# Detecting Change-Points in Epidemic Models

Zhenmin Chen*, Zihao Li and Min Zhou

Department of Mathematics and Statistics, Florida International University, Miami, FL 33199, U.S.A.
Email: chenzh@fiu.edu

**Abstract.** The purpose of this research is to propose a new method for detecting change points with an epidemic alternative (in the form of a step function). There are several parametric approaches and nonparametric approaches in the literature that can be used for detecting change-points in epidemic models. Yan [16] summarized some existing parametric approaches. The approaches summarized in Yao's paper are based on the assumption of known population variances. The proposed test statistic in this research does not depend on the assumption of known population variances. This better fits the real world situation. Monte-Carlo simulation was used to find the critical values of the test. The power study was also based on Monte-Carlo simulation. The simulation result shows that the test statistic proposed in this research provides quite decent power compared with other existing statistical procedures, especially for the case that the step is large and duration of the epidemic is long. In 1997, a likelihood ratio test was proposed by Csorgo and Horvath [24]. Compared with the likelihood ratio test, the method proposed in this paper is easier to use by the statistics users.

**Keywords:** Change-point, epidemic alternative, power comparison, Monte Carlo simulation, unknown variances.

## 1 Introduction

The problem of detecting changes in the characteristics of a random process is referred to as the change-point problem. This problem has become a fast developing research area in statistics mainly due to its important applications and newly developed theoretical and computational methods. Change-point problems occur in a wide variety of fields including detecting shifts in production processes, comparing and matching DNA sequences, examining the impact of social programs, and studying structural shifts in one or more parameters of the models in economics, engineering, political science and other empirical sciences. Many authors have considered the single change-point problems for the univariate and multivariate normal distributions. Page [1, 2] found a procedure for detecting a single change in the distribution of a sequence of independent random variables. The test is based on cumulative sums called cusums. Chernoff and Zacks [3] and Kander and Zacks [4] studied sequences of normal random variables and found a Bayesian test to detect a change in the mean. Sen and Srivastava [5] and Worseley [6] derived tests for the likelihood ratio in the single change situation. Chen and Gupta [7, 8] discussed procedures of detecting change-points using the Schwarz information criterion for the normal distribution. Gombay and Hovarth [9] derived the asymptotic null distribution of the likelihood ratio statistic for the exponential family. The asymptotic non-null distribution for the single change-point was derived by Gombay and Hovarth [10].

Change-points with epidemic alternatives were formulated by Levin and Kline [11] to model the changes over time in the proportion of abortions. In the epidemic change model, the random process is assumed to be stable initially, and then at an unknown time point it will exhibit an abrupt change in the characteristics, which will continue for an unknown duration before stabilizing again to the initial state. The standard normal of a neuron exhibiting a modulated activity during a time period and then reverting to its spontaneous activity is an example of this model, which was described by Commenges, et al [12]. Bromeling and Tsurumi [13] described a number of applications of this model in econometrics. Later, Siegmund [14, 15] and Yao [16] proposed some test statistics and large deviation approximations to the significance levels and powers, for the normal distribution. Furthermore, some semiparametric tests for change-points with epidemic alternatives were recommended by Guan [17].

The problem considered in this study is to detect an epidemic alternative in the mean value of a sequence of normally distributed independent random variables. Yao [16] summarized several test statistics for detecting change-points for epidemic alternatives in the literature. The test statistics are (i) Levin and Kline's statistic; (ii) the semi-likelihood ratio statistic; (iii) the test derived using the idea of the likelihood ratio test; (iv) the score-like statistic, which can be regarded as an analogue of Pettitt's method [18] in testing an epidemic alternative; and (v) the recursive residual statistic, which is inspired by Brown, Durbin and Evans [19]. Approximations to the significance levels have been developed by Hogan and Siegmund [20], Siegmund [14, 21] and Yao [16, 22, 23]. A more general parametric likelihood test was proposed by Csorgo and Horvath [24]. The likelihood ratio test is generally believed to be the most powerful test. The test can be adopted for the epidemic model mentioned in this paper as a special case.

A new approach for detecting epidemic change is proposed in this paper. Critical values of the proposed test statistic are obtained using Monte Carlo simulation. Power study for the test statistic is conducted to compare the performance of the proposed test statistic with the test statistics mentioned in Yao [16]. Compared with the likelihood ratio test, the method proposed in this paper is easier to use for the statistics users.

## 2    Model and Assumptions

Let $X_1, X_2, \cdots, X_n$ be a sequence of normally distributed independent random variables. Consider the following model:

$$X_i = \begin{cases} \mu + e_i & i = 1, \cdots, p, q+1, \cdots, n \\ \mu_a + e_i & i = p+1, \cdots, q \end{cases} \tag{1}$$

for $1 \le p < q < n$. Here $\mu_a = \mu + \delta$; $\mu$ and $\delta$ are unknown parameters; $e_1, \cdots, e_n$ are independent and identically distributed (*i.i.d.*) random variables with $E(e_i) = 0$ and $0 < Var(e_i) = \sigma^2 < \infty$. This model describes the situation that the normal state with the mean value $\mu$ runs up to the $p$th observation, then it changes to the epidemic one with the mean value $\mu_a = \mu + \delta$ at the $(p+1)$th observation and stays at this level through the $q$th observation before the normal state is restored at the $(q+1)$th observation.

The test proposed in this paper is to be used to check if an epidemic change has occurred in an unknown time period. More specifically, the hypotheses can be described as follows:

$H_0 : X_1, X_2, \cdots, X_n$ are normally distributed with mean $\mu$ and standard deviation $\sigma$,

$H_1 : X_1, \cdots, X_p, X_{q+1}, \cdots, X_n$ are normally distributed with mean $\mu$ and standard deviation $\sigma$,

and $X_{p+1}, \cdots, X_q$ are normally distributed with mean $\mu_a = \mu + \delta$ (2)

and standard deviation $\sigma$ for some $1 \le p < q < n$.

Here $\mu$ and $\delta$ play the role of nuisance parameters. To simplify the discussion, only one-sided alternative will be considered, i.e., it is assumed that the sign of $\delta$ is known, say, $\delta > 0$. The method can then be used for a two-sided test without any technical difficulties. The two-sided case will be mentioned and commented later.

The test statistics reviewed in Yao [16] assume that the variance of the underlying distribution is known. This is not acceptable to practitioners. To resolve this problem, point estimation was used to estimate the population variance. The point estimation is, however, based on samples that contain the effect of the alternative hypothesis. This may cause the point estimator of the variance to deviate from the real value significantly. Therefore, a new approach for detecting epidemic change which can bypass point estimation of the population variance is desired, and it will be introduced in the next section.

## 3    New Test Statistic

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent and normally distributed random variables. Among

them, the $k$th random variable has mean $\mu_k$ and variance $\sigma^2$. It is desired to test the hypotheses described in (2). According to model (1),

$$\mu_i = \begin{cases} \mu & i = 1, \cdots, p, q+1, \cdots, n \\ \mu_a = \mu + \delta & i = p+1, \cdots, q. \end{cases} \tag{3}$$

It should be mentioned that $\delta$ is an unknown parameter. In practice, no information about the locations of the start-point and end-point of the epidemic alternative will be given. It means that both $p$ and $q$ are unknown. It is reasonable to assume that $(p, q)$ has equal chance to fall at any possible points $p = 1, \cdots, n-2$ and $q = p+1, \cdots, n-1$. In order to simplify the discussion in the study, only a one-sided alternative is considered, for which it can be assumed that the sign of $\delta$ is known, say $\delta > 0$.

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent random variables, and let $i$ and $j$ be positive integers satisfying $1 \le i < j < n$. Denote

$$I_{i,j} = \left\{ i+1, \cdots, j \right\} \tag{4}$$

and

$$I_{i,j}^c = \left\{ 1, \cdots, i, j+1, \cdots, n \right\} \tag{5}$$

Then

$$I_{i,j} \cup I_{i,j}^c = \left\{ 1, \cdots, n \right\} \tag{6}$$

For any fixed $X_1, X_2, \cdots, X_n$, define

$$U_{i,j} = \sum_{k \in I_{i,j}} X_k \tag{7}$$

$$V_{i,j} = \sum_{k \in I_{i,j}^c} X_k \tag{8}$$

and

$$R_{i,j} = \frac{\left\{ n - (j-i) \right\} U_{i,j} - (j-i) V_{i,j}}{n} \tag{9}$$

Now suppose that $i_0$ and $j_0$ are two positive integers such that

$$R_{i_0, j_0} = \max_{1 \le i < j < n} R_{i,j} \tag{10}$$

Here the selected $(i_0, j_0)$ can actually serve as a point estimate of $(p, q)$. Suppose also that $X_{(1)}^{(i_0, j_0)} < X_{(2)}^{(i_0, j_0)} < \cdots < X_{(n - j_0 + i_0)}^{(i_0, j_0)}$ are the ordered elements of the set

$$\left\{ X_k, k \in I_{i_0, j_0}^c \right\}$$

Let

$$m = \text{int}\left( \frac{n - (j_0 - i_0)}{2} \right) \tag{11}$$

which is the integer part of $\dfrac{n - (j_0 - i_0)}{2}$. The test statistic for detecting the epidemic alternative can then be defined as

$$T = \frac{m R_{i_0, j_0}}{\displaystyle\sum_{k = n - j_0 + i_0 - m + 1}^{n - j_0 + i_0} X_{(k)}^{(i_0, j_0)} - \sum_{k=1}^{m} X_{(k)}^{(i_0, j_0)}} \tag{12}$$

It can be seen that the distribution of the statistic $T$ does not depend on the parameters $\mu$ and $\sigma^2$ under the null hypothesis $H_0$. Therefore, it is unnecessary to assume that the variance $\sigma^2$ is known.

Theoretically speaking, it is almost impossible (except a zero-probability event) that more than one pair of $(i_0, j_0)$ can be found such that the values of $R_{i_o, j_0}$ defined in (10) are the same. This is because

the underlying distribution is assumed to be a normal distribution which is a continuous distribution. However, in practice, one may encounter the above situation because of the fact that people cannot keep an infinite number of decimal places for their data. If that is the case, the authors would suggest that both pairs (or even more pairs although the probability will be much smaller) be used to calculate the value of $T$ in (12). Then compare the maximum value of $T$ with the critical value in Table 1. Similar approach can be used for the point estimates of $p_0$ and $q_0$. When more than one pair of $(i_0, j_0)$ are found such that the values of $R_{i_0, j_0}$ defined in (10) are the same, one might use the smallest value of $i_0$ as the point estimate of $p_0$, and use the largest value of $j_0$ as the point estimate of $q_0$.

**Table 1.** Values of $C_\alpha$ when $\alpha = 0.05$

| $n$ | $C_\alpha$ | $n$ | $C_\alpha$ | $n$ | $C_\alpha$ |
|-----|-----------|-----|-----------|-----|-----------|
| 20 | 5.79 | 47 | 7.58 | 74 | 9.16 |
| 21 | 5.82 | 48 | 7.62 | 75 | 9.22 |
| 22 | 5.89 | 49 | 7.70 | 76 | 9.26 |
| 23 | 5.93 | 50 | 7.78 | 77 | 9.30 |
| 24 | 6.02 | 51 | 7.83 | 78 | 9.35 |
| 25 | 6.10 | 52 | 7.89 | 79 | 9.39 |
| 26 | 6.15 | 53 | 7.95 | 80 | 9.47 |
| 27 | 6.23 | 54 | 8.00 | 81 | 9.51 |
| 28 | 6.33 | 55 | 8.07 | 82 | 9.58 |
| 29 | 6.40 | 56 | 8.15 | 83 | 9.61 |
| 30 | 6.44 | 57 | 8.19 | 84 | 9.69 |
| 31 | 6.52 | 58 | 8.25 | 85 | 9.73 |
| 32 | 6.56 | 59 | 8.32 | 86 | 9.77 |
| 33 | 6.65 | 60 | 8.38 | 87 | 9.80 |
| 34 | 6.73 | 61 | 8.43 | 88 | 9.85 |
| 35 | 6.77 | 62 | 8.47 | 89 | 9.88 |
| 36 | 6.86 | 63 | 8.52 | 90 | 9.93 |
| 37 | 6.91 | 64 | 8.60 | 91 | 9.99 |
| 38 | 7.00 | 65 | 8.68 | 92 | 10.02 |
| 39 | 7.06 | 66 | 8.73 | 93 | 10.07 |
| 40 | 7.12 | 67 | 8.77 | 94 | 10.11 |
| 41 | 7.19 | 68 | 8.82 | 95 | 10.18 |
| 42 | 7.27 | 69 | 8.88 | 96 | 10.22 |
| 43 | 7.33 | 70 | 8.92 | 97 | 10.30 |
| 44 | 7.39 | 71 | 8.98 | 98 | 10.37 |
| 45 | 7.45 | 72 | 9.05 | 99 | 10.41 |
| 46 | 7.49 | 73 | 9.11 | 100 | 10.46 |

In fact, there are several different ways to define the denominator in (12). For instance, the following test statistic was studied:

$$T^* = \frac{R_{i_0, j_0}}{\max_{k \in I^c_{i_0, j_0}} X^{(i_0, j_0)}_{(k)} - \min_{k \in I^c_{i_0, j_0}} X^{(i_0, j_0)}_{(k)}} \tag{13}$$

This test statistic is similar to the test statistic defined in (12) except that it has a different denominator. When the test statistic $T^*$ is used, one may have some concern about the effect of possible outliers. In the next section, statistical simulation will be conducted to compare the power of the test statistics $T$ and $T^*$.

For the two-sided alternative hypothesis in a more general situation, the restriction $\delta > 0$ will be removed. Some modification is needed in (9). Actually, the following can be used to replace the expression of $R_{i,j}$ in (9)

$$R_{i,j} = \frac{\left| \left\{ n - \left( j - i \right) \right\} U_{i,j} - \left( j - i \right) V_{i,j} \right|}{n} \tag{14}$$

The test statistics in (12) and (13) can be defined in the same way. Upper critical values of $T$ can be obtained by Monte Carlo simulation. The upper critical values $C_\alpha$ for different values of $n$ are listed in Table 1 where $\alpha = 0.05$.

## 4    Example and Power Study

To illustrate the use of the statistical test provided in this research, the following flu data set is used.
**Example 1.** The numbers of newly reported flu cases were recorded weekly in Mexico between June 2005 and May 2006. The data are listed in Table 2.

**Table 2.** Mexico flu data

| Date | Number of New Cases | Date | Number of New Cases |
|------|---------------------|------|---------------------|
| 2005-06-05 | 740 | 2005-12-04 | 1407 |
| 2005-06-12 | 710 | 2005-12-11 | 1393 |
| 2005-06-19 | 713 | 2005-12-18 | 1472 |
| 2005-06-26 | 654 | 2005-12-25 | 1540 |
| 2005-07-03 | 727 | 2006-01-01 | 1397 |
| 2005-07-10 | 690 | 2006-01-08 | 1383 |
| 2005-07-17 | 748 | 2006-01-15 | 1475 |
| 2005-07-24 | 664 | 2006-01-22 | 1310 |
| 2005-07-31 | 690 | 2006-01-29 | 1222 |
| 2005-08-07 | 698 | 2006-02-05 | 1169 |
| 2005-08-14 | 641 | 2006-02-12 | 1064 |
| 2005-08-21 | 641 | 2006-02-19 | 1137 |
| 2005-08-28 | 630 | 2006-02-26 | 1131 |
| 2005-09-04 | 811 | 2006-03-05 | 1069 |
| 2005-09-11 | 786 | 2006-03-12 | 1088 |
| 2005-09-18 | 855 | 2006-03-19 | 1038 |
| 2005-09-25 | 845 | 2006-03-26 | 950 |
| 2005-10-02 | 822 | 2006-04-02 | 980 |
| 2005-10-09 | 877 | 2006-04-09 | 880 |
| 2005-10-16 | 948 | 2006-04-16 | 933 |
| 2005-10-23 | 1027 | 2006-04-23 | 844 |
| 2005-10-30 | 1180 | 2006-04-30 | 855 |
| 2005-11-06 | 1294 | 2006-05-07 | 869 |
| 2005-11-13 | 1279 | 2006-05-14 | 909 |
| 2005-11-20 | 1311 | 2006-05-21 | 894 |
| 2005-11-27 | 1304 | 2006-05-28 | 823 |

Data Source: Google Flu Trends (http://www.google.org/flutrends)

Suppose the model proposed in this paper is used to fit the data. It can be found that the value of $R_{i,j}$ reaches its maximum 5894.3462 when $i_0 = 20$ and $j_0 = 42$. Then the value of the test statistic in (12) is 32.2565 which is much higher than the upper critical value 7.89 found in Table 1. Thus one may reject the hypothesis that no epidemic change in means had occurred, and conclude that there was a shift in the means at level of significance 0.05.

Monte-Carlo simulation has been carried out to study the power of the statistical procedures using $T$ and $T^*$. To compare the power of $T$ and $T^*$ with $Z_1, Z_2, Z_3, Z_4$ and $Z_5$ mentioned in Yao [23], one

hundred thousand pseudo-random samples were used for each triplet $(p, q, \delta)$. Here $p, q$ and $\delta$ were defined in (2). The selection of $(p, q, \delta)$ is the same as the one used in Yao [16]. To keep the table size reasonable, only the case of sample size $n = 60$ and significance level $\alpha = 0.05$ is considered. Table 3 lists the result of power comparison for the test statistics $T$ and $T^*$ with $Z_1, Z_2, Z_3, Z_4$ and $Z_5$ described in in Yao [16]. The selected value of $\delta_0$ in the power comparison was $\delta_0 = 0.2$ for both $Z_1$ and $Z_2$. The selected values of $(n_0, n_1)$ in the power comparison for $Z_3$ were $(1, 59)$ and $(6, 54)$. In the power comparison, the maximum likelihood estimator $\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$ to estimate the population variance $\sigma^2$ when the test statistics $Z_1, Z_2, Z_3, Z_4, Z_5$ were used.

**Table 3.** Power comparison; $n = 60$, $\alpha = 0.05$

| $\delta$ | $q - p$ | $T$ | $T^*$ | $Z_1$ $\delta_0 = 0.2$ | $Z_2$ $\delta_0 = 0.2$ | $Z_3$ $n_0 = 1$ $n_1 = 59$ | $Z_3$ $n_0 = 6$ $n_1 = 54$ | $Z_4$ | $Z_5$ $n_0 = 6$ $p + q = n$ |
|---|---|---|---|---|---|---|---|---|---|
| 0.8 |  | 0.14 (0.16) | 0.13 (0.18) | 0.13 (0.12) | 0.12 (0.18) | 0.12 (0.25) | 0.14 (0.21) | 0.12 (0.17) | 0.14 (0.12) |
| 1.2 | 6 (54) | 0.25 (0.25) | 0.21 (0.29) | 0.25 (0.13) | 0.21 (0.19) | 0.28 (0.23) | 0.32 (0.23) | 0.20 (0.18) | 0.33 (0.20) |
| 1.6 |  | 0.41 (0.33) | 0.33 (0.43) | 0.43 (0.09) | 0.33 (0.13) | 0.54 (0.15) | 0.61 (0.16) | 0.29 (0.12) | 0.60 (0.33) |
| 0.8 |  | 0.27 (0.31) | 0.22 (0.33) | 0.29 (0.16) | 0.25 (0.23) | 0.21 (0.23) | 0.27 (0.26) | 0.25 (0.23) | 0.26 (0.19) |
| 1.2 | 10 (50) | 0.53 (0.54) | 0.43 (0.58) | 0.58 (0.15) | 0.50 (0.22) | 0.51 (0.20) | 0.60 (0.23) | 0.47 (0.21) | 0.56 (0.35) |
| 1.6 |  | 0.81 (0.77) | 0.69 (0.81) | 0.84 (0.07) | 0.76 (0.11) | 0.80 (0.10) | 0.87 (0.11) | 0.71 (0.11) | 0.84 (0.55) |
| 0.8 |  | 0.58 (0.61) | 0.51 (0.60) | 0.57 (0.36) | 0.56 (0.46) | 0.38 (0.30) | 0.47 (0.37) | 0.56 (0.46) | 0.40 (0.33) |
| 1.2 | 20 (40) | 0.91 (0.92) | 0.85 (0.91) | 0.88 (0.40) | 0.86 (0.52) | 0.70 (0.30) | 0.80 (0.39) | 0.86 (0.53) | 0.76 (0.62) |
| 1.6 |  | 0.99 (0.99) | 0.98 (0.99) | 0.98 (0.21) | 0.97 (0.30) | 0.90 (0.13) | 0.94 (0.19) | 0.97 (0.31) | 0.95 (0.85) |
| 0.4 |  | 0.25 | 0.24 | 0.2 | 0.24 | 0.14 | 0.18 | 0.24 | 0.14 |
| 0.8 | 30 | 0.69 | 0.64 | 0.56 | 0.61 | 0.38 | 0.48 | 0.62 | 0.41 |
| 1.2 |  | 0.96 | 0.94 | 0.81 | 0.84 | 0.59 | 0.71 | 0.84 | 0.75 |
| 0 |  | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| $C_\alpha$ |  | 8.38 | 3.21 | 9.32 | 10.30 | 3.60 | 3.40 | 11.65 | 3.24 |

The row at the bottom of Table 3 provides the cutoff point $C_\alpha$ for each test. The simulated values for the significance level are also reported, from which one can find that approximately 5% of the sample points fell in the rejection region for each test when there was no epidemic change, i.e. under the null hypothesis. Here $\delta = 0$ means the situation is under the null hypothesis $H_0$. The columns of Table 3 are organized as follows:

(i). Column 1 contains epidemic change values $\delta > 0$,
(ii). Column 2 gives durations of the epidemic state $q - p$,
(iii). Columns 3 and 4 list the power of the proposed test statistics $T$ and $T^*$ for certain different situations,
(iv). Columns 5 through 10 list the power of those test statistics reviewed in Section 1 for several different situations.

Columns 7 and 8 in Table 3 involve the test statistic $Z_3$ with different choices of $n_0$ and $n_1$. Since the power of $Z_5$ depends on $(p, q)$ not only through $q - p$, the results were reported merely for a special case that the epidemic state appears in the middle, i.e., $p + q = n$.

The powers of $Z_1, Z_2, Z_3, Z_4$ and $Z_5$ are not symmetric with respect to the duration of the epidemic state, i.e. an epidemic alternative with duration $q - p$ may have different power from an alternative with duration $n - (q - p)$. Thus, the figures within parentheses in the columns of $Z_1, Z_2, Z_3, Z_4$ and $Z_5$ exhibit the powers for the duration values greater than 30. Roughly speaking, for a given epidemic

change value $\delta$, the powers of $Z_1, Z_2, Z_3, Z_4$ and $Z_5$ corresponding to certain durations $q - p < n/2$ are considerably greater than those corresponding to durations $n - (q - p)$ except for the special cases with $\delta$ very small and $q - p$ near zero. The reason of the above fact is that the tests $Z_1, Z_2, Z_3, Z_4$ and $Z_5$ have used the point estimator $\widehat{\sigma^2}$ as the population variance $\sigma^2$ during standardization of the random samples. The point estimator $\widehat{\sigma^2}$ is based on the sample data contaminated by the epidemic alternative. This will, of course, affect the accuracy of the estimation, and affect the power of the tests. Therefore, for a given epidemic change value $\delta$, the powers of the tests $Z_1, Z_2, Z_3, Z_4$ and $Z_5$ would be affected. For the same reason, the power of $Z_1, Z_2, Z_3, Z_4$ and $Z_5$ would also be subjected to the effect of epidemic change value $\delta$. It is obvious that, for a given $q - p$ and $n - (q - p)$ in Table 3, the difference between the power values of $Z_1, Z_2, Z_3, Z_4$ and $Z_5$ outside the parentheses and those between the parentheses increase as the epidemic change value $\delta$ increases, say from 0.8 to 1.2 and then 1.6. Furthermore, an increase in the epidemic change value $\delta$ would increase the powers of $Z_1, Z_2, Z_3, Z_4$ and $Z_5$ for a certain duration $q - p$ less than $n/2$; whereas for a certain epidemic duration greater than $n/2$, i.e. a certain value between the parentheses in column 2 such as 40, 50 or 54, the simulated results in columns 4, 5, 6, 7 and 8 show that the power values of $Z_1, Z_2, Z_3$ and $Z_4$ between the parentheses tend to decrease when the epidemic change value $\delta$ increases. When the epidemic change value $\delta$ increases to 1.6 or more, for some duration close to n, the powers of $Z_1, Z_2, Z_3$ and $Z_4$ would become even tiny.

One can intuitively imagine that, when the epidemic state has longer duration or when the epidemic change is stronger, it should be easier for statistical procedures to detect the existence of the epidemic alternative. In other words, the power of a statistical test should be higher when the duration of the epidemic state is longer or when the value $\delta$ is bigger. However, Table 3 shows that the performance of $Z_1, Z_2, Z_3, Z_4$ and $Z_5$ is not as good as expected.

To apply those statistical test procedures described in Section 1, some additional conditions or assumptions are needed. These conditions or assumptions will restrict the usage of those procedures. Some cases have been described by Yao [16]. For instance, there is an additional parameter $\delta_0$ in the expressions of test statistics $Z_1$ and $Z_2$. Here $\delta_0$ is taken as the smallest increment $\delta$ in means which is considered important to detect. Presumably an increase in the value of $\delta_0$ would increase the power of $Z_1$ when $q - p$ is near zero, and also that of $Z_2$ when $q - p$ is near zero; but it would decrease the power of $Z_1$ when $q - p$ is large, and also that of $Z_2$ when $q - p$ is near $n/2$. Another example is that before using the test statistic $Z_3$, parameters $n_0$ and $n_1$ should be preset such that $n_0 \le j - i \le n_1$. From Table 3, it can be seen that the test statistic $Z_3$ with $n_0 = 6$ and $n_1 = 54$ has greater power than $Z_3$ with $n_0 = 1$ and $n_1 = 59$. The case with the recursive residual statistic $Z_5$ is more complicated. Some other simulation shows that, for fixed values of $\delta$ and $q - p$, the power will increase as the value of $q + p$ increases.

As mentioned above, the power of the test procedures reviewed in Yao [16] drops from the case $q - p = 6$ to the case $q - p = 54$ for a fixed value of $\delta$. This does not fit the real world situation. It can be seen from the third and fourth columns of Table 3 that for a given value of epidemic change $\delta$ (e.g. 0.8, 1.2, or 1.6), the powers of test statistics $T$ and $T^*$ with some epidemic duration $q - p < n/2$ (the value outside the parentheses in column 3) are closer to the powers of the tests with corresponding epidemic duration $n - (q - p) > n/2$ (the values inside the parentheses). In other words, the power of $T$ or $T^*$ is more symmetric with respect to the duration of the epidemic state. It means that an epidemic alternative with duration $q - p$ possesses a similar power to an alternative with duration $n - (q - p)$. On the other hand, for a fixed duration $q - p$, an increase of the value of epidemic change $\delta$ increases the powers of tests $T$ and $T^*$. Additionally, when the epidemic duration $q - p$ is closer

to $n/2$, the power of tests $T$ and $T^*$ would become greater. These better fit the real application situation.

Comparing the proposed test statistics $T$ and $T^*$ with $Z_1, Z_2, Z_3, Z_4$ and $Z_5$ on their performance, it can be found $T$ and $T^*$ provide quite decent power in many cases. Especially, when the epidemic change $\delta$ is not very small, say 1.2, and the epidemic duration $q-p$ is around or greater than $n/2$, the performance of $T$ and $T^*$ is better than the rest test statistics.

The following example shows how the method presented in this paper and the existing methods described above perform when they are used to find point estimates of $p$ and $q$ .

**Example 2.** A computer simulated random sample of size 30 was generated from a standard normal distribution. To simulate the epidemic shift on the means, 13 observations starting from observation 11 were artificially raised by $\delta = 1$. The following is the data set after $\delta = 1$ was added to observations 11 to 23:

−1.37, 0.49, −1.31, 0.88, −0.15, 0.26, 1.12, −0.51, 0.25, −1.49, 0.27, 2.04, 0.24, 1.99, 0.13, 1.03, −1.20, 0.89, 2.83, 0.32, 0.74, 1.66, 1.42, −1.33, 0.66, 0.62, −0.05, −1.35, −0.69, –0.04.

Table 4 shows the point estimates of $p$ and $q$ when the method described in this paper and the methods based on $Z_1, Z_2, Z_3, Z_4, Z_5$ are used. It can be found that the estimates are quite close to each other except the case when $Z_1$ is used. It should be mentioned that the value of $\delta_0$ is assumed to be known, i.e., $\delta_0 = 1$ .

**Table 4.** Point estimates of $p$ and $q$ in Example 2

| Method | $\hat{p}$ | $\hat{q}$ | Max Value |
|--------|-----------|-----------|-----------|
| $R_{i,j}$ | 11 | 23 | $\max R_{i,j} = 8.718$ |
| $Z_1$ | 17 | 23 | $\max Z_1 = 3.174$ |
| $Z_2$ | 11 | 23 | $\max Z_2 = 5.118$ |
| $Z_3$ | 11 | 23 | $\max Z_3 = 3.249$ |
| $Z_4$ | 11 | 23 | $\max Z_4 = 8.718$ |
| $Z_5$ | 10 | 23 | $\max Z_5 = 2.730$ |

## 4    Conclusion and Discussion

A new method for detecting change-points with epidemic alternative is studied in this research. Since this method is independent of the population mean $\mu$ and the population standard deviation $\sigma$, it is unnecessary to assume that the variance is known. The method can then be used to obtain insight into more general problems. Moreover, by numerical comparison with other five test statistics summarized in Yao [16], the statistical tests presented in this study provide quite decent power, especially for the case that the step is large and duration of the epidemic is long. Comparing test statistics $T$ and $T^*$, the power of $T$ is slightly better than $T^*$. Also it is possible that the test statistic $T^*$ could be affected by outliers. Therefore, the proposed test $T$ should be recommended to practitioners for detecting epidemic change in means.

As mentioned in Section 1, a more general parametric likelihood test was proposed in Csorgo and Horvath [24]. For the model discussed in this paper, the likelihood ratio test is equivalent to

$$T_{LR} = \max_{1 \le p < q \le n} \frac{1}{\widehat{\sigma}_{p,q}} \left( \frac{n}{(q-p)(n-q+p)} \right)^{1/2} \left( S_q - S_p - \frac{q-p}{n} S_n \right)$$

where

$$\left(\widehat{\sigma_{p,q}}\right)^2 = \frac{1}{n}\left(\sum_{i \in I_{p,q}^c}\left(X_i - \widehat{\mu_{p,q}}\right)^2 + \sum_{i \in I_{p,q}}\left(X_i - \widehat{\eta_{p,q}}\right)^2\right),$$

$$\widehat{\mu_{p,q}} = \frac{1}{n - q + p}\sum_{i \in I_{p,q}^c} X_i ,$$

$$\widehat{\eta_{p,q}} = \frac{1}{q - p}\sum_{i \in I_{pq}} X_i .$$

It is generally believed that the likelihood ratio tests usually have competitive power. The problem for the likelihood ratio test is that the computation is relatively lengthier. Statistics users may use the method presented in this paper if they need an easy test for detecting possible mean shift in the epidemic models.

The method used in this paper can be used for some models similar to the epidemic model. For examples, the models with $X_i$ being defined as

$$X_i = \begin{cases} \mu + \varepsilon_i & i = 1, 2, \cdots, p \\ \mu_a + \varepsilon_i & i = p+1, p+2, \cdots, n \end{cases}$$

or

$$X_i = \begin{cases} \mu_a + \varepsilon_i & i = 1, 2, \cdots, p \\ \mu + \varepsilon_i & i = p+1, p+2, \cdots, n. \end{cases}$$

These two models will not be appropriate models to fit epidemic disease data. However, they might be used in some other application areas. The authors hope the current paper will attract researchers to make contribution to those areas.

## References

1. Page ES. Continuous inspection schemes. Biometrika. 1954; 41: 100-155.
2. Page ES. A test for a change in a parameter occurring at an unknown point. Biometrika. 1955; 42: 523-527.
3. Chernoff H, Zacks S. Estimating the current mean of a normal distribution which is subjected to changes over time. Annals of Mathematical Statistics. 1964; 35: 999-1018.
4. Kander Z, Zacks S. Test procedures for possible changes in parameters of statistical distributions occurring at unknown time points. Annals of Mathematical Statistics. 1966; 37: 1196-1210.
5. Sen A, Srivastava MS. Some one-sided tests for change in level. Technometrics. 1975; 17: 61-64.
6. Worseley KJ. Confidence regions and tests for a change-point in a sequence of exponential family random variables. Biometrika. 1986; 73: 91-105.
7. Chen J, Gupta AK. Likelihood procedure for testing change-points hypothesis for multivariate Gaussian model. Random Operators and Stochastic Equations. 1995; 3: 235-244.
8. Chen J, Gupta AK. Testing and locating variance change-points with applications to stock prices. Journal of the American Statistical Association. 1997; 92: 739-747.
9. Gombay E, Hovarth L. An application of the maximum likelihood test to the change-point problem. Stochastic processes and applications. 1994; 50: 161-171.
10. Gombay E, Hovarth L. Approximations for the time of change and the power function in change-point models. Journal of Statistical Planning and Inference. 1996; 52: 43-66.
11. Levin B, Kline J. The cusum test of homogeneity with an application in spontaneous abortion epidemiology. Statistics in Medicine. 1985; 4: 469-488.
12. Commenges D, Seal J, Pinatal F. Inference about a change-point in experimental neurophysiology. Mathematical Biosciences. 1986; 80: 81-108.
13. Broemeling LD, Tsurumi H. Econometrics and Structural Change. New York: Marcel Dekker; 1987.

14. Siegmund D. Approximate tail probabilities for the maxima of some random fields. Annals of Probability. 1988; 16: 67-80.

15. Siegmund D. Confidence sets in change-point problems. International Statistical Review. 1988; 56: 31-48.

16. Yao Q. Tests for change-points with epidemic alternatives. Biometrika. 1993; 80: 179-191.

17. Guan Z. Semiparametric tests for change-points with epidemic alternatives. Journal of Statistical Planning and Inference. 2007; 137: 1748-1764.

18. Pettitt AN. A simple cumulative sum type statistic for the change-point problem with zero-one observations. Biometrika. 1980; 67: 79-84.

19. Brown RT, Durbin J, Evans JM. Techniques for testing the constancy of regression relations over time. Journal of the Royal Statistical Society, Series B (Methodological). 1975; 37: 149-192.

20. Hogan ML, Siegmund D. Large deviations for the maxima of some random fields. Advances in Applied Mathematics. 1986; 7: 2-22.

21. Siegmund D. Boundary crossing probabilities and statistical applications. Annals of Statistics. 1986; 14: 361-404.

22. Yao Q. Large deviations for boundary crossing probabilities of some random fields. Journal Mathematical Research & Exposition. 1989; 9: 181-192.

23. Yao Q. Boundary crossing probabilities of some random fields related to likelihood ratio tests for epidemic alternatives. Journal of Applied Probabilities. 1993; 30: 52-65.

24. Csorgo M, Horvath L. Limit Theorems in Change-Point Analysis. Chichester: John Wiley & Sons; 1997.