TRUNCATED ZERO INFLATED BINOMIAL CONTROL CHART FOR MONITORING RARE HEALTH EVENTS

Amir Afshin Fatahi*, Rassoul Noorossana, Pershang Dokouhaki & Massoud Babakhani

Industrial Engineering Department, Iran University of Science & Technology (IUST), Narmak, Tehran, Iran
E-mails: afshinfatahi@iust.ac.ir, noorossana@iust.ac.ir, dokouhaki@iust.ac.ir, m.babakhani@iust.ac.ir

ABSTRACT

Development of miscellaneous statistical methods for monitoring rare health events shows significance of the matter in health engineering domain. Rare health events, as attribute quality characteristics can not be monitored by traditional Shewhart based np charts since overdispersion occurs.
One relatively new approach to the 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 paper develops a truncated ZIB control chart (TZIB) applying probability limits in lieu of Shewhart based control limits for monitoring ZIB distributed observations. As the most prevalent criteria, average run length approach is used to evaluate the performance of the proposed chart. The applicability of TZIB control chart in phase I and phase II control charting is also investigated by a real case study, using the number of patients who go under surgery in a hospital and contract a wound infection. Results are compared with np chart as an inefficient control chart for our case study.

Key words: *Health Quality Engineering; Overdispersion; Rare Health Events; Truncated Distribution; Zero Inflated Binomial (ZIB) Distribution*

1. INTRODUCTION

Conventional 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 health surveillance (or monitoring) exclusively, as reviewed by Sonesson and Bock (2003) and Woodal (2006).
Some quality characteristics, which the researchers are recently interested in to monitor, are for example, infection rates, rates of patient falls, number of congenital malformations in a society, various sorts of waiting times and so forth (Benneyan 1998.a,b, Benneyan 2003, Lee & McGreevey 2002). Most of the statistical methods are developed based on the type of interested quality characteristic observations; whether they are 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 control chart (Benneyan 2001), g-type CUSUM control chart, sets method (Chen 1978), two modifications of the sets method as CUSCORE (Wolter 1987 , Munford 1980) and SHDA (Sitter et al. 1990), and the Bernoulli CUSUM (Sego et al. 2008). However, Sego believes that the performance of Bernoulli CUSUM and g-type CUSUM are better than sets method and its two modifications. The approach used for such methods assumes that all data are gathered and investigated one after another separately (Sego et al. 2008), but if you have to gather a grouped sample as aggregated data, these methods would fail to be implemented and some other methods should be substituted.
Applying traditional np-chart for 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 is the occurrence of a large number of zero failures. In industrial terminology, this condition is named "high yields processes" (Noorossana et al. 2007), while in health engineering it is called "rare health events" (Woodal 2006, Sego et al. 2008). It can be observed that when there are large numbers of zero data for an attribute quality characteristic, overdispersion (Woodal 2006) 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 Lambert (1997), 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.

In this paper, first a new type of health-related attribute characteristic, named ZIB, is taken into account. Then, a proper control chart is developed for monitoring such distributed observations values. Performance evaluation of the proposed control chart, based on average run length (ARL) is computed and illustrated in the next section. In the subsequent section, a motivating real health engineering case study, related to a hospital infection counts is presented to show the applicability of the new method. Finally, our conclusion about the matter is presented.

2. 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 Equation 1. In this equation, it is assumed that random shocks (like beginning of a special season, incidence of an epidemic disease, Changing surgery procedures and …) occur with probability θ, 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.

$$f_X(x;\theta,p)=\begin{cases}\theta\binom{n}{x}p^x(1-p)^{n-x} & ;x=1,2,\ldots,n\\ (1-\theta)+\theta(1-p)^n & ;x=0\end{cases} \quad (1)$$

Since for all X values, $f_X(x_i;\theta,p)\geq 0$ and $\sum_X f_X(x_i;\theta,p)=1$, this function is exactly a probability distribution function.

For the above distribution, mean and variance of the number of failures (X) can be calculated by Equations 2 and 3.

$$E(X)=\sum_{x=1}^{n}x\theta\binom{n}{x}p^x(1-p)^{n-x}+0\times\left[(1-\theta)+\theta(1-p)^n\right]=\theta\sum_{x=1}^{n}x\binom{n}{x}p^x(1-p)^{n-x}$$

$$=\theta\sum_{x=0}^{n}x\binom{n}{x}p^x(1-p)^{n-x}=np\theta \quad (2)$$

$$E(X^2)=E(X(X-1))+E(X)$$

$$=\sum_{x=1}^{n}x(x-1)\theta\binom{n}{x}p^x(1-p)^{n-x}+0(0-1)\times\left[(1-\theta)+\theta(1-p)^n\right]+np\theta$$

$$=\theta\sum_{x=1}^{n}x(x-1)\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}+np\theta$$

$$=\theta\sum_{x=2}^{n}\frac{n!}{(x-2)!(n-x)!}p^x(1-p)^{n-x}+np\theta$$

$$=\theta.n(n-1)p^2\sum_{x=2}^{n}\binom{n-2}{x-2}p^{x-2}(1-p)^{n-x}+np\theta$$

$$=\theta.n(n-1)p^2\sum_{y=0}^{s}\binom{s}{y}p^y(1-p)^{s-y}+np\theta$$

$$=\theta.n(n-1)p^2+np\theta=\theta\left[n(n-1)p^2+np\right]$$

$$Var(X)=E(X^2)-(E(X))^2=\theta.n(n-1)p^2+np\theta-n^2p^2\theta^2 \quad (3)$$

The parameters p and θ moment method estimates (MME) for the distribution can be obtained as Equation 4.

$$\begin{cases} E(X)=\dfrac{\sum_{t=1}^{m}X_t}{m} \\ E(X^2)=\dfrac{\sum_{t=1}^{m}X_t^2}{m} \end{cases} \Rightarrow \begin{cases} np\theta=\dfrac{\sum_{t=1}^{m}X_t}{m} \\ \theta.n(n-1)p^2+np\theta=\dfrac{\sum_{t=1}^{m}X_t^2}{m} \end{cases} \Rightarrow \begin{cases} \hat{p}=\dfrac{\sum_{t=1}^{m}X_t^2-\sum_{t=1}^{m}X_t}{(n-1)\sum_{t=1}^{m}X_t} \\ \hat{\theta}=\dfrac{(n-1)(\sum_{t=1}^{m}X_t)^2}{nm(\sum_{t=1}^{m}X_t^2-\sum_{t=1}^{m}X_t)} \end{cases} \quad (4)$$

Equation 5 shows the covariance between $\hat{p}$ and $\hat{\theta}$.

$$\begin{aligned} Cov(\hat{p},\hat{\theta}) &= E\left[\left(\frac{\sum_{t=1}^{m}X_t^2-\sum_{t=1}^{m}X_t}{(n-1)\sum_{t=1}^{m}X_t}\right)\left(\frac{(n-1)(\sum_{t=1}^{m}X_t)^2}{nm(\sum_{t=1}^{m}X_t^2-\sum_{i=1}^{m}X_t)}\right)\right] \\ &\quad - E\left(\frac{\sum_{t=1}^{m}X_t^2-\sum_{t=1}^{m}X_t}{(n-1)\sum_{t=1}^{m}X_t}\right).E\left(\frac{(n-1)(\sum_{t=1}^{m}X_t)^2}{nm(\sum_{t=1}^{m}X_t^2-\sum_{i=1}^{m}X_t)}\right) \\ &= \frac{1}{nm}E\left(\sum_{t=1}^{m}X_t\right)-\frac{1}{nm}E\left(\frac{\sum_{t=1}^{m}X_t^2-\sum_{i=1}^{m}X_t}{\sum_{t=1}^{m}X_i}\right).E\left(\frac{(\sum_{t=1}^{m}X_t)^2}{\sum_{t=1}^{m}X_t^2-\sum_{i=1}^{m}X_t}\right) \end{aligned} \quad (5)$$

Generating $m\times l$ simulated ZIB data (m: number of subgroups and l: number of simulation runs) we can obtain l estimate values for $\hat{p}_i$ and $\hat{\theta}_i$ ($i=1,2,\ldots,l$). Applying estimated $\hat{p}_i$ and $\hat{\theta}_i$ in Equation 5, l estimates are obtained for $Cov(\hat{p},\hat{\theta})$. Using an statistical test with a suitable significance level for the null hypothesis $H_0:\rho_{\hat{p},\hat{\theta}}=0$ against $H_0:\rho_{\hat{p},\hat{\theta}}\neq 0$ can not be rejected. So, we can have sensitivity analysis by shifting p and θ separately or simultaneously to calculate ARLs and evaluate the performance of the proposed control chart.

3. ZERO INFLATED BINOMIAL CONTROL CHARTS

3.1. Shewhart based ZIB control chart

Based on general Shewhart statistical control chart principles, if w is a statistic that measures a quality characteristic, and if mean and variance of w are equal to μ_w and σ_w^2 respectively, then the general model for the Shewhart control chart is as (Montgomery 2005):

$$\begin{cases} UCL=\hat{\mu}_w+L\hat{\sigma}_w \\ CL=\hat{\mu}_w \\ LCL=\hat{\mu}_w-L\hat{\sigma}_w \end{cases} \quad (6)$$

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 Equation 1, we obtain trial control limits as Equation 7.

$$\begin{cases} UCL_{ZIB} = n\hat{p}\hat{\theta} + L\left[n(n-1)\hat{p}^2\hat{\theta} + n\hat{p}\hat{\theta} - n^2\hat{p}^2\hat{\theta}^2\right]^{0.5} \\ CL_{ZIB} = n\hat{p}\hat{\theta} \\ LCL_{ZIB} = n\hat{p}\hat{\theta} - L\left[n(n-1)\hat{p}^2\hat{\theta} + n\hat{p}\hat{\theta} - n^2\hat{p}^2\hat{\theta}^2\right]^{0.5} \end{cases} \quad (7)$$

For different p and θ 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.

Applying probability limits instead of control limits neither would be helpful, since equation $\sum_{x=0}^{[LPL_{ZIB}]^+} f_X(x) = \frac{\alpha}{2}$ would not be satisfied for any rational α values; where $[\ ]^+$ is the integer round up function. To overcome this problem, we can calculate probability limit only for the first part of ZIB probability distribution function ($\theta\binom{n}{x}p^x(1-p)^{n-x}$) as a truncated distribution, Named truncated ZIB (TZIB) control chart.

In order to calculate TZIB probability limit ($UPL_{T\,ZIB}$), it is needed to solve Equation 8.

$$\sum_{x=[UPL_{T\,ZIB}]^+}^{n} \theta\binom{n}{x}p^x(1-p)^{n-x} = \frac{\alpha}{2} \quad (8)$$

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, and finally substituting them in Equation 8. In phase II, applying phase I upper probability limit, sampling n units at definite intervals and showing the number of failures, we can figure out whether the process is in control or out-of-control.

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.

3.2. 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

$$ARL = \frac{1}{Probability\ (\text{One point plots out of control})} \quad (9)$$

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 10.

$$ARL^+ = \frac{\text{Number of } X \text{ positive values, equal to or greater than } UCL_{T\,ZIB}}{\text{Number of } X \text{ positive}} \quad (10)$$

Applying different values for n in Equations 4 and 7, we can have a set of ARL curves. Also, using different values for type I error (α) leads to a set of ARL curves. Greater values for α leads to more false out-of-control points (false alarms) and decreases false in-control points. But since in health engineering applications, related costs of false alarms (costs of more inspections) is much less than the costs of false in-control points (costs of ignoring

mortality or acceleration of epidemic diseases), usually it's more rational to choose greater values for α than in industrial applications.

Applying typical values $p = 0.05$, $\theta = 0.1$ (for such situation, binomial failure rate is equal to 0.005, which the related events are considered as rare ones), $n = 50$ and $\alpha = 0.005$ for TZIB chart, UCL_{TZIB} is equal to 7. While applying the same parameters values for np-chart results in $LCL_{np} = 0$, $UCL_{np} = 0.93$ and in-control $ARL^{+} = 1$ which could not be useful at all. The ARL^{+} curves for the proposed TZIB chart are computed by simulation, base on Equation 10, and illustrated in Figure 1. Since ARL^{+} curves are computed by simulation, they are not smooth.

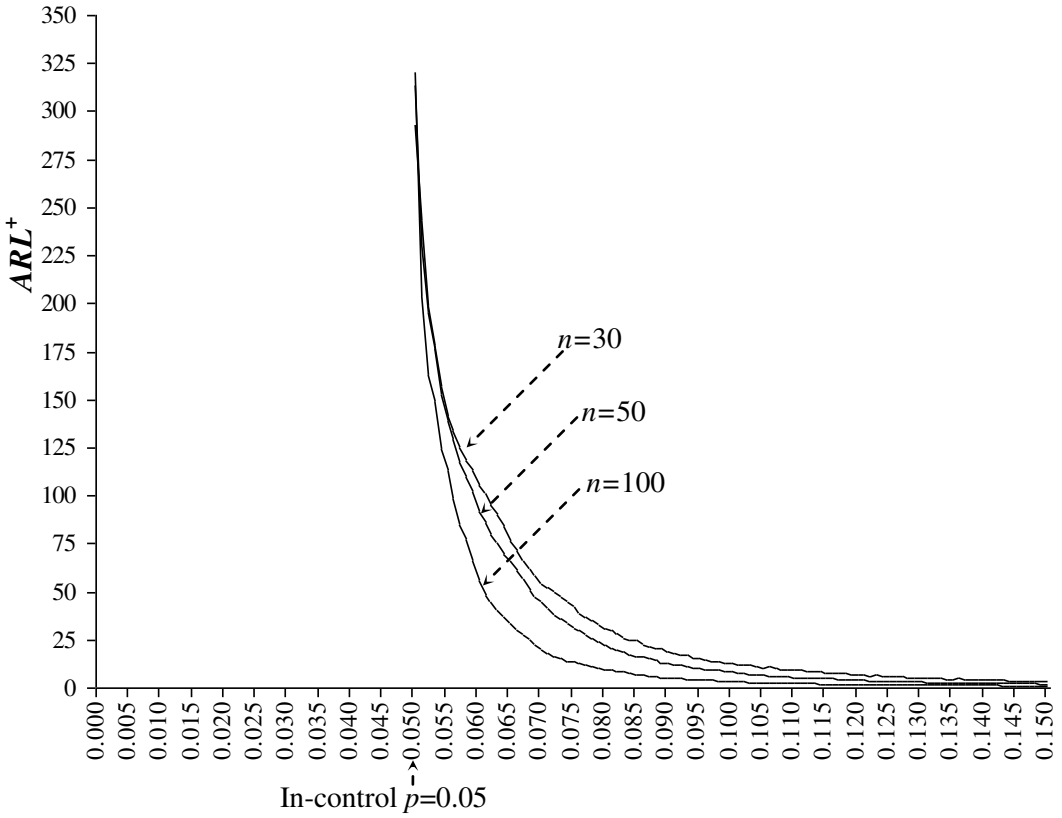

Figure 1. TZIB average run length with in-control parameters $\theta = 0.1$, $p = 0.05$

4. MOTIVATING CASE STUDY

Infection rates of patients in hospitals have always been one of the most important quality attribute characteristics. (Martone et al. 1991, Benneyan 1998.b) This nosocomial characteristic can be defined and measured in different parts of a hospital by epidemiologists. As a motivating case, in a hospital, we have concentrated on every day number of patients who go under a surgery and contract wound infection during four consecutive days. So, for X and the two parameters n, m we have

n : Number of every day sampling patients who go under surgery (sample size)

m : Number of consecutive sampling days to run phase I and estimating ZIB parameters p and θ.

X_t : Every day ($t = 1,2,\ldots,m$) number of patients who go under surgery and contract wound infection during next four consecutive days.

Data gathering for 100 consecutive days is presented in Table 1. Number of patients whom are scheduled to go under surgery on each day are different from 45 to 53. Based on Montgomery (2005), since the samples have small

variation in size, we can use average numbers of sample sizes for our case calculations. So, the sample size n is considered equal to 49 in the above case study.

Table 1. Every day number of patients who have gone under surgery and have contracted wound infection along 100 consecutive days

t:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X_t:	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0
t:	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
X_t:	1	0	0	0	0	1	0	0	0	0	0	0	2	1	0
t:	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
X_t:	1	0	0	0	0	0	0	0	0	0	0	1	0	0	2
t:	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
X_t:	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
t:	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
X_t:	0	0	1	0	0	0	0	0	1	1	0	2	0	1	0
t:	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
X_t:	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
t:	91	92	93	94	95	96	97	98	99	100					
X_t:	0	0	0	0	0	2	0	0	0	0					

Based on Equation 4 and the data lain in Table 1, we can obtain $\hat{p}$ and $\hat{\theta}$ values equal to 0.0114 and 0.4375 respectively. Applying Chi-square goodness of fit test for X_t ($t = 1,2,3,\ldots,100$) to test if it is distributed as a ZIB variable with parameters $\hat{p} = 0.0114$ and $\hat{\theta} = 0.4375$, results in a large enough P-value to assume it as a ZIB variable.

As it was mentioned before, since in health engineering applications, related costs of false alarms are much less than the costs of false in-control points, usually it's more rational to choose greater values for α than in industrial applications. So, in our case we have considered $\alpha = 0.005$, which is more than the value usually used in industrial applications ($\alpha = 0.0027$). According to Equation 8, UCL_{TZIB} is obtained equal to 3. Phase I of the proposed TZIB is illustrated in Figure 2. Since none of X_ts, presented in Table 1, are equal to or more than UCL_{TZIB}, we can assume that the process is statistically in control. So this proposed control chart can be considered for use in phase II.

As we can see in Figure 2, in comparison with TZIB, np-chart is just inefficient, showing many of the observations above UCL_{np} leading to exceeding false alarms.

From subgroup 101 to subgroup 250 (for 150 consecutive days), patients going under surgery on each day are evaluated for wound infections. These observations are presented in Figure 3 as phase II of TZIB control chart. As it can be seen, the values associated with days 145 and 236 are plotted out-of-control indicating more attention. Complementary investigations show an unexpected change in recovery procedure of patients after surgery for day 236, but for day 145 no especial cause could be found. Therefore, it was considered as a false alarm.

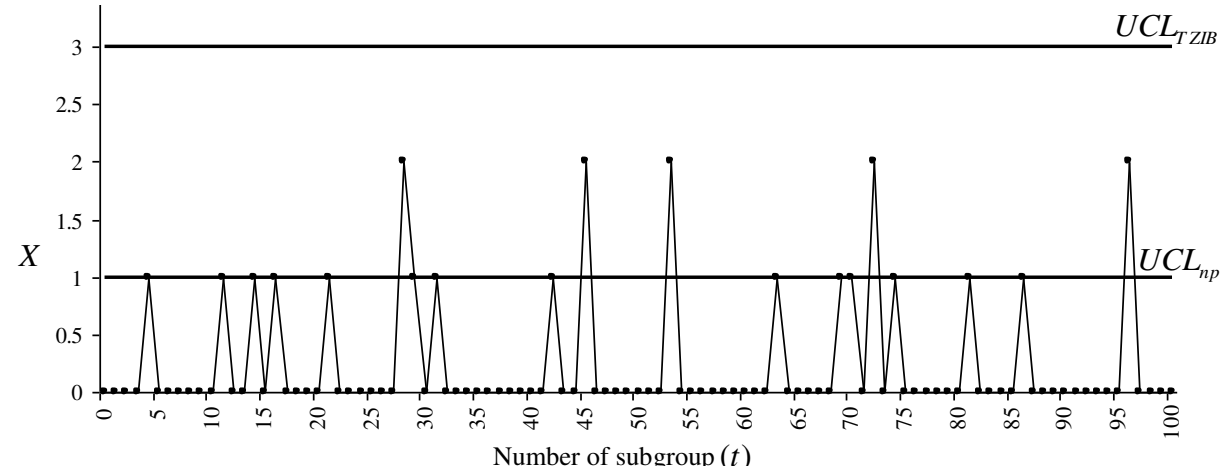

Figure 2. Phase I of TZIB control chart and the plotted X_t s ($t = 1,2,3,\ldots,100$) in comparison with np-chart

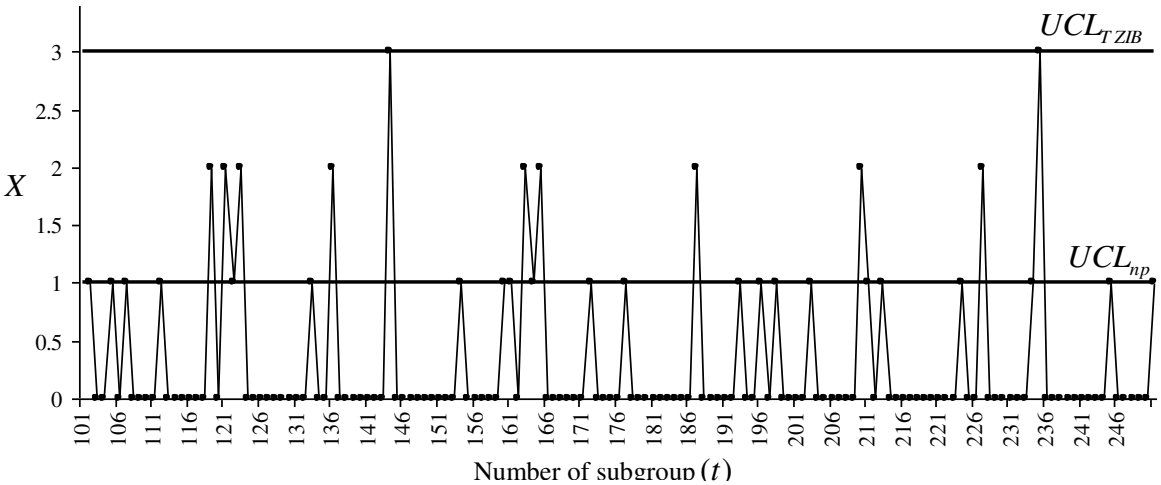

Figure 3. Phase II of TZIB control chart and the plotted X_t s ($t = 101,102,\ldots,250$) in comparison with np-chart

The ARL^+ curve related to the aforementioned case study is computed by simulation and depicted in Figure 4.

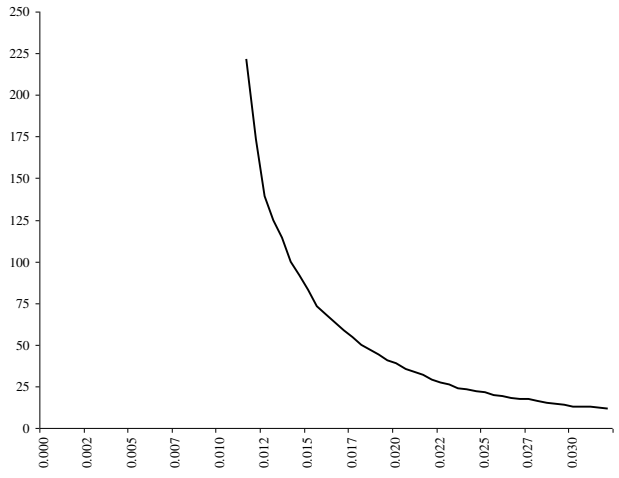

Figure 4. The ARL curve for TZIB chart with in-control parameters $p = 0.0114$ and $\theta = 0.4375$, and average sample size 49

5. CONCLUSION

Recently, miscellaneous statistical methods have been proposed in the literature to monitor rare health events. One relatively new approach to the problem is the use of control charts based on zero inflation in a binomial distribution. This paper develops a truncated ZIB control chart, applying probability limit in lieu of Shewhart based control limits for ZIB distributed observations. As the most prevalent criteria, average run length approach is used to evaluate the performance of the proposed chart. But 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 concentrate only on X positive values (not zero ones) leading to calculation of ARL^{+}. The applicability of the TZIB control chart in phase I and phase II is also investigated using the number of patients who go under surgery and contract a wound infection in a high technology hospital in Tehran. Results are compared with np-chart as an inefficient control chart for our case study.

REFERENCES

[1]. **Benneyan, J.C. 1998.** Statistical Quality Control Methods in Infection Control and Hospital Epidemiology, Part I: Introduction and Basic Theory. Infection Control and Hospital Epidemiology **19**: 194–214.

[2]. **Benneyan, J.C. 1998.** Statistical Quality Control Methods in Infection Control and Hospital Epidemiology, Part II: Chart Use, Statistical Properties, and Research Issues. Infection Control and Hospital Epidemiology **19**: 265–283.

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

[4]. **Benneyan, J.C. 2001.** Performance of Number-Between g-Type Statistical Quality Control Charts for Monitoring Adverse Events. Health Care Management Science **4**: 319-336.

[5]. **Benneyan, J.C. & Borgman, A.D. 2003.** Risk-adjusted Sequential Probability Ratio Tests and Longitudinal Surveillance Methods. International Journal for Quality in Health Care **15**: 5–6.

[6]. **Chen, R. 1978.** A surveillance system for congenital malformations. Journal of the American Statistical Association **73**: 323–327.

[7]. **Lambert, D. 1992.** Zero-infated Poisson regression with application to defects in manufacturing. Technometrics **34**: 1–14.

[8]. **Lee, K.Y. & McGreevey, C. 2002.** Using Control Charts to Assess Performance Measurement Data. Journal on Quality Improvement **28**: 90–101.

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

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

[11]. **Munford, A.G. 1980.** A control chart based on cumulative scores. Applied Statistics **29**: 252–258.

[12]. **Noorossana R, Saghaei A, Paynabar K, et al. 2007.** On the conditional decision procedure for high yield processes. Computers & Industrial Engineering **53**: 469- 477.

[13]. **Sego, L.H. Woodall, W.H. & Reynolds, M.R. 2008.** A comparison of surveillance methods for small incidence rates Statistics In Medicine **28**:1225- 1247.

[14]. **Sitter, R.R. Hanrahan, L.P. Demets, D. & Andereson, H.A. 1990.** A monitoring system to detect increased rates of cancer incidence. American Journal of Epidemiology **132**: 123–130.

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

[16]. **Wolter, C. 1987.** Monitoring intervals between rare events: a cumulative score procedure compared with Rina Chen's sets technique. Methods of Information in Medicine **26**: 215–219.

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