Journal of Applied Statistics
Vol. 35, No. 4, April 2008, 421–429

Multivariate attribute control chart using Mahalanobis D^2 statistic

Arup Ranjan Mukhopadhyay*

SQC & OR Unit, Indian Statistical Institute, Kolkata, India

Process control involves repeated hypothesis testing based on several samples. However, process control is not exactly hypothesis testing as such since it deals with detection of non-random patterns of variation as well in a fleeting kind of population. Compare this with hypothesis testing which is principally meant for a stagnant population. Dr Walter A. Shewhart introduced a graphic method for doing this testing in a fleeting population in 1924. This graphic method came to be known as control chart and is widely used throughout the world today for process management purposes. Subsequently there was much advancement in process control techniques. In particular, when more than one variable was involved, process control techniques were developed mainly by Hicks (1955), Jackson (1956 and 1959) and Montgomery and Wadsworth (1972) based on the pioneering work of Hotelling in 1931. Most of them have worked in the area of multivariate variable control chart with the underlying distribution as multivariate normal. When more than one attribute variables are involved some works relating to test of hypothesis was done by Mahalanobis (1946). These works were also based on the Hotelling T^2 test. This paper expands the concept of 'Mahalanobis Distance' in case of a multinomial distribution and thereby proposes a multivariate attribute control chart.

Keywords: Euclidean distance; Mahalanobis distance; multinomial distribution; correlation matrix; variance covariance matrix

Introduction

There are many situations when inspection classifies the products into several categories of non-conformities. A control scheme is required to exercise simultaneous control of all the categories. With the existing tools one can apply several proportion defective (p) charts – one for each category of defect. However, this will be equivalent to testing several equality of proportion defective hypotheses independently. The problem here is of the type H_0: $p_i = p$ where $p_i \sim$ multinomial (n_i, p). So it is easy to check that in the several p charts case the type I error and consequently the power of the test will suffer a distortion. To take care of the above situation a simultaneous test of $p_i = p$ was thought of.

There are two ways of looking at the data matrix. One may be interested in comparing the columns of the data matrix, i.e. the variables. This leads to techniques known as R-techniques,

*Email: armukherjee@yahoo.co.in

so called because the correlation matrix R plays an important role in this approach. Principal component analysis, factor analysis, canonical correlation analysis fall under this group of techniques. But in the present case the interest is of comparing rows of the data matrix, i.e. the different objects that are time points here. This leads to techniques such as discriminant analysis, cluster analysis, multidimensional scaling which are known as Q techniques. The Mahalanobis' concept of distance between objects plays an important role in all these Q techniques [1,4].

The Mahalanobis distance

For a data matrix whose columns represent variables and the rows represent the objects, a natural way to compare two rows $\mathbf{X}_r$ and $\mathbf{X}_s$ is to look at the Euclidean distance between them.

$$\|\mathbf{X}_r - \mathbf{X}_s\|^2 = (\mathbf{X}_r - \mathbf{X}_s)^{\mathrm{T}}(\mathbf{X}_r - \mathbf{X}_s).$$

But when the variation in $\mathbf{X}$ is stochastic in nature it is better to look at a transformation of the form

$$\mathbf{Z}_r = \mathbf{S}^{-1/2}(\mathbf{X}_r - \bar{\mathbf{X}}), \quad r = 1, 2, \ldots, n.$$

This enables one to eliminate the correlation between the variables and standardize the variances of each variable.

$$\mathbf{S} = \frac{1}{n}\sum_{r=1}^{n}(\mathbf{X}_r - \bar{\mathbf{X}})(\mathbf{X}_r - \bar{\mathbf{X}})^{\mathrm{T}}.$$

After the aforesaid transformation one can look at the Euclidean distance between the transformed rows. Such distances play a role in cluster analysis. The most important of these distances is the Mahalanobis distance given by

$$D_{rs}^2 = \|\mathbf{Z}_r - \mathbf{Z}_s\|^2 = (\mathbf{X}_r - \mathbf{X}_s)^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{X}_r - \mathbf{X}_s).$$

Mahalanobis distance can be of different kinds.

(i) Let $\mathbf{X} \sim (\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and let $\mathbf{Y} \sim (\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ then $D_{\mu 1 \mu 2}$ is a Mahalanobis distance between the parameters.
(ii) Let $\mathbf{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The Mahalanobis distance between X and μ, $D_{X\mu}$, is here a random variable.
(iii) Let $\mathbf{X} \sim (\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $\mathbf{Y} \sim (\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. The Mahalanobis distance between X and Y is D_{XY} [2,3].

The present problem (qualitative data)

Consider a classification of individuals into K categories. For each $i = 1, 2, 3, \ldots, n$ let $(p_{i1}, p_{i2}, p_{i3}, \ldots, p_{ik}) = \mathbf{p}_i^{\mathrm{T}}$ denote the observed proportions from a population of size N_i subject to $\mathbf{p}_i^{\mathrm{T}} \sim$ multinomial with parameters $\bar{\mathbf{p}}^{\mathrm{T}}$. For example, $\mathbf{p}_i^{\mathrm{T}}$ might denote the proportions of defectives which are segregated defect-wise in a painting shop of ceiling fans as 'poor covering', 'overflow', 'patty defect', etc. and the corresponding proportion of good items in each of n days. Hence, $\sum_{j=1}^{K} p_{ij} = 1, i = 1, 2, 3, \ldots, n$. Consequently, the variance-covariance matrix of the vector $\mathbf{p}_i^{\mathrm{T}}$ is singular. So a Mahalanobis distance will not exist. However, the requirement under the present circumstance is Mahalanobis-like distance of the second kind, which has been

discussed earlier. Hence a possible distance formula could be

$$D_i^2 = (\mathbf{p}_i - \bar{\mathbf{p}})^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-} (\mathbf{p}_i - \bar{\mathbf{p}}).$$

Here $\boldsymbol{\Sigma}_i$ is the variance-covariance matrix of the vector $\mathbf{p}_i$ and is equal to $N_i^{-1}\boldsymbol{\Sigma}$, where

$$\boldsymbol{\Sigma} = [\sigma_{ij}],$$

$$\sigma_{ij} = \begin{cases} \bar{\mathbf{p}}_i(1 - \bar{\mathbf{p}}_i) & \text{for } i = j, \\ -\bar{\mathbf{p}}_i \bar{\mathbf{p}}_j & \text{for } i \neq j. \end{cases}$$

Since, $\mathbf{p}_i$ lies on a hyper plane $\boldsymbol{\Sigma}$ is singular [6]. However, it is easy to check that a g-inverse of $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma}^{-} = \text{diag.}[\bar{p}_1^{-1}, \bar{p}_2^{-1}, \bar{p}_3^{-1}, \ldots, \bar{p}_K^{-1}].$$

Thus, a generalized Mahalanobis distance can be defined by $D_i^2 = (\mathbf{p}_i - \bar{\mathbf{p}})^{\mathrm{T}} \Sigma_i^{-} (\mathbf{p}_i - \bar{\mathbf{p}})$.

$$\Sigma_i = \frac{1}{N_i} \begin{bmatrix} \bar{p}_1(1 - \bar{p}_1) & -\bar{p}_1\bar{p}_2 & -\bar{p}_1\bar{p}_3 \ldots & -\bar{p}_1\bar{p}_K \\ \ldots & \bar{p}_2(1 - \bar{p}_2) & -\bar{p}_2\bar{p}_3 \ldots & -\bar{p}_2\bar{p}_K \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \bar{p}_K(1 - \bar{p}_K) \end{bmatrix}.$$

Now,

$$\begin{aligned} \boldsymbol{\Sigma}_i^{-} &= N_i \boldsymbol{\Sigma}^{-} \\ &= N_i \text{diag.}[\bar{p}_1^{-1}, \bar{p}_2^{-1}, \bar{p}_3^{-1}, \ldots, \bar{p}_K^{-1}] \\ &= \begin{bmatrix} \frac{N_i}{\bar{p}_1} & 0 & 0 & 0 \\ 0 & \frac{N_i}{\bar{p}_2} & 0 & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \frac{N_i}{\bar{p}_K} \end{bmatrix} \end{aligned}$$

Thus,

$$D_i^2 = [(p_{i1} - \bar{p}_1) \ \ldots \ \ldots \ (p_{iK} - \bar{p}_K)] \begin{bmatrix} \frac{N_i}{\bar{p}_1} & 0 & 0 & 0 \\ 0 & \frac{N_i}{\bar{p}_2} & 0 & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \frac{N_i}{\bar{p}_K} \end{bmatrix} \begin{bmatrix} p_{i1} - \bar{p}_1 \\ p_{i2} - \bar{p}_2 \\ \ldots \\ \ldots \\ \ldots \\ p_{iK} - \bar{p}_K \end{bmatrix}$$

$$= \left[\frac{N_i(p_{i1} - \bar{p}_1)}{\bar{p}_1} \frac{N_i(p_{i2} - \bar{p}_2)}{\bar{p}_2} \ldots \frac{N_i(p_{iK} - \bar{p}_K)}{\bar{p}_K} \right] \begin{bmatrix} p_{i1} - \bar{p}_1 \\ p_{i2} - \bar{p}_2 \\ \ldots \\ \ldots \\ \ldots \\ p_{iK} - \bar{p}_K \end{bmatrix}$$

$$= \sum_{j=1}^{K} \frac{N_i(p_{ij} - \bar{p}_j)^2}{\bar{p}_j}.$$

The distributional aspects of Mahalanobis distance

Mahalanobis distance (D^2) underlies Hotelling's T^2 statistic [5]. In the present case, as mentioned earlier, the interest is to compare the rows of the data matrix, i.e. the proportion of defective vector corresponding to a particular time point with the average proportion defective vector.

If $\bar{\mathbf{X}}$ and $\mathbf{S}$ are the mean vector and covariance matrix of a sample of size n from $N_P[\boldsymbol{\mu}, \Sigma]$, then $(n-1)(\bar{\mathbf{X}}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{S}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu}) \sim \mathbf{T}^2(P, n-1)$, where $\bar{\mathbf{X}}$ is a $(P \times 1)$ vector [7].

The present distance formula is the following:

$$D^2_{\mathbf{p}_i,\bar{\mathbf{p}}} = N_i \sum_{j=1}^{K} \frac{(p_{ij}-\bar{p}_j)^2}{\bar{p}_j},$$

where $\mathbf{p}_i' = [p_{i1}, p_{i2}, \ldots, p_{iK}]$ and $\bar{\mathbf{p}}_j = [\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_K]$ and N_i is the sample size at the ith time point.

When $\mathbf{X} \sim N_P[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, $D^2_{\bar{X},\mu} = (\bar{\mathbf{X}}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{S}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu})$ and $(n-1)D^2_{\bar{X},\mu}$ follows $T^2_{P,(n-1)}$. Whereas when $p_i \sim$ multinomial$[\bar{\mathbf{p}}, N_i]$ with K categories and variance-covariance matrix is given by $\mathbf{S}_i$, $D^2_{\mathbf{p}_i,\bar{\mathbf{p}}} = N_i(\mathbf{p}_i-\bar{\mathbf{p}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-}(\mathbf{p}_i-\bar{\mathbf{p}})$ and $D^2_{\mathbf{p}_i,\bar{\mathbf{p}}} \sim \mathbf{T}^2_{K-1,N_i}$.

D^2 (multi-normal case) between sample average and population parameter uses the estimate of population $\boldsymbol{\Sigma}$ given by $\mathbf{S}$ and not the variance-covariance matrix of the sample $\bar{\mathbf{X}}$, which will actually be given by $(n-1)^{-1}\mathbf{S}$. That is why the need arises for multiplying D^2 by $(n-1)$ so that $(n-1)D^2 \sim \mathbf{T}^2$. However, in the case of the multinomial situation the metric used is $\boldsymbol{\Sigma}_i$, which already involves the N_i factor. Hence, multiplication of D^2 by a N_i factor is no longer necessary.

The second degrees of freedom for T^2 in the multinomial case is N_i itself, not (N_i-1) as in the multi-normal case because $\boldsymbol{\Sigma}_i$ is an unbiased estimate of the population $\boldsymbol{\Sigma}$ with sample size N_i.

The first degrees of freedom for T^2 in multinomial case is $(K-1)$ and not K as in multi-normal case because $\sum_{j=1}^{K} p_{ij} = 1$. This sort of a constraint is absent in a multi-normal case. Here, $\mathrm{T}^2_{(K-1),N_i,\alpha} = [(N_i(K-1))/(N_i-K+2)]\mathbf{F}_{K-1,N_i-K+2,\alpha}$ which is the upper control limit for the D^2 control chart at the α level of significance. The lower control limit is obviously zero.

Discussion on the distributional assumption

As one looks at the distance formula for the multinomial case it may appear that a chi-square distribution is appropriate. But in order to use the chi-square distribution one needs to ensure that the expected frequency in each category is at least five. This may not be possible to obtain whenever sample size is small or the average proportion defective for some categories is too small a number. Of course one can club together some such adjacent categories but this will lead to undesirable complication of the matter because of variation in K.

Also the assumption of a chi-square distribution ignores the effect of sample size on the control limit, which a T^2 distribution will retain by resembling the traditional p-chart. Thus, the construction of control limits based upon the T^2 distribution is more justified.

Merits of the D^2 control chart

- D^2 control chart exercises simultaneous control on proportion defectives falling in various categories of defects in a single chart without any distortion in the advertised level of type I error.

- While constructing a p-chart it is a standard practice to draw p-charts for major defects only along with the overall p-chart. Hence, the conventional p-charts do not take care of the fluctuations involved in the categories of minor defects – the cumulative effect of which may very well destabilize the process at times. However, the D^2 control chart takes into account various categories of defects exhaustively and thereby enhances the sensitivity in the detection of a shift. Therefore, to understand the overall performance of a process by considering all categories of defects at the same time, the D^2 control chart is very effective.

A case example

This case example is concerned with the proportion defective data with regard to various kinds of paint defects of a ceiling fan cover. The defects, which are prevalent in painting of such covers, are poor covering, overflow, patty defect, bubbles, paint defects, buffing defects.

Apart from the six categories of defects the good items are also considered for computing the D^2 statistic. Hence, the value of K in this particular case is 7 and the value of n or the number of objects or time periods is 24. For any time period $\sum_{j=1}^{7} p_{ij} = 1, i = 1, 2, 3, \ldots, 24$. Here p_{ij} is the proportion defective or proportion good item at a particular time period i for each j. It can be seen from Table 1 that the sample size ranges from 20 to 404. The D^2 chart and the corresponding overall p-chart and p-charts for the individual defect categories are given in Figures 1 to 8 for ready comprehension. Note that this is a perfect multinomial case in the sense that as soon as a particular cover is found to contain a defect it is categorized into the most predominant defect it has within. Hence, a particular defective cover contains one and only one kind of defect.

Table 1. Data for paint defect.

Time period	No. inspected	Poor covering	Overflow	Patty defect	Bubbles	Paint defect	Buffing	Total
1	176	15	12	2	6	3	4	42
2	160	13	18	4	4	5	5	49
3	186	18	12	6	8	4	6	54
4	167	12	10	5	4	3	2	36
5	291	8	6	4	3	2	5	28
6	170	10	6	2	3	2	3	26
7	224	15	12	4	3	6	5	45
8	140	11	9	5	4	3	4	36
9	250	19	15	7	4	8	5	58
10	145	20	10	4	2	6	2	44
11	100	15	12	2	3	3	4	39
12	80	4	2	1	2	2	1	12
13	170	13	11	5	1	4	3	37
14	200	22	17	4	3	6	2	54
15	112	8	6	2	3	1	2	22
16	250	15	11	4	5	7	0	42
17	122	12	9	6	8	4	5	44
18	312	23	18	4	7	5	9	66
19	200	15	13	4	6	7	5	50
20	20	3	2	2	1	0	0	8
21	60	6	4	2	2	3	2	19
22	404	40	15	6	8	4	20	93
23	104	8	6	2	3	2	4	25
24	124	12	9	3	4	2	5	35
Total	4167	337	245	90	97	92	103	964

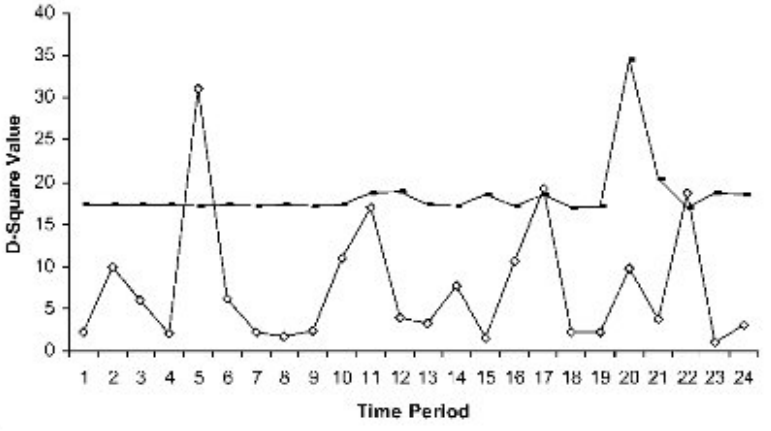

Figure 1. D^2 chart.

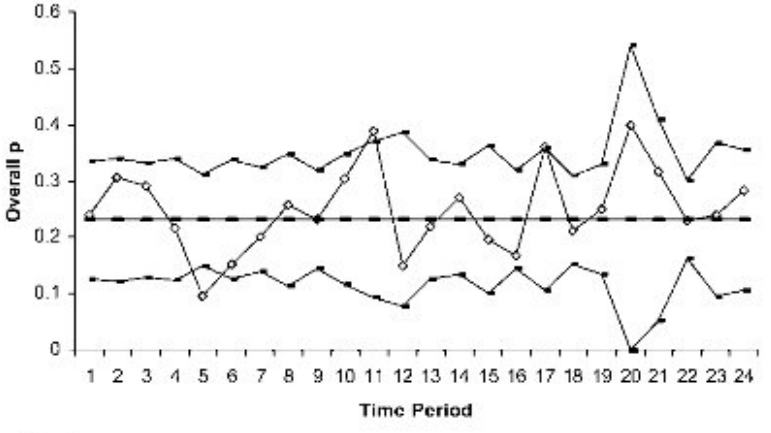

Figure 2. Overall p-chart.

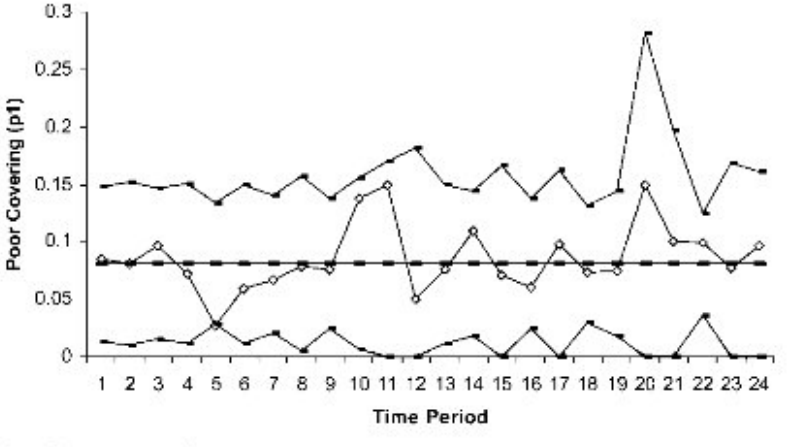

Figure 3. p-chart for poor covering.

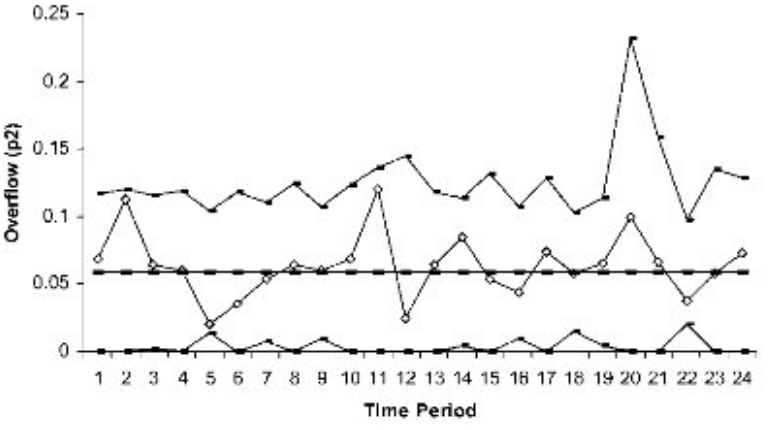

Figure 4. *p*-chart for overflow.

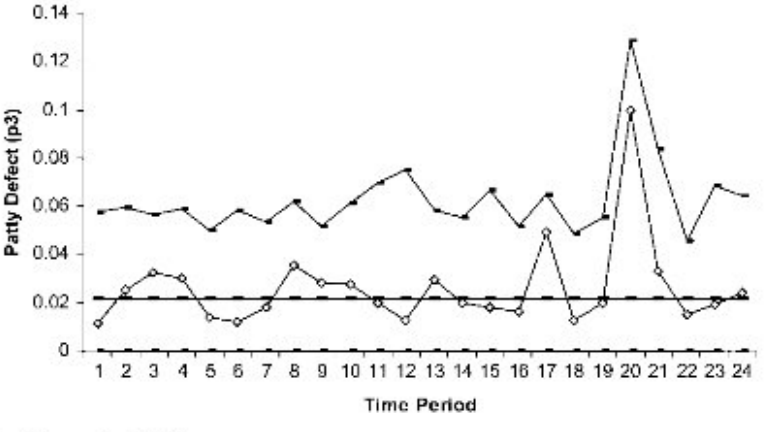

Figure 5. *p*-chart for patty defect.

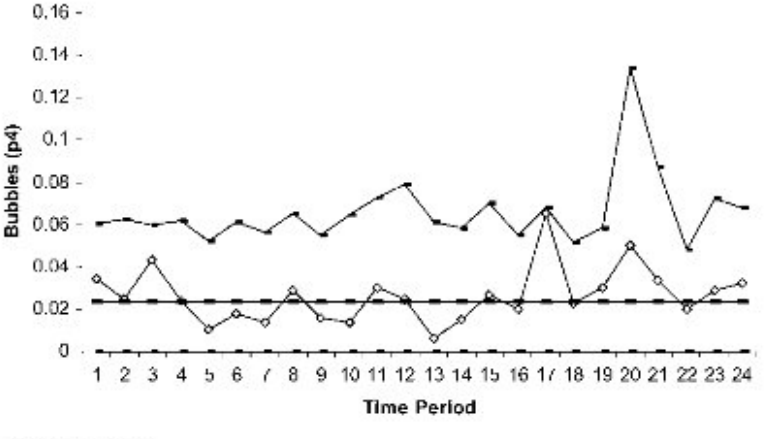

Figure 6. *p*-chart for bubbles.

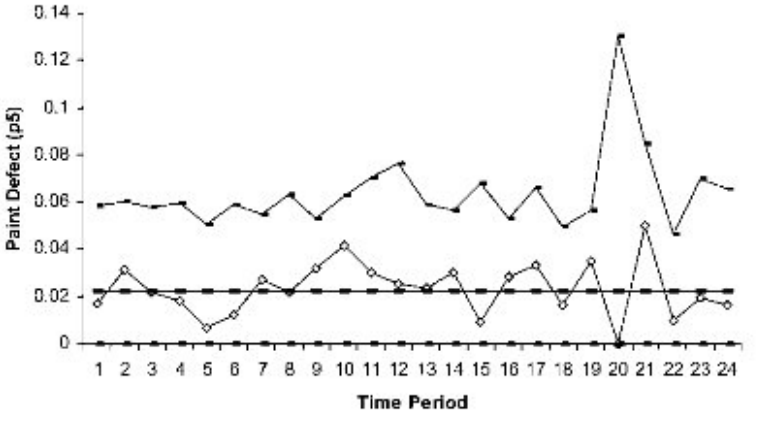

Figure 7. p-chart for paint defect.

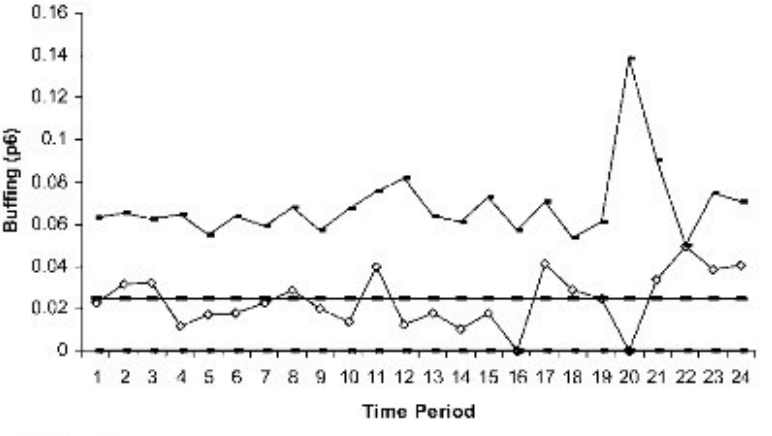

Figure 8. p-chart for buffing.

Observation and interpretation

It may be observed from the D^2 chart that the points 5, 17, and 22 have fallen outside the control limit (1%). In addition, it may also be noted that point number 11 (D^2 value 16.9897) has fallen outside the warning limit (5%), the value of which is 13.7053.

Note that if α' is the type I error for the D^2 chart and α is the type I error for the individual p-charts then $\alpha' = 1 - (1 - \alpha)^K$. For $\alpha' = 0.01$, α is found to be 0.001 for $K = 7$ categories. So for individual p-charts 3.19σ control limits have been used.

At the 5th time period the overall proportion defective and the proportion defective due to poor covering lie below their respective lower control limit. Other proportion defectives do not show any lack of statistical control but their simultaneous low values have got a conspicuous impact on the D^2 chart.

At the 17th time period the overall proportion defective has fallen outside the upper control limit and the 'patty defect' and 'bubbles' have fallen on the higher side nearer to their respective upper control limits. The D^2 chart has revealed that.

At the 22nd time period the defect named as 'buffing' shows an out of control situation since the corresponding proportion defective falls beyond the upper control limit.

References

[1] T.W. Anderson, *et al.*, *Multivariate Analysis and Its Application*, Institute of Mathematical Statistics, California, 1994.
[2] M. Jambu, *Exploratory and Multivariate Data Analysis*, Academic Press, Boston, 1991.
[3] A.M. Kshirsagar, *Multivariate Analysis*, Marcel Dekker, New York, 1972.
[4] K.V. Mardia, J.T. Kent, and J.M. Bibby, *Multivariate Analysis*, Academic Press, London, 1979.
[5] D.C. Montgomery, *Introduction to Statistical Quality Control*, Wiley, New York, 2001.
[6] A.R. Rao and P. Bhimasankaram, *Linear Algebra*, McGraw Hill, New Delhi, 1992.
[7] A.J. Richard and W.W. Dean, *Applied Multivariate Statistical Analysis*, Prentice Hall of India Pvt. Ltd, India, 1996.