MultiCraft

International Journal of Engineering, Science and Technology
Vol. 14, No. 2, 2022, pp. 34-40

INTERNATIONAL JOURNAL OF ENGINEERING, SCIENCE AND TECHNOLOGY

www.ijest1-ng.com
www.ajol.info/index.php/ijest

Process control for categorical (ordinal) data

Nandini Das

SQC&OR unit, Indian Statistical Institute, 203 BT Road, Kolkata-700108, INDIA
**Corresponding Author: e-mail: nandini@isical.ac.in, Tel+91-033-25753352*
ORCID iD: http:/orcid.org/0000-0002-6857-6490

Abstract

Quality improvement is playing the key role in the success of a business. Reduction of variability is the main step for improvement of quality. Control charts are developed for the purpose of monitoring the quality characteristics with the aim of reducing variability. In many industries instead of continuous variable categorical (ordinal) data are used to measure the quality characteristics of interest. Hence developing control charts techniques for monitoring ordinal data has become a recent research focus. Quality control practitioners often face a problem to select the appropriate technique for monitoring ordinal data in the practical field since there are quite a few techniques available in the literature for this purpose. In this paper we have studied the various techniques for monitoring ordinal data and compared their performance to detect the shift in location parameter. Data were simulated from Normal distribution and average run length (ARL) were computed for different values of shift in mean (both in positive and negative direction) using different methodologies under study. The best technique to detect the shift was identified with respect to ARL.

Keywords: Ordinal variable, Control chart, Average run length, Simulation

DOI: http://dx.doi.org/10.4314/ijest.v14i2.4

Cite this article as:
Das N. 2022. Process control for categorical (ordinal) data. *International Journal of Engineering, Science and Technology*, Vol. 14, No. 2, pp. 34-40. doi: 10.4314/ijest.v14i2.4

Received: October 10, 2022; Accepted: October 19, 2022; Final acceptance in revised form: October 27, 2022

1. Introduction

Control chart is a useful quality control tool to monitor different quality characteristics. It is giving alarm when the assignable cause of variation occurs in the process. Control charts are of two types: Control chart for variable data and Control chart for attribute data. Though variable data always contains more information about the quality characteristics it is not always possible to collect variable data due to time and cost constraint. Hence attribute data is also playing an important role in all practical purpose. When attribute data are classified into several categories according to some thresholds they are called categorical data. Ordinal data is a categorical variable with some intrinsic order of the specified categories. Some typical examples are: Education (no high school degree, HS degree, under graduate, post graduate), Agreement (strongly disagree, moderately disagree, neutral, moderately agree, strongly disagree), Rating (excellent, good, fair, poor), Frequency of occurrence (always, often, sometimes, never). It is clear from above examples that ordinal data is having a clear order in the categories.

Ordinal data is commonly used in the field of social science, marketing, medical and public health domain. It is worthwhile to note here that any type of measurement, even if it is a continuous variable, can be converted to categorical data constructing different categories within the range of measurement. Hence in the field of quality control ordinal variable is always drawing researcher's attention.

The research on the topic of control chart for controlling ordinal variable has its starting point since 1950 originated by Duncan. With the passage of time a number of different control chart techniques have been developed by different researchers adopting different theoretical considerations. Quality control practitioners are often facing the problem to choose the right one from the different methods available in the literature. Hence this paper aims to compare the performance of different ordinal control charts to detect the shift and summarized the result.

The paper is written in the following sections. Section-2 describes the different methods available in the literature for controlling ordinal variable in a nut shell. Since it is not possible to cover each and every methodology for comparison purpose, some important techniques were selected for the study. Section-3 describes the methodologies under study elaborately. Performance comparison was done simulating data from Normal distribution and computing average run length (ARL) for different methodologies for different magnitudes of shift. This analysis is given in section-4. Conclusions drawn from the simulation study are provided in section-5. Section-6 is throwing some light on future direction of research in this topic.

2. Literature survey

The p-chart is the easiest technique for controlling attribute data. Hence using separate p-chart for each category may be one option for handling data with several categories. Dealing with several p-chart may be difficult to implement in the practical situation. Hence, Duncan (1950) proposed the use of chi-square statistics in-stead of using several p-chart. Marcucci (1985) and Nelson (1987) also proposed a similar approach. Fuchs and Kenett (1980) proposed a more powerful technique than chi square technique using M-test. Woodall, Tsui, and Tucker (2002) used maximum likelihood estimate assuming different underlying distribution and provided a new technique for monitoring ordinal variable. Duran and Albin (2009) developed a new method for controlling k categories using a probability tree methods with k-1 binary stages. A multinomial cumulative sum control chart techniques using likelihood ratio statistics was suggested by Ryan, Wells, and Woodall (2011). An excellent review article on multinomial control chart was written by Topalidou and Psarakis (2009). Weiß (2012) assumed that under in control state the quality characteristics are identically and independently distributed and suggested three different techniques for controlling a purely categorical variable to detect the violation of assumptions. Weiß (2012) proposes a chart based on Gini index and another control chart based on (*k*, *r*) runs.

Yashchin (2012) developed a change-point detection scheme using generalized likelihood ratio tests. An ordered samples control charts was proposed by Franceschini et al. (2005) based on a dominance criterion considering a position in the ordered sample space to each sample. Jian et al. (2014) suggested a technique for detecting shifts in the location parameter of the latent variable using the attribute level counts without losing the ordinal information. Bai and Li (2021) developed a control chart for controlling ordinal variable using log linear model. Fullerton & Xu (2018) developed a new and improved version of the constrained partial adjacent category model considering unconstrained and constrained versions of the partial adjacent category logit model, relaxing the proportional odds assumption for a subset of independent variables. Fernandez et al. (2019) modelled the outcomes using the ordered stereotype model, which is not so popular as other models such as linear regression and proportional odds models. They compared the performance of the ordered stereotype model with other more commonly used models among researchers and practitioners.

Pal & Gauri (2021) introduced a technique by using appropriately designed control chart for the sample estimate of the process area of proportions obtained from a sample of size n collected at regular intervals. The uniqueness of their technique is that they do not require any assumption about the latent variable. Hakimi et al. (2019) developed a new control charts using ordinal-normal statistic to control the ordinal log-linear model based processes in Phase II. Two new control charts based on the WALD and Stuart score test statistics was designed and developed by Kamranrad et al. (2017a) for monitoring of contingency table-based processes in Phase-II. Kamranrad et al. (2017b) suggested two control charts based on the generalized linear test (GLT) and contingency table for Phase-II monitoring of multivariate categorical processes.

3. Different Control Chart Procedures under Study

In this study we will consider the following methodologies and compare their performance with respect to the average run length (ARL) for different shifts in location parameters.

3.1 Duncan's method (extension of the p-chart)

3.1.1 Methodology

Duncan (1950) proposed a chi-square distribution based statistics to monitor categorical data. He assumed that the samples are independent. The test statistics used by him is as follows:

$$CHISQ_j = \sum_{i=1}^{k} \frac{(n_{ij}-s_j p_i)^2}{s_j p_i} \tag{1}$$

where, n_{ij} = the number of observation in ith category corresponding to jth sample

p_i = Proportion in ith category

$s_j = \sum_{i=1}^{k} n_i$

The chart statistic is following chi-square with $k-1$ degrees of freedom asymptotically. Hence upper control limit can be obtained from the χ^2 table. If the observed value of the statistic goes beyond the limit it gives an out of control signal.

3.2 Control chart based on MLE (proposed by Woodall, Tsui, and Tucker (2002))

3.2.1 Assumptions

- Assuming an underlying distribution of quality measurements.
- Assuming that as the change in location parameter will not affect the proportion of the category in the poorest level of quality.
- It is more likely that change in the distribution of quality measurements introduces quality improvement.

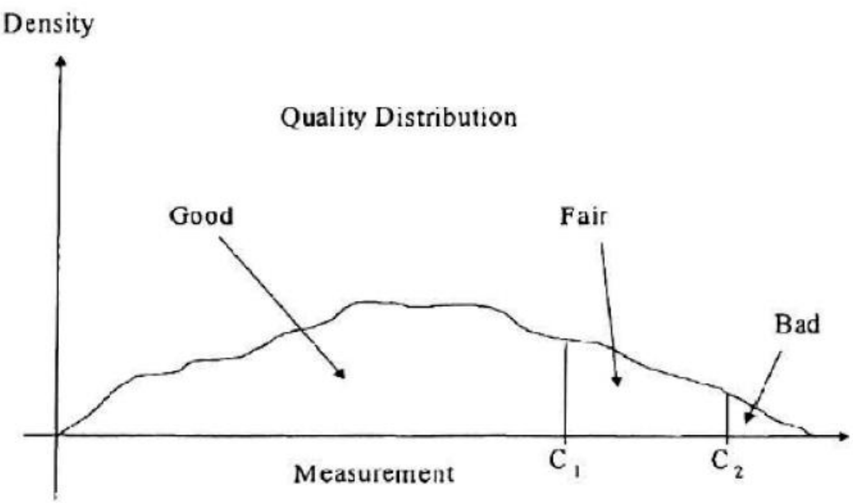

Figure 1. Distribution of quality characteristic measurements

Figure 1 shows an underlying distribution where higher the value poorer the quality level. Woodall, Tsui, and Tucker (2002) proposed a technique to monitoring ordinal variable using maximum likelihood estimate (MLE) of the unknown parameters of the basic distribution of the quality characteristic.

3.2.2 Methodology

The basic assumption made by McCullagh and Neider (1989) is that location parameter of the underlying distribution is 0 and scale parameter is 1 when the process is under statistical control. Let p_i be the proportion in the ith category when the process is under in-control state. One can easily get the cut points $C_0, C_1, \dots C_i, \dots C_k$ from the inverse cumulative distribution function $F^{-1}(p_1+\dots+p_i) = C_i$. The range of the underlying probability density function is partitioned into k ordered categories by the cut points. The interval (C_{i-1}, C_i) denotes the i^{th} category. The number of observations in ith category can be observed from the future sample and denoted by n_i .

The probability of an item belongs to ith category is

$$p_{i,\theta} = \int_{C_{i-1}}^{C_i} f(\theta, u)du \tag{2}$$

where θ is the location parameter.

The likelyhood function is

$$L(\theta) = \prod_{i=1}^{k} (p_{i,\theta})^{n_i} \tag{3}$$

Let , $\hat{\theta}$ be the maximum likelihood estimate (MLE) of θ. As the number of categories and sample size increases the distribution of $\hat{\theta}$ tends to Normal distribution.

The chart statistics based on MLE is

$$T = \hat{\theta}/STD(\hat{\theta}) \tag{4}$$

where STD$(\hat{\theta})$ is the asymptotic standard deviation of $\hat{\theta}$.

Control limits can be determined by assuming different underlying distribution.

3.3 Simple Ordinal Chart (SOC)(proposed by Jian et al. (2014))

3.3.1 Assumptions

- The methodology can be applied to the grouped ordinal data
- The underlying distribution and the cut-points are unknown.
- The numerical value of the underlying continuous variable shows the order of the categories.
- The change in MLE of location parameters reflects the shift in the underlying continuous distribution.

3.3.2 Procedure

The basic assumption of the procedure proposed by Jian et al. (2014) is that there exists an underlying continuous variable with cumulative distribution function (CDF)$F(X)$ which decides the levels of the categories say Y. A particular set of numerical values of X denotes a particular attribute level of Y. The count of each level of Y is equal to the number of the number of continuous values of X lying in that category.

Let there be h labels in the ordinal data. This implies there must be h class intervals in the underlying distribution.

Such intervals are constructed by some cut-points $a_k(k = 0, \dots \dots, h)$.

Let y_{kt} $(k = 1, \dots \dots, h)$ be an indicator variable denoting whether the value x_t $(t = 1,2, \dots \dots)$ is falling in the k^{th}interval $(a_{k-1}, a_k]$,

$$y_{kt} = \boldsymbol{I}(x_t \in (a_{k-1}, a_k])$$

For a sample of size N the number of counts of level k of Y is

$$n_k = \sum_{t=1}^{N} y_{kt} \quad (5)$$

Assuming the CDF of X as $F(X)$ the probability that X will lie in the kth interval is $p_k = F(a_k) - F(a_{k-1})$.

For sample of size N, the joint distribution of n_k $(k = 1, \dots \dots, h)$ is the multinomial distribution $MN(N; p)$ with $p = [p_1, \dots \dots, p_h]^T$

To test whether there is a shift in the location parameter of ordinal data Y is equivalent to test the shift in location parameter of X. This implies to test whether δ is zero or not assuming CDF F(x−δ) under the out of control state.

Hence the test statistic boils down to

$$H_0: \delta = 0 \ versus \ H_1: \delta > 0$$

Let $c_k = \sum_{j=1}^{k} p_j$ $(k = 1, \dots \dots, h)$ be the cumulative probability up to k and $c_0 = 0$ and pdf of X is f.

The asymptotically most powerful test statistic is given below

$$l = \sum_{k=1}^{h} \beta_k n_k, \quad (6)$$

$$\beta_k = f[(F^{-1}(c_{k-1}) - (F^{-1}(c_k)]/p_k \quad (7)$$

The ratio of the cumulative count up to level k to the sample size provides the value of c_k.

The above test procedure is meant for one sided test. When the process is under in control state i.e. with δ = 0, we h have $E[l] = 0$. Hence $|\, l \,|$ can be plotted if the shift direction is unknown. $|\, l \,|$ is having the memory less property. A more efficient statistic is based on EWMA statistics which is given in equation (8).

$$z_i = a_{0,i,\lambda}^{-1} \sum_{j=1}^{i} (1-\lambda)^{i-j} n_j \quad (8)$$

where $a_{,t_0 t_1, \lambda} = \sum_{j=t_0+1}^{t_1} (1-\lambda)^{t_1-j}$ is a sequence of constants such that sum of the weights is unity, 0 <λ< 1.

3.4 Using Gini index (proposed by Weiß (2012))

3.4.1 Assumptions

- The categories are mutually exclusive.
- 100% inspection is carried out.
- The quality characteristics are iid variable $with \ p = p_0$ under in control state.

3.4.2 Computation

Weiß (2012) assumes $(X_t)_N$ as the basic categorical variable to be monitored. He used the moving average (MA) estimator which combined the characteristic properties of the underlying distributions into a single real number and compares it with that under the in control state. The test statistics is given in equation (9)

$$\widehat{p_t}^{(w)} = \frac{\sum_{r=0}^{w-1} X_{t-r}}{w}, for\ t \geq w \tag{9}$$

The chart statistic $T_t \coloneqq T_{p_0}(\widehat{p_t}^{(w)})$ is developed using the marginal distribution of $(X_t)_N$ and control limits are computed appropriately, where $T_{p_0} = [0;1]^{m+1} \to \mathbb{R}$ is a comparative function corresponding to the marginal distribution under in control state. One efficient measure of dispersion of categorical data is Gini index. Weiß (2012) proposed a statistic based on Gini index. It is quite likely that shift in p will introduce shift in dispersion. The relevant chart statistic to be plotted is given in equation 10.

$$T_{p_0}\left(\widehat{p_t}^{(w)}\right) = \frac{1-s_2(\widehat{p_t}^{(w)})}{1-s_2(p_0)} - 1 \tag{10}$$

where, $s_k(p) = \sum_{j=1}^{m+1} {p_j}^k$ for $k \in \mathbb{N}$ and for a probability vector $p = (p_1, \dots \dots, p_{m+1})^T$ and $s_1(p) = 1$

The control limits can be computed using simulation.

3.5 Ordered sample control charts (proposed by Fiorenzo et al. (2005))

3.5.1 Assumptions

- It does not require any distributional assumption.
- Dominance criteria is used to form the Sample space ordering

3.5.2 Methodology

The concept of ordered sample values and ordered sample ranges is used in this methodology proposed by Fiorenzo et al. (2005). According to the definition given by Fiorenzo et al., a sample A is said to be preferred to sample B if A dominates B. Using this concept a new ordinal scale is defined whose levels are the positions of samples in the ordered sample space. If there does not exist any dominance relationship between the sample then they can form a same equivalence class. The selection of dominance criteria affects the number of levels and order of the levels. Dominance criteria may be developed based on the specific process criteria.

3.5.3 Ordered sample control charts

At the outset analysing the Pareto-dominance criterion is explained. By definition , sample A is said to pareto dominates sample B if all the points in B do not exceed the corresponding points in A, and at least one point in A exceeds the corresponding one in B. The formal notation under this situation is $A \rhd$B . If samples A and B belongs to the same equivalence class, the notation will be: $X \approx$Y. Since their intersection is not null set It is not possible to assign a well-defined position to samples A, B and C.

The Rank dominance criterion is another criterion which uses the concept of optimal sample. If all the elements are in the highest level of ordinal scale it is called optimal. Rank index for any sample is developed based on its positioning with regard to the optimal sample. The index is computed by adding the numbers of scale levels contained between each sample value and the corresponding value of the optimal sample. A 'bad' sample is having high value of rank index. The samples belonging to same equivalence class are having the same index. The rank dominance criterion provides $n(t-1)+1$ numbers of equivalence classes, where t is the number of levels and sample size is n.

Integrating the rank dominance criterion with the dispersion dominance criterion results in a greater resolution. For this chart only lower control limit exists. Central line can be taken as the median of sample distribution. An empirical distribution can be obtained using the initial samples. Control limit can be computed using this empirical distribution for a fixed type I error.

4. Performance Assessment

Performance comparison of the above mentioned methods, described in the section 3.0, is computed following the steps given below.

- Consider a process having 5 categories, with known in control probabilities p_1=0.85, p_2=0.08, p_3=0.04, p_4=0.02, p_5=0.01.
- Assuming Normal distributions the control limits of all the control charts are determined based on the random samples (of size 200) for a fixed value location parameter (0), scale parameter (1) and for a fixed value of type-1 error.
- Shifts considered in the location parameter are δ= ±0.02, ±0.05, ±0.1, ±0.2, ±0.5, ±1, ±1.5, ±2, ±2.5, ±3.

- Random Samples were simulated from with location parameter at some fixed shift(s) mentioned above and scale parameter = 1. Chart statistics are plotted for each of the samples and 5000 run lengths were simulated, which was followed by computation of ARL.
- Performance comparison with respect to average run length (ARL) of Control Chart based on MLE, Duncan's method, SOC Method, Control chart using Gini's coefficient and ordered sample control chart is shown in the Table 1.

Table 1. Performance of different control charts with respect to ARL for various shift in location parameters

Shift in mean	Control Chart based on MLE	Duncan's CHISQ	SOC	Using Gini's coefficient	Ordered sample control chart
3	1	1	1	1	1
2.5	1	1	1	1	1
2	1	1.8	1	1.4	1
1.5	1	2.1	1	3.5	3.0
1	1	5.9	1	4.1	2.5
0.5	1.1	43.7	1	33.8	24.8
0.2	8.5	115.1	8.3	50.2	48.3
0.1	20.5	283.2	21.2	183.7	171.2
0.05	64.8	369.8	65.5	388.8	265.5
0.02	185.3	426.6	191.7	421.6	391.7
-0.02	309.9	412.2	308.7	410.6	311.7
-0.05	246.0	124.2	247.0	201.2	291.0
-0.1	72.6	68.9	73.2	65.4	88.2
-0.2	8.9	18.5	8.8	17.6	18.8
-0.5	1.01	2.9	1	2.5	2.1
-1	1	2.8	1	2.9	1.8
-1.5	1	1.5	1	1.3	1.1
-2	1	1	1	1	1
-2.5	1	1	1	1	1
-3	1	1	1	1	1

5. Conclusion

It is worthwhile to note here that lower the Average run length higher the performance of the control chart to detect the shift. Based on the simulation results summarised in Table 1 following conclusions can be drawn.

- For detecting higher shift (> 1.5) in location parameter, both in positive and negative direction, all the techniques under study perform almost at par.
- For detecting moderate shift (≥ 1 but ≤ 1.5) Control Chart based on MLE and SOC method are better than the other 3 methods i.e. Ordered sample control chart, Duncan's CHISQ and Gini's coefficient method. But for negative moderate shift Ordered sample control chart is slightly better than Duncan's CHISQ and Gini's coefficient method.
- For detecting lower shift (< 1) Control Chart based on MLE and SOC method are better than the other 3 methods i.e. Ordered sample control chart, Duncan's CHISQ and Gini's coefficient method. but Ordered sample control chart is better than Duncan's CHISQ and Gini's coefficient method for detecting lower shift in positive direction. But for negative direction Gini's coefficient method is observed to perform slightly better than Duncan's CHISQ method and Ordered sample control chart.

For future scope of work, other methods described in the literature survey can be studied in more detail and their performance can be assessed in a similar way. The performance of the different charts can be investigated by simulating data from non-normal distribution and their robustness can be explored.

Acknowledgment
The author is grateful to the unknown reviewer for his valuable suggestion which enables the paper in this final form.

References

Bai K., Li J., 2021. Location-scale monitoring of ordinal categorical processes, *Naval Research Logistics*, Vol. 68, No. 7, pp. 937-950. https://doi.org/10.1002/nav.21973

Cozzucoli, P., 2009. Process monitoring with multivariate p-control chart, *International Journal of Quality, Statistics, and Reliability*, Volume 2009, Article ID 707583, pp. 1-11. https://doi.org/10.1155/2009/707583

Duncan, A. J., 1950. A chi-square chart for controlling a set of percentages, *Industrial Quality Control*, Vol. 7, No. 11, pp. 11-15

Duran, I., and S. L. Albin., 2009. Monitoring and accurately interpreting service processes with transactions that are classified in multiple categories. *IIE Transactions* Vol. 42, No. 2, pp.136–45. https://doi.org/10.1080/07408170903074908

Fernandez D, Liu I, Costilla R., 2019. A method for ordinal outcomes: The ordered stereotype model, *International Journal of Methods in Psychiatric Research*, Vol. 28, No. 4, e1801, https://doi.org/10.1002/mpr.1801

Franceschini, F., M. Galetto, and M. Varetto. 2005. Ordered samples control charts for ordinal variables, *Quality and Reliability Engineering International,* Vol. 21, No. 1, pp. 177–195. https://doi.org/10.1002/qre.614

Fuchs, C., and R. Kenett. 1980. A test for detecting outlying cells in the multinomial distribution and two-way contingency tables. *Journal of the American Statistical Association*, Vol. 75, No. 370. pp. 395–398. https://doi.org/10.1080/01621459.1980.10477483

Fullerton A.S., & Xu J. 2018. Constrained and unconstrained partial adjacent category logit models for ordinal response variables. *Sociological Methods and Research*, Vol. 47, No. 2, pp. 169–206. https://doi.org/10.1177/0049124115613781

Hakimi A, Farughi H, Amiri A, Arkat J, 2019. New phase II control chart for monitoring ordinal contingency table based processes, *Journal of Industrial and Systems Engineering*, Vol. 12, Special issue on Statistical Processes and Statistical Modeling, pp. 15-34.

Jian, L., Tsung, F. and Zou, C. 2014. A simple categorical chart for detecting location shifts with ordinal information, *International Journal of Production Research*, Vol. 52 No. 2, pp. 550-562. https://doi.org/10.1080/00207543.2013.838329

Kamranrad, R., Amiri, A., & Niaki, S.T.A. , 2017b. Phase II monitoring and diagnosing of multivariate categorical processes using generalized linear test-based control charts, *Communications in Statistics-Simulation and Computation*, Vol. 46. No. 8, pp. 5951-5980. https://doi.org/10.1080/03610918.2016.1186186

Kamranrad, R., Amiri, A., & Niaki, S.T.A., 2017a. New approaches in monitoring multivariate categorical processes based on contingency tables in phase II. *Quality and Reliability Engineering International*, Vol. 33, No. 5, pp. 1105-1129. https://doi.org/10.1002/qre.2103

Marcucci, M. 1985. Monitoring multinomial processes. *Journal of Quality Technology*, Vol. 17, No. 2, pp. 86-91. https://doi.org/10.1080/00224065.1985.11978941

McCullagh, P. and Neider, J. A. 1989, Generalized Linear Models, Chapman and Hall, New York, NY.

Nelson, L. S. 1987. A chi-square control chart for several proportions. *Journal of Quality Technology*, Vol. 19, No. 4, pp. 229-231.

Pal S & Gauri S K, 2021 Monitoring processes with ordinal data: an area-based approach, *Communications in Statistics - Simulation and Computation*, published online, https://doi.org/10.1080/03610918.2021.1882494

Ryan, A. G., L. J. Wells, and W. H. Woodall. 2011. Methods for monitoring multiple proportions when inspecting continuously. *Journal of Quality Technology* Vol. 43, No. 3, pp. 237-248. https://doi.org/10.1080/00224065.2011.11917860

Topalidou, E., and S. Psarakis. 2009. Review of multinomial and multi attribute quality control charts. *Quality and Reliability Engineering International,* Vol. 25, No. 7, pp. 773–804. https://doi.org/10.1002/qre.999

Tucker, G. R., Woodall, W. H. & Tsui, K.L. 2002. A control chart method for ordinal data, *American Journal of Mathematical and Management* Sciences, Vol. 22. No. 1, pp. 31-48. https://doi.org/10.1080/01966324.2002.10737574

Weiß, C. H. 2012. Continuously monitoring categorical processes. *Quality Technology & Quantitative Management,* Vol. 9, No. 2, pp. 171–88. https://doi.org/10.1080/16843703.2012.11673284

Yashchin, E. 2012. On detection of changes in categorical data. *Quality Technology & Quantitative Management*, Vol. 9, No. 1, pp. 79–96. https://doi.org/10.1080/16843703.2012.11673279

Biographical notes

Nandini Das received her M. Stat. and M.Tech. (QROR) degree from Indian Statistical Institute. She received her Ph.D. degree from Jadavpur University. Presently she is working as a faculty in SQC&OR division of Indian Statistical Institute. She has 30 years of experience in teaching, research and consultancy. She has published more than 30 papers in international journals. Her area of interest is Statistical process control.