-chart with TZIB, np-chart is less efficient; showing many of the observations above its upper control limits leading to many false alarms. As the most prevalent criteria, the average run length approach is used to evaluate the performance of the TZIB control chart. Therefore, since zero values for X (as the number of failures) are desirable and the positive values for X are undesirable, to calculate the ARL we made use of only X positive values (not zero ones) leading to calculation of ARL.

5 Conclusions

Using both charts for monitoring such data set, the process is said to be in statistical control applying truncated zero inflated binomial (TZIB) control chart while the process is out of statistical control when the data set is configured in number proportion (np) control chart.

Conclusively, truncated zero inflated binomial (TZIB) may be preferably considered when minimizing the rate of false alarm when monitoring the process mean.

References

[1] Sonesson, C. and Bock, D., A review and discussion of prospective statistical surveillance in public health, *Journal of the Royal Statistical Society* **166**, (2003), 5–21.

[2] Woodal, W.H., The Use of Control Charts in Health-Care and Public-Health Surveillance, *Journal of Quality Technology*, **38**, (2006), 88-103.

[3] Lee, K.Y. and McGreevey, C., Using Control Charts to Assess Performance Measurement Data, *Journal on Quality Improvement*, **28**, (2002), 90-101.

[4] Benneyan, J.C., Number-Between g-Type Statistical Quality Control Charts for Monitoring Adverse Events, *Health Care Management Science*, **4**, (2001a), 305-318.

[5] Sego, L.H. Woodall, W.H. and Reynolds, M.R., A comparison of surveillance methods for small incidence rates, *Statistics in Medicine*, **28**, (2008), 1225-1247.

[6] Noorossana R, Saghaei A and Paynabar K., On the conditional decision procedure for high yield processes, *Computers & Industrial Engineering*, **53**, (2007), 469-477.

[7] Lambert, D., Zero-infated Poisson regression with application to defects in manufacturing, *Technometrics* **34**, (1997), 1-14.

[8] Montgomery, D.C., *Introduction to Statistical Quality Control*, (5th edn). Wiley, New York, 2005.

[9] Martone W.J. and Gaynes R.P., Nosocomial infection rates for interhospital comparison: limitations and possible solutions. Infect Control Hosp Epidemiol **12**, (1991), 609-621.

[10] Benneyan, J.C., Statistical Quality Control Methods in Infection Control and Hospital Epidemiology, Part I: Introduction and Basic Theory. *Infection Control and Hospital Epidemiology*, **19**, (1998a), 194-214.