# Count Data Monitoring: Parametric or Nonparametric?

Zhiqiong Wang[1] and Peihua Qiu[2]

[1]College of Management and Economics, Tianjin University

[2]Department of Biostatistics, University of Florida

### Abstract

Count data are common in practice, ranging from security protection, disease surveillance, to quality monitoring of a production process. To describe the distribution of a count data, we usually use a Poisson probability model or a similar parametric model (e.g., a Negative Binomial model). In practice, however, such a parametric model may not be able to describe the distribution of a count data well in some cases, because the count data are often affected by some confounding factors and such a confounding impact is difficult to accommodate by the parametric model. In this paper, we study the count data monitoring problem, and the consequence to use a parametric control chart in cases when the underlying parametric distribution model is invalid. Based on that study, we suggest using nonparametric charts to monitor count data when it is uncertain that the count data can be described by a parametric distribution model.

**Keywords:** Count data; Data categorization; Distribution free; Parametric probability models; Poisson distribution; Statistical process control.

## 1 Introduction

In many sequential processes, major outcome variables are often in the form of counts of certain events that are related to the quality or performance of the related processes. For instance, in manufacturing industries, the number of specific defects found in a product sampled from a production line could be a quality variable to monitor. In road or internet traffics, the number of accidents is often a major index of concern. In public health, we are often concerned about the number of daily occurrences of a specific disease (e.g., lung cancer) in a region or country. Therefore, proper monitoring of count data is an important research problem with broad applications, which is the focus of this paper.

In the statistical process control (SPC) literature, there have been some control charts for monitoring count data. In some applications, it is obvious that the distribution of the count data is binomial, because the count denotes the number of successes from a binomial experiment. Many control charts, including the classic $p$ and $mp$ charts, have been developed for monitoring binomial count data. See, for instance, Gan[1], Wu et al[2], Megahed et al[3], and some others. In many other applications, the count data record the number of events occurring in a given period of time and a given region. Most existing control charts for handling such count data are based on the Poisson distribution assumption. For instance, the classical $c$, $u$ and $D$ charts are based on batch data, and they are Shewhart charts that make decisions based on the observed data at the current time point only (cf Section 3.3, Qiu[4]). It has been well demonstrated that Shewhart charts are good in detecting large shifts and ineffective in detecting small shifts. So, in the literature, some authors proposed the cumulative sum (CUSUM) charts[5,6] and the exponentially weighted moving average (EWMA) charts[7,8] for monitoring homogeneous Poisson processes. Besides a Poisson distribution, some authors suggested using a negative binomial or another parametric distribution for describing count data. Control charts based on such assumptions include those discussed in Sheaffer and Leavenworth[9], Saghir and Lin[10], and many others.

In practice, however, it is often difficult to describe count data using a parametric distribution. One major reason is that the count data are often affected by many different factors. Some of them may not be our major concern so that they are not part of the collected data, some might be difficult to measure, some others might even be hard to be noticed of their existence. Thus, the impact of such factors on the count data is difficult to describe properly by a parametric model. Figure 1 shows the density histogram of the monthly counts of polio cases in US from January 1970 to December 1972, along with the estimated density curve (solid) and the density curve of a Poisson distribution (dashed) with the same mean as that of the data. From the plot, it can be seen that the distribution of the observed data is quite different from the Poisson distribution in that the observed data seem to have a heavier right tail. The Pearson's Chi-square goodness-of-fit test and the Fisher's index of dispersion test[11] give p-values of 0.014 and 0.000, respectively, which confirms that the distribution of the observed data is indeed significantly different from the Poisson distribution.

In cases when the quality variables are continuous variables, it has been well demonstrated in the literature that the parametric control charts are not reliable to use when the underlying parametric model is invalid (e.g. Chakraborti et al[12]; Chapters 8 and 9, Qiu[4]; Qiu[13]). In such cases, nonparametric or distribution-free charts are recommended. For count data, it will be shown in this paper that this is also true. In the next section, we will adapt several existing nonparametric control charts for monitoring count data. Then, we will evaluate their in-control (IC) and out-of-control (OC) performance in Section 3, together with several representative parametric and nonparametric control charts. Application of the related control charts to the US polio data is discussed in Section 4. Several remarks conclude the article in Section 5.

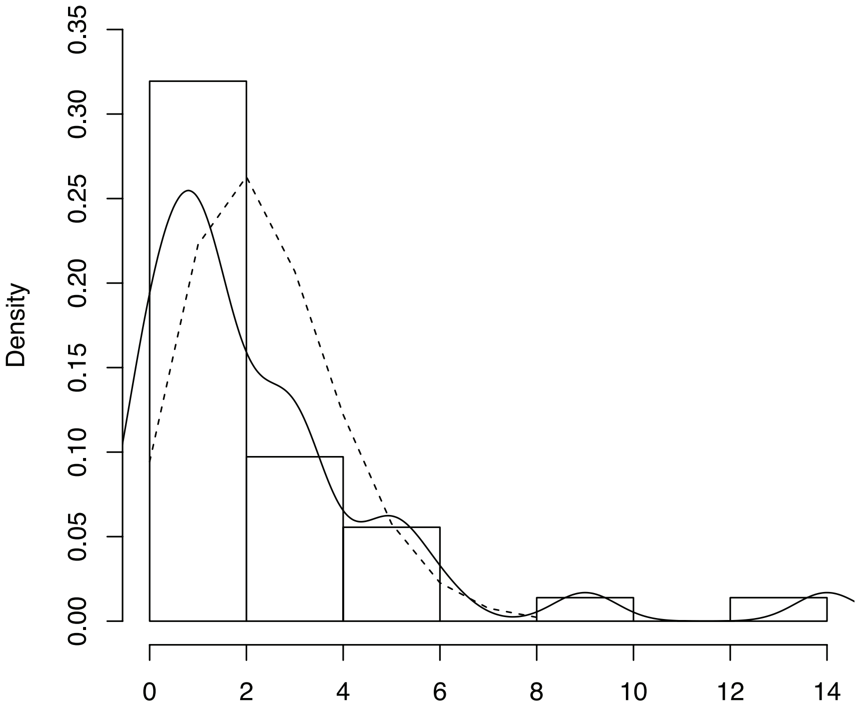

**Figure 1.** Density histogram with estimated density curve (solid) of the monthly counts of polio cases in US from January 1970 to December 1972, and the density curve of a Poisson distribution (dashed) with the same mean as that of the polio data.

## 2 Nonparametric Control Charts for Monitoring Count Data

In the literature, there are many nonparametric or distribution-free charts available. Most of them are designed for monitoring continuous numerical quality variables. They can be adapted properly for monitoring count data. In this section, we briefly discuss how to adapt the ones in Qiu and Li[14] for monitoring count data. Adaptation of other nonparametric charts can be discussed similarly.

Let $X(n)$ be an observation obtained at the $n$th time point during Phase II online monitoring, and $I_1 = [0, q_1)$, $I_2 = [q_1, q_2)$, $\ldots$, $I_p = (q_{p-1}, \infty)$ be a partition of $[0, \infty)$, where $0 < q_1 < q_2 < \ldots < q_{p-1} < \infty$ are $p-1$ boundary points of the partitioning intervals. Then, we first discretize $X(n)$ as follows:

$$Y_l(n) = \mathbf{I}(X(n) \in I_l), \quad \text{for } l = 1, 2, \ldots, p, \tag{1}$$

where $\mathbf{I}(a) = 1$ if $a$ is "true" and 0 otherwise. From Equation (1), we can see that $Y_l(n)$ indicates whether $X(n)$ belongs to the $l$th interval $I_l$. So, $\mathbf{Y}(n) = (Y_1(n), Y_2(n), \ldots, Y_p(n))'$ has one and only one component being 1, and the index of the component being 1 has a discrete distribution with probabilities $f_l = P(X(n) \in I_l)$, for $l = 1, 2, \ldots, p$. For convenience of presentation, the distribution $\mathbf{f} = (f_1, f_2, \ldots, f_p)'$

is also called the distribution of $\mathbf{Y}(n)$. Let $\mathbf{f}^{(0)} = (f_1^{(0)}, f_2^{(0)}, \ldots, f_p^{(0)})'$ be the IC distribution of $\mathbf{Y}(n)$ and $\mathbf{f}^{(1)} = (f_1^{(1)}, f_2^{(1)}, \ldots, f_p^{(1)})'$ be its OC distribution. Then, under some regularity conditions, it can be checked that $\mathbf{f}^{(1)}$ is different from $\mathbf{f}^{(0)}$ if there is a mean shift in $X(n)$.

In Qiu and Li[14], it was suggested that the parameters $\{q_1, q_2, \ldots, q_{p-1}\}$ were chosen such that $\mathbf{f}^{(0)} \approx (1/p, 1/p, \ldots, 1/p)'$. For count data, this is often difficult to achieve because of the discreteness of the data. In such cases, we suggest choosing the parameters such that $\mathbf{f}^{(0)}$ is as close to a uniform distribution as possible. The control chart suggested in Qiu and Li[14] is the CUSUM version of the Pearson's chi-squared test. More specifically, let $\mathbf{S}_0^{\text{obs}} = \mathbf{S}_0^{\text{exp}} = \mathbf{0}$ be two $p \times 1$ column vectors, and

$$\begin{cases} \mathbf{S}_n^{\text{obs}} = \mathbf{0}, & \text{if } C_n \leq k_P, \\ \mathbf{S}_n^{\text{exp}} = \mathbf{0}, & \text{if } C_n \leq k_P, \\ \mathbf{S}_n^{\text{obs}} = (\mathbf{S}_{n-1}^{\text{obs}} + \mathbf{Y}(n))(C_n - k_P)/C_n, & \text{if } C_n > k_P, \\ \mathbf{S}_n^{\text{exp}} = (\mathbf{S}_{n-1}^{\text{exp}} + \mathbf{f}^{(0)})(C_n - k_P)/C_n, & \text{if } C_n > k_P, \end{cases}$$

where

$$C_n = \left( \left(\mathbf{S}_{n-1}^{\text{obs}} - \mathbf{S}_{n-1}^{\text{exp}}\right) + \left(\mathbf{Y}(n) - \mathbf{f}^{(0)}\right) \right)' \times \left( \text{diag}\left(\mathbf{S}_{n-1}^{\text{exp}} + \mathbf{f}^{(0)}\right) \right)^{-1} \times \left( \left(\mathbf{S}_{n-1}^{\text{obs}} - \mathbf{S}_{n-1}^{\text{exp}}\right) + \left(\mathbf{Y}(n) - \mathbf{f}^{(0)}\right) \right),$$

$k_P \geq 0$ is the so-called allowance parameter, diag($\mathbf{a}$) denotes a diagonal matrix with its diagonal elements being the corresponding elements of the vector $\mathbf{a}$, and the superscripts "obs" and "exp" denote observed and expected counts, respectively. Define

$$u_{n,P} = \left(\mathbf{S}_n^{\text{obs}} - \mathbf{S}_n^{\text{exp}}\right)' \left(\text{diag}(\mathbf{S}_n^{\text{exp}})\right)^{-1} \left(\mathbf{S}_n^{\text{obs}} - \mathbf{S}_n^{\text{exp}}\right).$$

Then, a mean shift in $\mathbf{X}(n)$ is signaled if

$$u_{n,P} > h_P, \tag{2}$$

where $h_P > 0$ is a control limit chosen to reach a pre-specified $\text{ARL}_0$ value. When $k_P = 0$, it is not difficult to check that $\mathbf{S}_n^{\text{obs}}$ is a frequency vector with its $l$th element denoting the cumulative observed count of observations in the $l$th interval $I_l$ as of the time point $n$, for $l = 1, 2, \ldots, p$, and $\mathbf{S}_n^{\text{exp}}$ is the vector of the corresponding cumulative expected counts. Therefore, $u_{n,P}$ is the Pearson's chi-squared test statistic in such cases to measure the difference between the cumulative observed and expected counts as of the time point $n$. Because the charting statistic $u_{n,P}$ can only take some discrete values on the positive number line, we often cannot find a proper $h_P$ value so that a pre-specified nominal $\text{ARL}_0$ value is reached within a desired precision. To overcome this limitation, we can apply the modification procedure suggested in Qiu and Li[14] by adding a small random number generated from $\text{N}(0, s^2)$ to each component of $\mathbf{Y}(n)$ to alleviate the discreteness of $u_{n,P}$. As long as $s$ is chosen small, the OC behavior of the chart would not change much. But, most nominal $\mathrm{ARL}_0$ values can be reached within a desired precision after the modification. In all the simulation studies in this paper, we choose $s = 0.01$.

In the above CUSUM chart, usually the allowance constant $k_P$ is specified beforehand. Then, the control limit $h_P$ is chosen such that a pre-specified $\mathrm{ARL}_0$ value is reached. To compute $h_P$ from an IC data, we suggest using the following four-step, iterative, bisection searching algorithm based on the bootstrap resampling idea.

Step 1: In the $i$th iteration, $h_P$ is searched in the interval $[L^{(i)}, U^{(i)}]$. When $i = 1$, define $L^{(1)} = 0$ and $U^{(1)} = U$, where $U$ is an upper bound satisfying the condition that the $\mathrm{ARL}_0$ value of the P-CUSUM chart when $h_P = U$ is large than the pre-specified $\mathrm{ARL}_0$.

Step 2: A sequence of observations is selected randomly with replacement from the IC dataset and this sequence of data is sued as Phase II observations. Then, the P-CUSUM chart with $h_P = h^{(i)} = (L^{(i)} + U^{(i)})/2$ is applied to this sequence and the run length is recorded.

Step 3: Step 2 is repeated for $N$ times, and the actual $\mathrm{ARL}_0$ value is approximated by the average of the $N$ run lengths, denoted as ${\mathrm{ARL}_0}^{(i)}$.

Step 4: If $|{\mathrm{ARL}_0}^{(i)} - \mathrm{ARL}_0| < \varepsilon$, where $\varepsilon$ is a small number and denotes the required searching accuracy, then the whole algorithm stops and the searched value of $h_P$ is $h^{(i)}$. Otherwise, define

$$\begin{aligned}
&L^{(i+1)} = h^{(i)} \text{ and } U^{(i+1)} = U^{(i)}, \text{ if } {\mathrm{ARL}_0}^{(i)} < \mathrm{ARL}_0,\\
&L^{(i+1)} = L^{(i)} \text{ and } U^{(i+1)} = h^{(i)}, \text{ if } {\mathrm{ARL}_0}^{(i)} > \mathrm{ARL}_0,\\
&\text{and } h^{(i+1)} = (L^{(i+1)} + U^{(i+1)})/2,
\end{aligned}$$

and the algorithm executes the next iteration until the maximum number of iterations, say $Q$, is reached.

The above searching algorithm usually converges quickly. Although it is rare, if it does not stop before the $Q$th iteration, then define $h_P = h^{(Q)}$.

Besides the CUSUM chart (2) that is based on the Pearson's chi-squared test, we can construct a similar CUSUM chart based on the likelihood ratio test. More specifically, let $\widetilde{\mathbf{S}}_n^{\mathrm{obs}}$ and $\widetilde{\mathbf{S}}_n^{\mathrm{exp}}$ be quantities defined in the same way as $\mathbf{S}_n^{\mathrm{obs}}$ and $\mathbf{S}_n^{\mathrm{exp}}$ used in chart (2), except that $k_P$ is replaced by another constant $k_L$ and $C_n$ is replaced by

$$\widetilde{C}_n = 2\big(\widetilde{\mathbf{S}}^{\text{obs}}_{n-1} + \mathbf{Y}(n)\big)'\log\left(\frac{\widetilde{\mathbf{S}}^{\text{obs}}_{n-1} + \mathbf{Y}(n)}{\widetilde{\mathbf{S}}^{\text{exp}}_{n-1} + \mathbf{f}^{(0)}}\right),$$

where $\mathbf{a}/\mathbf{b}$ denotes a vector obtained by component-wise division of the vector $\mathbf{a}$ by the vector $\mathbf{b}$, and $\log(\mathbf{a}/\mathbf{b})$ denotes a component-wise operation as well. Then, the new charting statistic is defined as

$$u_{n,L} = 2\big(\widetilde{\mathbf{S}}^{\text{obs}}_{n}\big)'\log\left(\frac{\widetilde{\mathbf{S}}^{\text{obs}}_{n}}{\widetilde{\mathbf{S}}^{\text{exp}}_{n}}\right).$$

It gives a signal when

$$u_{n,L} > h_L, \tag{3}$$

where the control limit $h_L > 0$ can be chosen similarly to $h_P$.

## 3 Performance Assessment

We present some simulation results in this section to compare the performance of the nonparametric CUSUM charts (2) (denoted as P-CUSUM) and (3) (denoted as L-CUSUM) with some parametric and nonparametric charts for monitoring count data. In the simulation study, we investigate the performance of different control charts under various different discrete distributions. For describing count data, Poisson distribution is a standard model. One important property of a Poisson distribution with a rate parameter $\lambda$, denoted as Poisson$(\lambda)$, is that its mean and variance are the same to be $\lambda$. In practice, however, count data can be over-dispersed (i.e., the variance is larger than the mean) or under-dispersed (i.e., variance is smaller than the mean). In this section, we consider the following discrete distributions that are all different from the regular Poisson distribution:

(I) Binomial distribution: cases with under-dispersion

Let $X$ denote the number of successes in $r$ Bernoulli trials with probability of success being $\pi$. Then, $X \sim \text{Bin}(r,\pi)$, with mean $r\pi$ and variance $r\pi(1-\pi)$. The index of dispersion, defined as the ratio of variance to mean, is then $1-\pi$. Because $\pi$ is usually in the interval $(0,1)$, $0 < 1-\pi < 1$, implying that the Binomial distribution is under-dispersed.

(II) Negative binomial distribution: cases with over-dispersion

Let $X$ be the random variable representing the number of failures in a sequence of i.i.d. Bernoulli trials until the $r$th success. Then, $X$ has a negative binomial distribution, denoted as $X \sim \text{NB}(r,\pi)$. The mean and variance of $X$ are $r(1-\pi)/\pi$ and $r(1-\pi)/\pi^2$, respectively. So, the index of dispersion is $1/\pi$, which is larger than 1 if $\pi < 1$.

(III) Cases with mixed-dispersion

(i) Discrete uniform distribution

Let $X$ be the random variable that takes integer values in $\{0,1,2,\ldots,r\}$ with equal probabilities. Then, it has a discrete uniform distribution, denoted as $X \sim \mathrm{DU}(r)$. It can be checked that the mean and variance of $X$ are $r/2$ and $r(r+2)/12$, respectively, and the index of dispersion is $(r+2)/6$. Therefore, this index can be larger or smaller than 1, depending on the value of $r$.

(ii) Generalized Poisson distribution

The probability mass function of the generalized Poisson distribution $\mathrm{GP}(\eta,\theta)$ is given by

$$f(x;\eta,\theta) = \begin{cases} 0, & \text{when } x > m \text{ and } \theta < 0, \\ \frac{\eta(\eta+\theta x)^{x-1}\exp(-(\eta+\theta x))}{x!}, & \text{otherwise,} \end{cases}$$

where $\eta > 0$ and $\max(-1,-\eta/m) \le \theta < 1$ are two parameters, and $m \ge 4$.[15] When $\theta < 0$, $m$ is the largest positive integer so that $\eta + m\theta > 0$. If $\theta = 0$, then the distribution $\mathrm{GP}(\eta,\theta)$ reduces to the standard Poisson distribution with mean $\lambda = \eta$. It can be checked that the mean and variance of $X$ are $\eta/(1-\theta)$ and $\eta/(1-\theta)^3$, respectively. Thus, the index of dispersion is $1/(1-\theta)^2$, which can be larger or smaller than 1.

We first investigate the IC performance of the nonparametric charts P-CUSUM and L-CUSUM. For comparison purposes, besides these two charts, we also consider the traditional Poisson CUSUM chart[5], that is constructed and designed based on the assumption that the related process has a regular Poisson distribution. This chart is denoted as T-CUSUM. Also, the ideal CUSUM chart that is constructed and designed using the true process distribution is considered as a gold standard. This chart is denoted as I-CUSUM. Furthermore, we consider the nonparametric CUSUM chart based on the Wilcoxon rank-sum test that was discussed in Li et al[16]. This chart is denoted as W-CUSUM. The allowance constants in all CUSUM charts are chosen to be 0.5. The assumed $\mathrm{ARL}_0$ value is chosen to be 200 or 500. In charts P-CUSUM and L-CUSUM, the number of categories $p$ is chosen to be 5. Their control limits are searched by the bootstrap procedure with $Q = 100$ iterations. In each iteration, the $\mathrm{ARL}_0$ value is computed from $10{,}000$ replicated simulation runs, in each of which the bootstrap re-sampling procedure is applied to an IC dataset of size $M = 500$. The control limit of T-CUSUM is determined based on the Poisson distribution assumption, and the control limit of I-CUSUM is computed based on the true IC process distribution.

Then, we consider the following cases when the true process distribution is the standard version with mean 0 and standard deviation 1 of one of the following five distributions: $\mathrm{Bin}(20,0.75)$, $\mathrm{DU}(10)$, $\mathrm{NB}(20,0.75)$, $\mathrm{GP}(5,0.25)$ and $\mathrm{GP}(5,-0.25)$. These five cases represent different scenarios when the true process distribution is under-dispersed or over-dispersed. The actual $\mathrm{ARL}_0$ values of all five charts and their standard errors are summarized in Table 1.

**Table 1.** Actual $\mathrm{ARL}_0$ values and their standard errors (in parentheses) of the five charts when the nominal $\mathrm{ARL}_0$ values are fixed at 200 and 500.

| $\mathrm{ARL}_0$ | Chart | $\mathrm{Bin}(20,0.75)$ | $\mathrm{DU}(10)$ | $\mathrm{NB}(20,0.75)$ | $\mathrm{GP}(5,0.25)$ | $\mathrm{GP}(5,-0.25)$ |
|---|---|---|---|---|---|---|
| 200 | T-CUSUM | 322.7 (3.21) | 385.2 (3.82) | 188.2 (1.86) | 162.9 (1.60) | 273.9 (2.59) |
| | I-CUSUM | 199.9 (1.95) | 199.9 (1.94) | 199.9 (1.94) | 199.9 (1.95) | 200.0 (1.94) |
| | W-CUSUM | 199.9 (1.84) | 200.0 (1.86) | 199.9 (1.85) | 200.1 (1.87) | 200.0 (1.82) |
| | L-CUSUM | 200.0 (1.91) | 200.0 (1.94) | 200.1 (1.93) | 199.9 (1.93) | 200.1 (1.93) |
| | P-CUSUM | 200.0 (2.06) | 200.0 (2.06) | 199.9 (2.06) | 199.5 (2.05) | 199.9 (2.05) |
| 500 | T-CUSUM | 939.9 (9.38) | 1241.7 (12.20) | 393.1 (3.89) | 336.9 (3.21) | 746.9 (7.30) |
| | I-CUSUM | 500.0 (4.95) | 500.1 (4.87) | 499.8 (4.92) | 500.0 (4.81) | 500.0 (4.87) |
| | W-CUSUM | 499.7 (4.76) | 500.1 (4.75) | 500.1 (4.81) | 500.0 (4.71) | 500.0 (4.70) |
| | L-CUSUM | 499.9 (4.92) | 499.9 (4.95) | 499.9 (4.91) | 500.0 (4.88) | 500.0 (4.91) |
| | P-CUSUM | 500.0 (5.00) | 500.0 (5.03) | 500.0 (5.02) | 500.0 (5.17) | 499.9 (5.08) |

From Table 1, it can be seen that the actual $\mathrm{ARL}_0$ values of the I-CUSUM chart are indeed close to the nominal $\mathrm{ARL}_0$ values, as expected. The actual $\mathrm{ARL}_0$ values of P-CUSUM, L-CUSUM and W-CUSUM charts perform as well as the I-CUSUM chart. These results confirm that the IC performance of all P-CUSUM, L-CUSUM and W-CUSUM charts does not depend on the true IC process distribution. Therefore, they are indeed robust to the true IC process distribution. Now, for the chart T-CUSUM that is based on the Poisson distribution assumption, we can see that its actual $\mathrm{ARL}_0$ values are quite different from the nominal values. More specifically, its actual $\mathrm{ARL}_0$ values are much larger than the nominal $\mathrm{ARL}_0$ values when the true process distribution is the standard version of $\mathrm{Bin}(20,0.75)$, $\mathrm{DU}(10)$ and $\mathrm{GP}(5,-0.25)$, and much smaller than the nominal $\mathrm{ARL}_0$ values when the true process distribution is the standard version of $\mathrm{NB}(20,0.75)$ and $\mathrm{GP}(5,0.25)$. In the former case when the actual $\mathrm{ARL}_0$ values are larger than the nominal values, some true process distributional shifts could be detected much later than what we would expect, which is not good because many defective products will be produced in such cases.[17] In the latter case when the actual $\mathrm{ARL}_0$ values are smaller than the nominal values, the related process would be stopped too often by the control chart when it is actually IC.[14,17] From these results, we can see that when process observations do not follow a pre-specified parametric distribution, the traditional CUSUM chart based on the pre-specified distribution is not reliable. To further illustrate this conclusion, Figure 2 shows the actual

$\mathrm{ARL}_0$ values of the T-CUSUM chart based on the assumptions that the IC process distribution is Poisson and $\mathrm{ARL}_0 = 500$, in cases when the actual IC process distribution is $\mathrm{GP}(\eta, \theta)$, where $\eta$ is fixed at 1, 5 or 10, and $\theta$ changes in $[-0.4, 0.4]$. From the figure, it can be seen that the actual $\mathrm{ARL}_0$ values could be substantially different from the nominal $\mathrm{ARL}_0$ value when $\theta$ moves away from 0. We would like to point out that when $\theta = 0$, $\mathrm{GP}(\eta, 0)$ becomes a regular Poisson distribution. Therefore, T-CUSUM performs well in such cases.

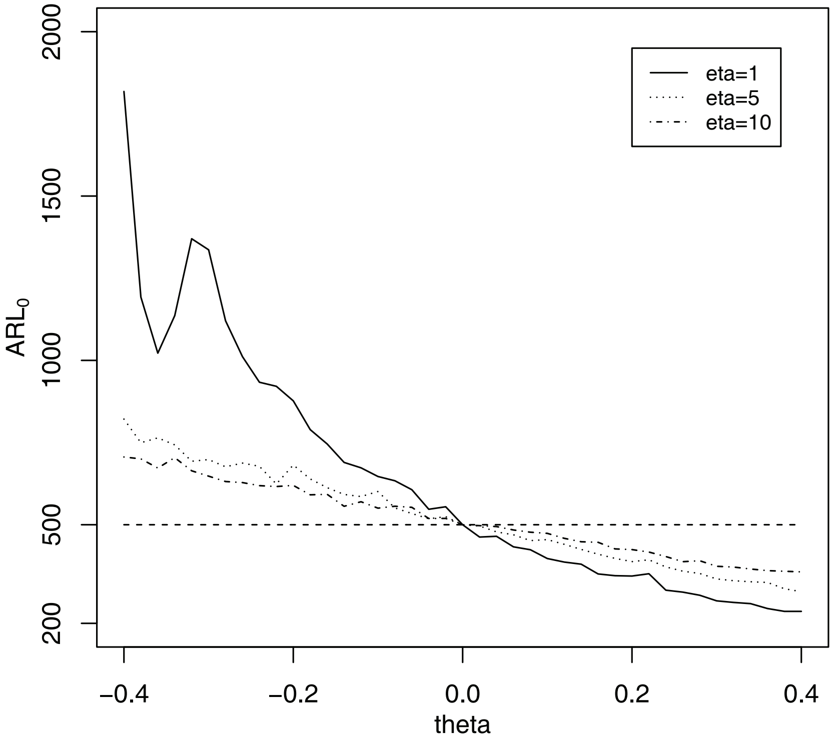

**Figure 2.** Actual $\mathrm{ARL}_0$ values of the the T-CUSUM chart in cases when the true IC process distribution is $\mathrm{GP}(\eta, \theta)$, with $\eta$ fixed at 1, 5 or 10 and $\theta$ changing in $[-0.4, 0.4]$.

To further demonstrate the distribution-free property of the nonparametric charts, the control limits of the P-CUSUM chart obtained based on 10,000 replicated simulations in cases when the allowance constant $k_P$ is chosen to be 0.01, 0.05 or 0.1, the nominal $\mathrm{ARL}_0$ is chosen to be 200 or 500, and when the true process distribution is one of the five parametric distributions considered in Table 1 are presented in Table 2. Results for the L-CUSUM chart are similar. So, they are omitted here. In the P-CUSUM chart, the number of categories $p$ is chosen to be 5, as before, and the IC distribution $\mathbf{f}^{(0)}$ of the categorized data is multinomial with probabilities $(0.2, 0.2, 0.2, 0.2, 0.2)$ for the five categories. Results in the first row labled as "Distribution-Free" are obtained when the categorized data $\mathbf{Y}(n)$ are generated from the IC distribution $\mathbf{f}^{(0)}$ directly, and the ones in the remaining rows are obtained when $\mathbf{Y}(n)$ are generated from the original process observations that follow the related distributions specified in the row labels. From the table, it can be seen that the control limit values of the P-CUSUM chart are indeed very close to each other in different rows when the true process distribution takes different parametric forms. Also, the actual $\text{ARL}_0$ values are very close to the nominal $\text{ARL}_0$ values as well in different scenarios.

**Table 2.** Control limits of the P-CUSUM chart when the nominal $\text{ARL}_0$ is fixed at 200 or 500, the allowance constant $k_P$ is chosen to be 0.01, 0.05 or 0.1, and the true process distribution takes different parametric forms. Numbers in parentheses are the corresponding actual $\text{ARL}_0$ values.

| | $k = 0.01$ | | $k = 0.05$ | | $k = 0.1$ | |
|---|---|---|---|---|---|---|
| | $\text{ARL}_0 = 200$ | $\text{ARL}_0 = 500$ | $\text{ARL}_0 = 200$ | $\text{ARL}_0 = 500$ | $\text{ARL}_0 = 200$ | $\text{ARL}_0 = 500$ |
| Distribution free | 6.722 (200.1) | 7.977 (500.0) | 7.923 (200.0) | 9.360 (500.1) | 8.472 (200.0) | 10.248 (500.0) |
| Bin(20,0.75) | 7.085 (200.1) | 8.197 (500.3) | 8.076 (199.9) | 9.573 (500.1) | 8.525 (199.7) | 10.392 (499.9) |
| DU(10) | 7.017 (200.0) | 8.169 (500.0) | 8.013 (200.0) | 9.549 (500.1) | 8.546 (200.0) | 10.408 (500.0) |
| NB(20,0.75) | 6.756 (200.0) | 7.953 (499.9) | 7.962 (200.0) | 9.379 (499.9) | 8.477 (200.0) | 10.283 (500.2) |
| GP(5,0.25) | 7.236 (200.0) | 8.461 (500.0) | 8.189 (200.0) | 9.854 (500.0) | 8.643 (200.0) | 10.554 (500.0) |
| GP(5,-0.25) | 7.188 (199.9) | 8.416 (500.1) | 8.079 (199.9) | 9.667 (500.0) | 8.577 (200.1) | 10.440 (500.1) |

Next, we compare the OC performance of the related control charts. From Table 1 and Figure 2, it can be seen that the T-CUSUM chart has unacceptable IC performance in various cases when the true process distribution is not Poisson. In such cases, its shift detection power may be irrelevant because a good power could be due to an overly small $\text{ARL}_0$ value. This is similar to the situation in hypothesis testing, where we should never consider a testing procedure whose actual significance level is larger than the nominal significance level. For this reason, the T-CUSUM chart is not considered in the comparison of the OC performance. Because the I-CUSUM chart needs to know the true IC distribution, which is often unrealistic in practice, it is not considered here either. For comparison purposes, we modify the T-CUSUM chart by using the bootstrap method described in Section 2 to determine its control limit so that the nominal $\text{ARL}_0$ value is reached. This modified T-CUSUM chart (denoted as T-CUSUM(adj)) is then included in the comparison.

In the OC performance comparison study, the true process distribution is assumed to be the standardized version with mean 0 and variance 1 of one of the following five distributions: $\text{Bin}(20,0.75)$, $\text{DU}(10)$, $\text{NB}(20,0.75)$, $\text{GP}(5,0.25)$ and $\text{GP}(5,-0.25)$. The process mean shift ranges from $-1.0$ to 1.0. For the P-CUSUM chart, we consider three versions with $p = 2,5$, or 10. For the L-CUSUM chart, $p$ is fixed at 5. For both of them, the IC distribution of the categorized data $\mathbf{f}^{(0)}$ is estimated from an IC dataset with size $M = 500$. For the W-CUSUM chart, an IC dataset of size 500 is used as a reference sample and all observations are batched with the batch size 5. The same IC dataset is used for the T-CUSUM(adj) chart to determine the control limit by the bootstrap method. Due to the fact that performance of different control charts with a same allowance constant may not be comparable, we choose to compare the optimal performance of all charts when detecting a specific shift, by selecting their allowance constants to minimize the $\mathrm{ARL}_1$ values while maintaining their $\mathrm{ARL}_0$ values all at 200. Based on 10,000 replications, the calculated $\mathrm{ARL}_1$ values of the related control charts are shown in Figure 3, where the y-axis is in natural logarithm scale to better demonstrate the difference among different control charts. From the plots of Figure 3, we can see that (i) the P-CUSUM charts almost always outperform the L-CUSUM, W-CUSUM and T-CUSUM(adj) charts, (ii) the W-CUSUM chart outperforms the L-CUSUM chart in most cases considered while the L-CUSUM chart has better performance for detecting downward shifts when the true distribution is $\mathrm{Bin}(20, 0.75)$ and for detecting all shifts when the true distribution is $\mathrm{GP}(5, -0.25)$, (iii) the performance of the T-CUSUM(adj) chart is generally comparable with the winner of the L-CUSUM and W-CUSUM charts, and (iv) for the P-CUSUM chart, it seems that $p$ can be simply chosen 2 for detecting relatively large shifts (e.g., the absolute shift size is larger than 0.5) and 5 for detecting relatively small shifts.

To use the P-CUSUM chart, an IC dataset of size $M$ is required to estimate the IC quantiles $\{q_1, q_2, \ldots, q_{p-1}\}$. Therefore, its performance depends on $M$. To study the impact of $M$ on the performance of the P-CUSUM chart, next we compute its optimal $\mathrm{ARL}_1$ values when $M$ is chosen 200, 500, 1000 or 2000 in the setup of Figure 3, and the results are presented in Figure 4. From the plots, we can see that (i) it is good enough to choose $M = 500$ in most cases except the case with $\mathrm{Bin}(20, 0.75)$ where $M$ should be chosen 1000 or larger for detecting small positive shifts, and (ii) the optimal $\mathrm{ARL}_1$ values are stable when $M \geq 1000$.

In the above OC performance comparison, we select the optimal allowance parameter $k_P$ for the P-CUSUM chart and use the optimal $\mathrm{ARL}_1$ as the performance metric for fair comparison. However, when using the P-CUSUM chart in practice, a specific value of the allowance parameter $k_P$ is usually needed. Next, we study the impact of $k_P$ on the OC performance of the P-CUSUM chart and provide some practical guidelines on choosing this parameter. In this example, the value of $k_P$ can change among 0.001, 0.01, 0.05 and 0.1, the nominal $\mathrm{ARL}_0$ is fixed at 200, $p = 5$ and $M = 500$. The true process distribution is the same as those in the previous example. The $\mathrm{ARL}_1$ values of the P-CUSUM chart are computed based on 10,000 replicated simulations and they are presented in Figure 5. As shown by the plots in the figure, we can see that i) the $\mathrm{ARL}_1$ values are larger when $k_P = 0.1$ or 0.05, compared to their values when $k_P = 0.001$ or 0.01, ii) the $\mathrm{ARL}_1$ values are close to each other in most cases when $k_P = 0.001$ or 0.01, and iii) it seems that the chart performs better for detecting small shifts when $k_P = 0.001$, and it performs better for detecting large

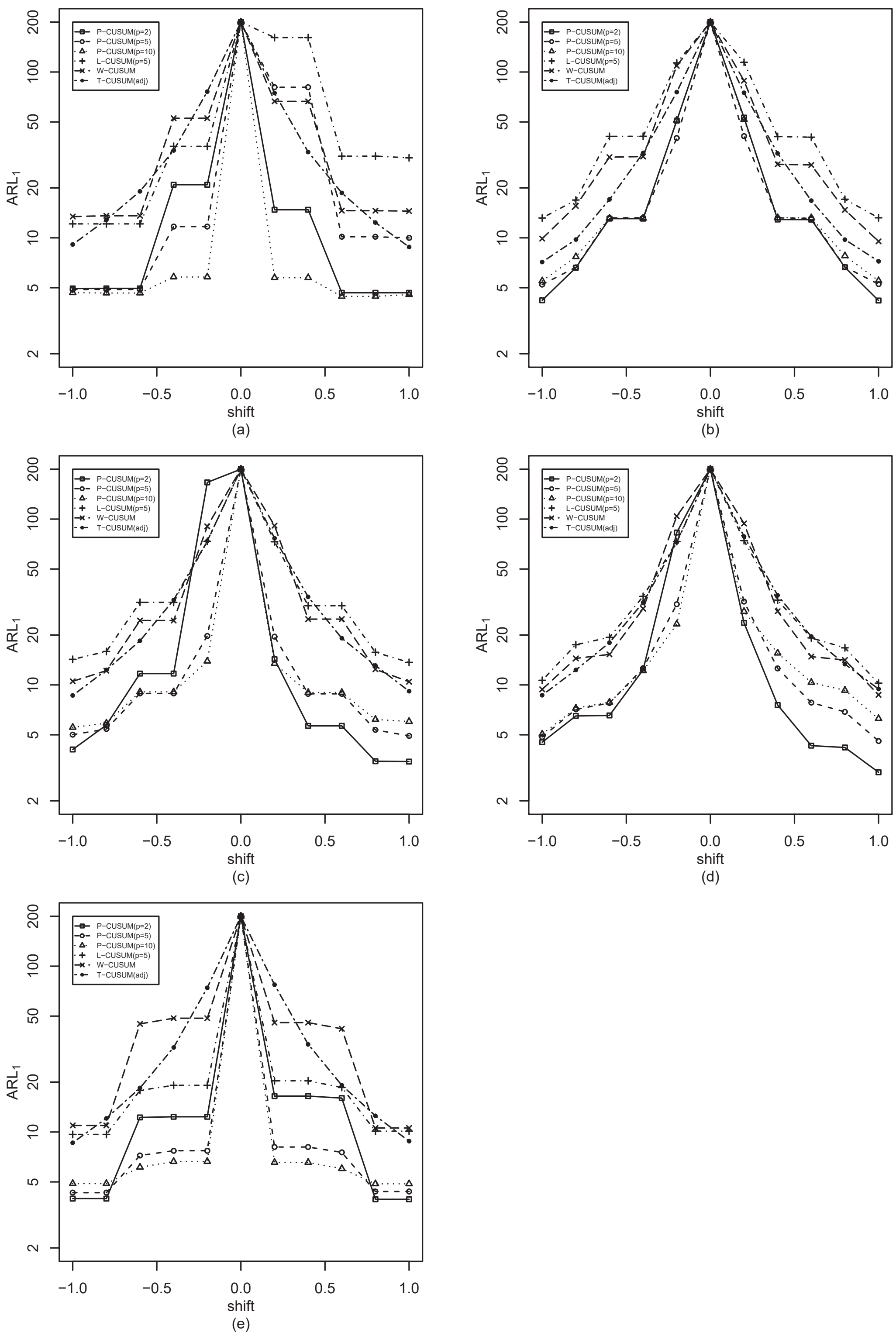

**Figure 3.** The optimal $\text{ARL}_1$ values of the control charts when $\text{ARL}_0 = 200$, $M = 500$, and the actual IC process distribution is the standardized version of $\text{Bin}(20, 0.75)$ (plot (a)), $\text{DU}(10)$ (plot (b)), $\text{NB}(20, 0.75)$ (plot (c)), $\text{GP}(5, 0.25)$ (plot (d)), and $\text{GP}(5, -0.25)$ (plot (e)).

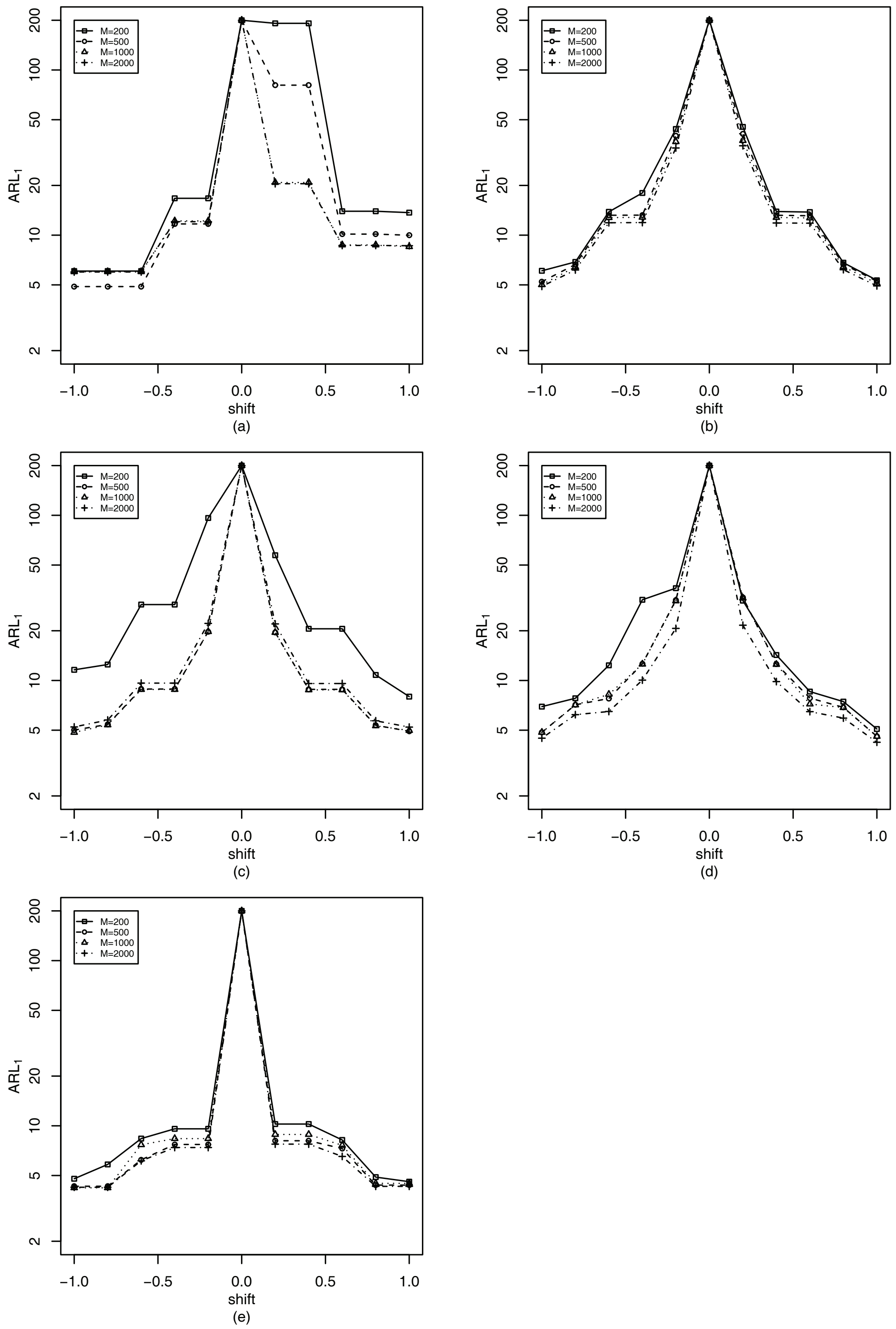

**Figure 4.** The optimal $\text{ARL}_1$ values of the P-CUSUM chart when $M = 200, 500, 1000, 2000$, $p = 5$, and the actual IC process distribution is the standardized version of $\text{Bin}(20, 0.75)$ (plot (a)), $\text{DU}(10)$ (plot (b)), $\text{NB}(20, 0.75)$ (plot (c)), $\text{GP}(5, 0.25)$ (plot (d)), and $\text{GP}(5, -0.25)$ (plot (e)).

shifts when $k_P = 0.01$. Based on this example, we suggest choosing $k_P$ in the range $[0.001, 0.01]$.

## 4 An Application

In this section, we apply the nonparametric control charts discussed in the previous sections to a real-data example about the monthly counts of polio cases in the USA between January 1970 to November 1983. This dataset can be found at *https://datamarket.com/data/set/22u4/monthly-us-polio-cases*, and it is shown in Figure 6. It can be seen from the plot that the monthly counts of polio cases from January 1970 to December 1972 are higher than the monthly counts in later months, and it seems that there is a distributional shift around the beginning of 1973. As a matter of fact, the mean monthly count during January 1970 and December 1972 is 2.36, and it decreases to 1.06 for the time period afterwards, which are denoted by the two horizontal dot-dashed lines in the plot. The explanation for the decrease is that a new polio vaccine started to be available around January 1973.

For the polio data shown in Figure 6, we use the observations during January 1970 and December 1972 as an IC dataset, and the remaining for testing. As discussed in Section 1, the distribution of the IC dataset is significantly different from Poisson. So, we consider the three nonparametric charts P-CUSUM, L-CUSUM and W-CUSUM only in this example. When implementing the related control charts, we choose $\text{ARL}_0 = 200$ in all three charts, $p = 2$ in the P-CUSUM and L-CUSUM charts, the allowance constants for the P-CUSUM and L-CUSUM charts to be 0.01, and the allowance constant for the W-CUSUM chart to be the same as that in Li et al[16]. The main reason for choosing $p = 2$ is that the IC sample size is quite small in this case. When $p = 2$, we only need to estimate one parameter (i.e., $q_1$). The three control charts are shown in Figure 7, where the dashed horizontal lines denote the control limits of the corresponding control charts. From the plots, the P-CUSUM, L-CUSUM, and W-CUSUM charts give signals at the 7th, 47th and 14th Phase II observation times, respectively. So, the P-CUSUM chart is the most effective one in this example, and all three charts confirm that the monthly counts of polio in USA indeed had a mean shift.

## 5 Concluding Remarks

SPC for count data is important because count data is a basic data format in practice. In the literature, most existing control charts for monitoring count data are based on the Poisson or other parametric probability

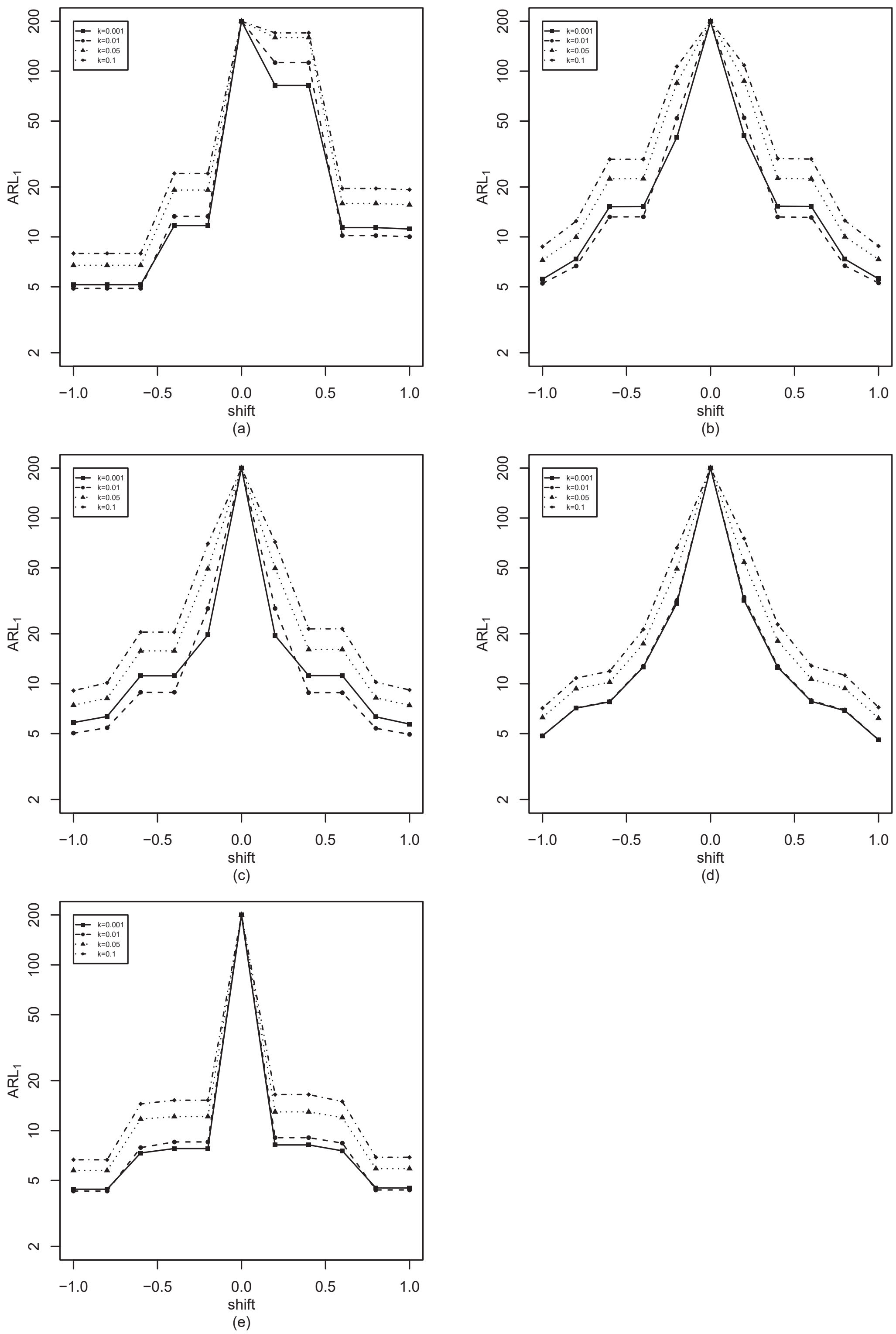

**Figure 5.** The calculated $\mathrm{ARL}_1$ values of the P-CUSUM chart when $k_P = 0.001, 0.01, 0.05, 0.1$, $p = 5$, $M = 500$ and the actual IC process distribution is the standardized version of $\mathrm{Bin}(20, 0.75)$ (plot (a)), $\mathrm{DU}(10)$ (plot (b)), $\mathrm{NB}(20, 0.75)$ (plot (c)), $\mathrm{GP}(5, 0.25)$ (plot (d)), and $\mathrm{GP}(5, -0.25)$ (plot (e)).

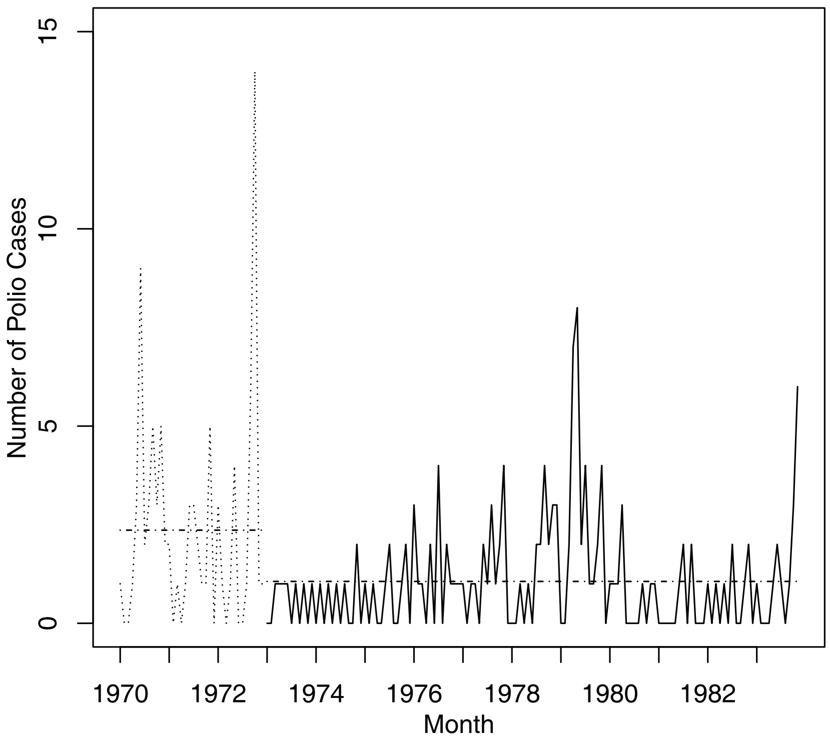

**Figure 6.** Monthly counts of polio cases in the USA between January 1970 to November 1983. The horizontal dot-dashed lines denote the sample means for the observations before and after January 1973.

models. In various applications, however, the true distribution of count data can hardly be described well by such parametric models, partly because the count data are affected by many confounding risk factors and such an impact is difficult to describe by a parametric model. In this paper, we have demonstrated that the existing parametric control charts are often inappropriate to use in cases when the assumed parametric models are invalid, and nonparametric control charts should be considered in such cases. This paper focuses on Phase II process monitoring in univariate cases. We believe that the related conclusions in the paper can be extended to Phase I process monitoring and to multivariate cases as well, which needs to be confirmed in our future research.

**Acknowledgments:** The authors thank the editors and one referee for many helpful comments and suggestions, which improved the quality of the paper greatly. This research is supported in part by an NSF grant.

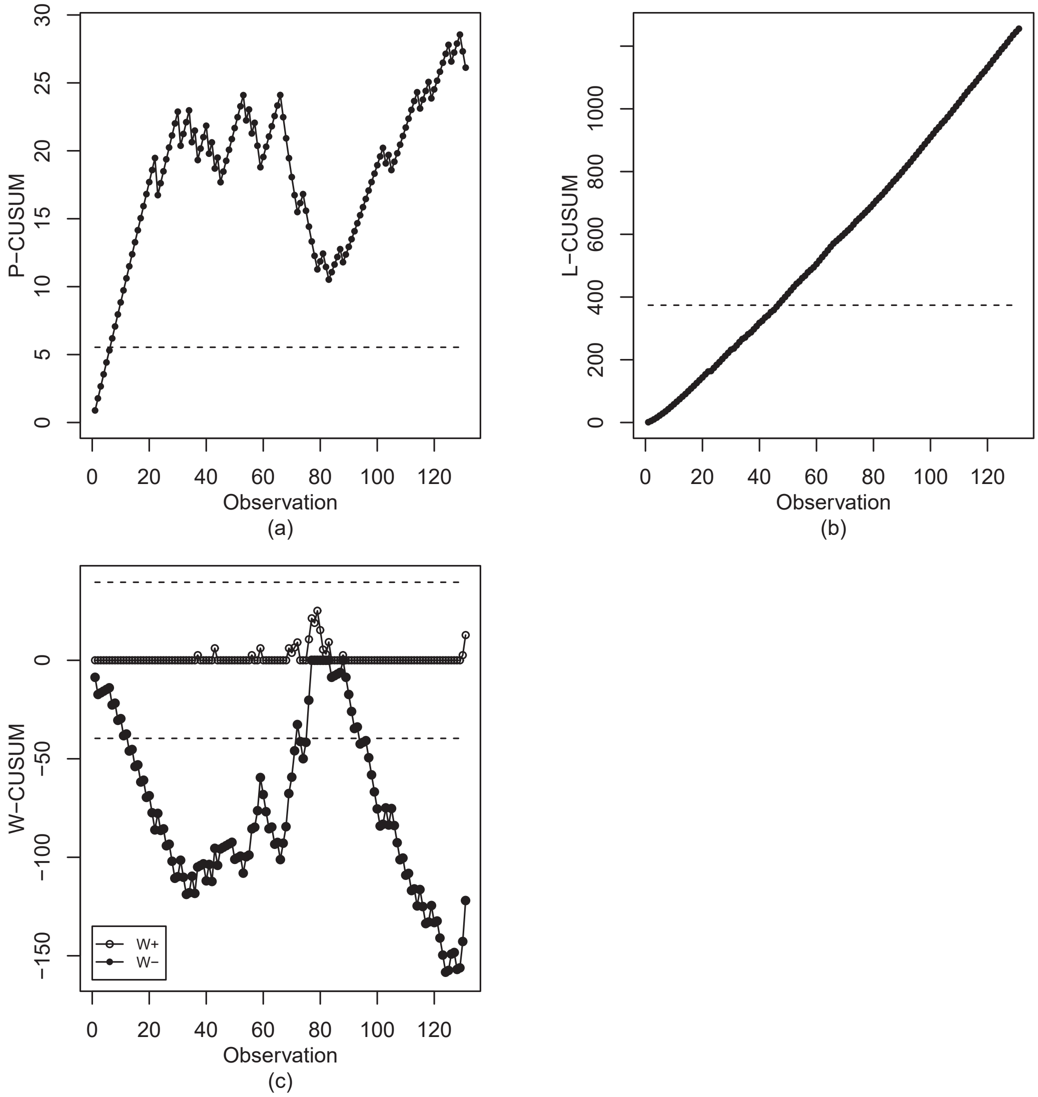

**Figure 7.** The P-CUSUM, L-CUSUM, and W-CUSUM charts for monitoring the monthly counts of polio in USA from January 1973 to November 1983. The horizontal dashed line in each plot denotes the control limit and the related chart.

# References

[1] Gan F. An optimal design of CUSUM control charts for binomial counts. Journal of Applied Statistics. 1993;20(4):445–460.

[2] Wu Z, Jiao J, Liu Y. A binomial CUSUM chart for detecting large shifts in fraction nonconforming. Journal of Applied Statistics. 2008;35(11):1267–1276.

[3] Megahed FM, Kensler JL, Bedair K, Woodall WH. A note on the ARL of two-sided Bernoulli-based CUSUM control charts. Journal of Quality Technology. 2011;43(1):43–49.

[4] Qiu P. Introduction to statistical process control. Chapman & Hall/CRC: Boca Raton, FL; 2014.

[5] Lucas JM. Counted data CUSUM's. Technometrics. 1985;27(2):129–144.

[6] White CH, Keats JB. ARLs and higher-order run-length moments for the Poisson CUSUM. Journal of Quality Technology. 1996;28(3):363–369.

[7] Gan F. Monitoring Poisson observations using modified exponentially weighted moving average control charts. Communications in Statistics-Simulation and Computation. 1990;19(1):103–124.

[8] Borror CM, Champ CW, Rigdon SE. Poisson EWMA control charts. Journal of Quality Technology. 1998;30(4):352–361.

[9] Sheaffer R, Leavenworth R. The negative binomial model for counts in units of varying size. Journal of Quality Technology. 1976;8(3):158–163.

[10] Saghir A, Lin Z. The negative binomial exponentially weighted moving average chart with estimated control limits. Quality and Reliability Engineering International. 2015;31(2):239–250.

[11] Fisher RA, Thornton H, Mackenzie W. The accuracy of the plating method of estimating the density of bacterial populations. Annals of Applied Biology. 1922;9(3-4):325–359.

[12] Chakraborti S, Qiu P, Mukherjee A. Special Issue on Nonparametric Statistical Process Control Charts. Quality and Reliability Engineering International. 2015;31(1):1–151.

[13] Qiu P. Some perspectives on nonparametric statistical process control. Journal of Quality Technology. 2018;50(1):49–65.

[14] Qiu P, Li Z. On nonparametric statistical process control of univariate processes. Technometrics. 2011;53(4):390–405.

[15] Consul PC, Jain GC. A generalization of the Poisson distribution. Technometrics. 1973;15(4):791–799.

[16] Li SY, Tang LC, Ng SH. Nonparametric CUSUM and EWMA control charts for detecting mean shifts. Journal of Quality Technology. 2010;42(2):209–226.

[17] Chatterjee S, Qiu P. Distribution-free cumulative sum control charts using bootstrap-based control limits. The Annals of applied statistics. 2009;3(1):349–369.