3-29-2013

Power Study on Testing Epidemic Alternatives

Zihao Li
Florida International University, zli006@fiu.edu

DOI: 10.25148/etd.FI13041502

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

POWER STUDY ON TESTING EPIDEMIC ALTERNATIVES

A thesis submitted in partial fulfillment of

the requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

by

Zihao Li

2013

To: Dean Kenneth Furton
College of Arts and Sciences

This thesis, written by Zihao Li, and entitled Power Study on Testing Epidemic Alternatives, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this thesis and recommend that it be approved.

_____
Gauri L. Ghai

_____
Florence George

_____
Zhenmin Chen, Major Professor

Date of Defense: March 29, 2013

The thesis of Zihao Li is approved.

_____
Dean Kenneth Furton
College of Arts and Sciences

_____
Dean Lakshmi N. Reddi
University Graduate School

Florida International University, 2013

ACKNOWLEDGMENTS

I wish to thank the members of my committee for their support, patience, and instructions. Especially thank my major professor Dr. Zhenmin Chen who guided me through the critical theories and pointed out the direction toward the qualitative methods. Mr. Tieyong Hu and Mr. Zeyi Wang, two recent graduates from statistics program, were particularly helpful in software coding discussion. Finally, I would like to thank my family taking care of my daughter and new born son during my hard work in the thesis study.

My coursework throughout the curriculum was very helpful and thoughtful, providing me with tools to explore the ideas in this thesis.

ABSTRACT OF THE THESIS

POWER STUDY ON TESTING EPIDEMIC ALTERNATIVES

by

Zihao Li

Florida International University, 2013

Miami, Florida

Professor Zhenmin Chen, Major Professor

Detecting change points in epidemic models has been studied by many scholars. Yao (1993) summarized five existing test statistics in the literature. Out of those test statistics, it was observed that the likelihood ratio statistic showed its standout power. However, all of the existing test statistics are based on an assumption that population variance is known, which is an unrealistic assumption in practice. To avoid assuming known population variance, a new test statistic for detecting epidemic models is studied in this thesis. The new test statistic is a parameter-free test statistic which is more powerful compared to the existing test statistics. Different sample sizes and lengths of epidemic durations are used for the power comparison purpose. Monte Carlo simulation is used to find the critical values of the new test statistic and to perform the power comparison. Based on the Monte Carlo simulation result, it can be concluded that the sample size and the length of the duration have some effect on the power of the tests. It can also be observed that the new test statistic studied in this thesis has higher power than the existing test statistics do in all of cases.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE PAGE

1. Background

Detecting change points in epidemic models has become an interesting topic in recent years. Change-points with epidemic alternatives were originally formulated by Levin and Kline [2] to model the changes over time in the proportion of abortions. Yao [1] summarized several statistical tests for detecting epidemic changes. The shortcoming of the methods summarized in [1] is that all the methods assume the population variance to be known, which is an unrealistic assumption in practice. To solve this problem, a new statistical test will be proposed in this research. The proposed test procedure does not depend on unknown population variance.

2. Introduction

The epidemic change model used in this research is described as follows.

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

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

for $1 \le p < q < n$. Here μ and δ are unknown parameters. $e_1, \cdots, e_n$ are independent and identically distributed 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 μ runs up to the p th observation. It changes to the epidemic one with the mean value $\mu_a = \mu + \delta$ at the $(p+1)$th observation and stays with this level through the q th observation before the

original normal state is restored at the $(q+1)$th observation if q is less than the total observation size.

The thesis is organized as follows. Section 3 reviews several existing test statistics and indicates the problems we concern. Section 4 studies a new test statistic based on the test statistic (Z_3). Section 5 extends the study to power comparison across the two test statistics. Section 6 draws conclusion that the new test statistic supreme over the existing test statistics.

3. Existing Test Statistics in the literature

The purpose of this research is to detect epidemic changes. It is desired to check if an epidemic change has occurred in an unknown time period. It is assumed that all observations are independent. The hypotheses can be described as follows:

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

H_1: $X_1, \cdots, X_p, X_{q+1} \cdots, X_n$ are normally distributed with mean μ and standard deviation σ, while $X_{p+1}, \cdots, X_q$ are normally distributed with mean $\mu_a = \mu + \delta$ and standard deviation σ.

Yao[1] summarized five existing test statistics in the literature.

3.1 Levin & Kline's Statistic

In the test procedure proposed by Levin& Kline [2], the mean is estimated by MLE namely $\mu = S_m / m$. Population variance is unknown and needs to be estimated. The test statistic is formulated as

$$Z_1 \equiv \max_{1 \le i < j \le m} \left\{ S_j - S_i - \frac{j-i}{m} S_m - (j-i)\frac{1}{2}\delta_0 \right\} \tag{1}$$

The unknown parameter, δ , is estimated using the smallest increment in means, say δ_0. Here m is the observed sample size and S_m is the total summation of observed sample data. The estimates of three parameters make applications difficult.

3.2 The semi-likelihood ratio statistic

Siegmund [3] considered the likelihood ratio when $\delta = \delta_0$ and μ is unknown.

The test statistic is formulated as

$$Z_2 \equiv \max_{1 \le i < j \le m} \left\{ S_j - S_i - \frac{j-i}{m} S_m - \frac{1}{2}(j-i)\left(1 - \frac{j-i}{m}\right)\delta_0 \right\}. \tag{2}$$

3.3 The likelihood ratio statistic

For the case that μ and δ are both unknown with δ >0, the square root of a slightly generalized log likelihood ratio statistic is calculated as

$$Z_3 \equiv \max_{n_0 \le j-i \le n} \left(S_j - S_i - \frac{j-i}{n} S_n \right) \Bigg/ \left\{ (j-i)\left(1 - \frac{j-i}{n}\right) \right\}^{\frac{1}{2}}. \tag{3}$$

The test statistic sheds some light on our interest in parameter-free test. To show the distribution of Z_3 does not depend on μ , note that

$$S_j - S_i = \sum_{k=1}^{j} X_k - \sum_{k=1}^{i} X_k = \sum_{k=i+1}^{j} X_k \ .$$

Let $t_1, t_2, \cdots, t_n$ be iid random variables from $N(0,1)$ distribution.

$$X_k = \mu + t_k \sigma \ \ (k = 1, 2, \cdots, n)$$

$$S_j - S_i - \frac{j-i}{n} S_n = (j-i)\mu + \sigma \sum_{k=i+1}^{j} t_k - \frac{j-i}{n}\left(n\mu + \sigma \sum_{k=1}^{n} t_k \right)$$

$$= (j-i)\mu - (j-i)\mu + \sigma \sum_{k=i+1}^{j} t_k - \frac{j-i}{n} \sigma \sum_{k=1}^{n} t_k$$

$$= \sigma \left[\sum_{k=i+1}^{j} t_k \left(1 - \frac{j-i}{n} \right) - \frac{j-i}{n} \left(\sum_{k=1}^{i} t_k + \sum_{k=j+1}^{n} t_k \right) \right]$$

Here n is the observed sample size and S_n is the total summation of observed sample data. Clearly, the parameter μ is eliminated inherently in the test.

3.4 The score-like statistic

In the general model, δ is unknown. By setting δ =0, the invariant efficient score statistic becomes

$$Z_4 \equiv \max_{1 \le i < j \le n} \left(S_j - S_i - \frac{j-i}{n} S_n \right) \qquad (4)$$

By setting $j \equiv n$, It is a special form of Pettitt's method testing one change-point hypotheses (Pettitt, 1980[4]; James et al., 1987[5]).

3.5 The recursive residual statistic

Brown et al. [6] developed the so-called recursive statistic for testing a change-point in a linear model. We considered the standard residual of y_k from the mean value

$$\overline{y}_{k-1} \equiv \left(y_1 + \cdots + y_{k-1}\right) / \left(k-1\right).$$

Denote

$$w_k \equiv \left\{(k-1)/k\right\}^{\frac{1}{2}} \left(y_k - y_{k-1}\right) \quad (k = 2, \cdots, m).$$

Define cumulative sum as

$$S_k \equiv w_2 + \cdots + w_k .$$

The test statistic is formulated as

$$Z_5 \equiv \max_{n_0 \le j-i \le n} \left(S_j - S_i\right) / \left(j-i\right)^{\frac{1}{2}}. \tag{5}$$

3.6 Comparison of the existing test statistics

The performance of Z_2 shows a little more robust over Z_1 if the duration of the epidemic state $q-p$ is close to m. If δ_0=0, Z_1 becomes Z_4. Actually, Z_4 is a special

case of the Levin & Kline statistic when people try to detect the change in means even though the increment is very small.

It seems simpler to study Z_4 than to study Z_1, Z_2, and Z_3. Yao (1993) specified all other conditions being the same, the Z_4 shown the favorite outcome. However, we cannot ignore the unknown parameter δ_0 in test statistics Z_1, Z_2, and Z_4. When δ_0 is very small, Z_2 and Z_4 perform similarly. Yao (1993) conducted numerical comparisons across Z_1, Z_2, Z_3, Z_4, and Z_5. Yao experimented 10,000 repetitions Monte Carlo to layout the results. The power of Z_1 and Z_5 is not symmetric during the epidemic status. Then the two-side epidemic detections by Z_1 and Z_5 have to be complicate by comparing data with either side significance level. Again, for δ_0=0.2 the semi-likelihood statistic Z_2 and the score-like statistic Z_4 perform roughly the same. Comparing with Z_2 and Z_4, the Levin & Kline's Z_1 has higher power when $q-p$ is near 0. The power becomes lower when $q-p$ tends to be larger. Only the extreme case makes Z_1 functional. Furthermore, the likelihood ratio statistic Z_3 has greater power than Z_1 for small values of $q-p$. Levin & Kline's statistic seems not our preferable in any case. The case with recursive residual statistic Z_5 is more complicated. By fixing δ and the value of $q-p$, the power will increase along with the increase of value of $p+q$. The unstable performance of Z_5 brought its difficulty to apply the statistic in practice. Yao (1993) drew a tentative conclusion that the recursive residual statistic is not demonstrably

inferior to the others. Through the rigor required conditions among Z_2, Z_4, and Z_5, Z_3 stands alone out of the list. My thesis attempts to build up the new argument over this likelihood ratio statistic.

The test statistics reviewed in Yao [1] assume that the variance of the underlying distribution is known. Practically, this assumption is unrealistic when working with actual data. When the population variance is unknown, it is suggested in Yao's paper that the unknown variance be replaced by its point estimate. However, the point estimation is calculated by using samples that contain the effect of H_1. It may cause the point estimator to deviate from the true value significantly. The deviation can be seen from the following proposition.

Proposition 1. Let $X_1, X_2, \cdots, X_n$ be an *iid* sample from a population distribution with mean μ and variance σ^2. For $1 \le i < j < n,$ define $Y_k = X_k$ $(k = 1, 2, \cdots, i, j+1, \cdots, n)$ and $Y_k = X_k + c$ $(k = i+1, i+2, \cdots, j).$ Then

$$E(S_Y^2) = E(S_X^2) + \frac{\big(n-(j-i)\big)\big(j-i\big)}{n(n-1)}c^2.$$

Here $S_X^2 = \dfrac{\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2}{n-1}$ and $S_Y^2 = \dfrac{\sum_{k=1}^{n}\left(Y_k - \overline{Y}\right)^2}{n-1}$.

Proof. Denote

$$K = \{i+1, i+2, \cdots, j-1, j\} \qquad \text{and} \quad K^c = \{1, 2, \cdots, i, j+1, \cdots, n\}$$

Then

$$\overline{Y} = \overline{X} + \frac{(j-i)c}{n}.$$

$$S_Y^2 = \frac{\sum_{k\in K}\left((X_k + c) - \left(\overline{X} + \frac{(j-i)c}{n}\right)\right)^2 + \sum_{k\in K^c}\left(X_k - \left(\overline{X} + \frac{(j-i)c}{n}\right)\right)^2}{n-1}$$

$$\sum_{k\in K}\left((X_k + c) - \left(\overline{X} + \frac{(j-i)c}{n}\right)\right)^2 + \sum_{k\in K^c}\left(X_k - \left(\overline{X} + \frac{(j-i)c}{n}\right)\right)^2$$

$$= \sum_{k\in K}\left(\left(X_k - \overline{X}\right) + \left(\frac{n-(j-i)}{n}c\right)\right)^2 + \sum_{k\in K^c}\left(\left(X_k - \overline{X}\right) - \frac{(j-i)}{n}c\right)^2$$

$$= \sum_{k\in K}\left(X_k - \overline{X}\right)^2 + \sum_{k\in K^c}\left(X_k - \overline{X}\right)^2 + 2\sum_{k\in K}\left(X_k - \overline{X}\right)\left(\frac{n-(j-i)}{n}c\right)$$

$$-2\sum_{k\in K^c}\left(X_k - \overline{X}\right)\left(\frac{(j-i)}{n}c\right) + \sum_{k\in K}\left(\frac{n-(j-i)}{n}c\right)^2 + \sum_{k\in K^c}\left(\frac{(j-i)}{n}c\right)^2$$

$$= \sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2 + 2c\sum_{k\in K}\left(X_k - \overline{X}\right) - 2\sum_{k=1}^{n}\left(\left(X_k - \overline{X}\right)\frac{(j-i)}{n}c\right)$$

$$+(j-i)\left(\frac{n-(j-i)}{n}c\right)^2 + \left(n-(j-i)\right)\left(\frac{(j-i)}{n}c\right)^2$$

$$= \sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2 + 2c\sum_{k\in K}\left(X_k - \overline{X}\right) + (j-i)\left(\frac{n-(j-i)}{n}c\right)^2$$

$$+\left(n-(j-i)\right)\left(\frac{(j-i)}{n}c\right)^2$$

$$S_Y^2 = \frac{\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2 + 2c\sum_{k\in K}\left(X_k - \overline{X}\right) + (j-i)\left(\frac{n-(j-i)}{n}c\right)^2 + \left(n-(j-i)\right)\left(\frac{(j-i)}{n}c\right)^2}{n-1}$$

$$= S_X^2 + \frac{2c\sum_{k\in K}\left(X_k - \overline{X}\right) + (j-i)\left(\frac{n-(j-i)}{n}c\right)^2 + \left(n-(j-i)\right)\left(\frac{(j-i)}{n}c\right)^2}{n-1}$$

$$E\left(S_Y^2\right) = E(S_X^2) + \left(\frac{j-i}{n-1}\right)\left(\frac{n-(j-i)}{n}c\right)^2 + \left(\frac{n-(j-i)}{n-1}\right)\left(\frac{(j-i)}{n}c\right)^2$$

$$= E(S_X^2) + \frac{(j-i)}{n-1}\left(c^2 - 2c^2\frac{(j-i)}{n} + \left(\frac{(j-i)}{n}c\right)^2\right) + \frac{n-(j-i)}{n-1}\left(\frac{(j-i)}{n}c\right)^2$$

$$= E(S_X^2) + \frac{(j-i)}{n-1}c^2 - 2c^2\frac{(j-i)^2}{n(n-1)} + \frac{(j-i)}{n-1}\left(\frac{(j-i)}{n}c\right)^2 + \frac{n-(j-i)}{n-1}\left(\frac{(j-i)}{n}c\right)^2$$

$$= E(S_X^2) + \frac{(j-i)}{n-1}c^2 - 2c^2\frac{(j-i)^2}{n(n-1)} + \frac{n}{n-1}\left(\frac{(j-i)}{n}c\right)^2$$

$$= E(S_X^2) + \frac{n(j-i) - 2(j-i)^2 + (j-i)^2}{n(n-1)}$$

$$= E(S_X^2) + \frac{\left(n-(j-i)\right)(j-i)}{n(n-1)}c^2$$

End of the proof. □

Obviously the expected value of the alternative variance is larger than the expected value of the sample variance in the null distribution. The Monte Carlo simulation in Yao(1993) assumed the variance to be known. So the simulations process missed the

important division step. We will address the adjustment in later of this thesis. We need to point out the soundness of the $Z_3/\hat{\sigma}$ trying to reduce influence from the larger variance. Let us focus on the nominator of Z_3.

Let $t_1, t_2, \cdots, t_n$ be iid random variables from $N(0,1)$ distribution.

$$X_k = \mu + t_k \sigma \quad (k = 1, 2, \cdots, n)$$

$$S_j - S_i - \frac{j-i}{n} S_n = (j-i)\mu + \sigma \sum_{k=i+1}^{j} t_k - \frac{j-i}{n}\left(n\mu + \sigma \sum_{k=1}^{n} t_k \right)$$

$$= (j-i)\mu - (j-i)\mu + \sigma \sum_{k=i+1}^{j} t_k - \frac{j-i}{n}\sigma \sum_{k=1}^{n} t_k$$

$$= \sigma \left[\sum_{k=i+1}^{j} t_k \left(1 - \frac{j-i}{n} \right) - \frac{j-i}{n}\left(\sum_{k=1}^{i} t_k + \sum_{k=j+1}^{n} t_k \right) \right] \tag{6}$$

We can easily observe the parameter mean μ was eliminated. So $Z_3/\hat{\sigma}$ somehow makes sense to expect the $\sigma/\hat{\sigma}$ offset at least part of variance increase during the epidemic period. Unfortunately, Z_3 is still a parametric test statistic. The estimator of σ is serious matter about the underlying unknown distribution in realistic epidemic cases. Yao(1993) skipped the estimation of variance. That makes the results higher power than those of the estimation process. Therefore, we need to find a new approach for detecting epidemic change which can bypass point estimation of the population variance. Meanwhile we also keep the new test free from depending on population mean.

4. New Test Statistic

A new test statistic for detecting epidemic changes has been proposed in [8] . The test statistic is described as follows.

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent and normally distributed random variables with unknown variance σ^2 . The mean of X_i is determined by the following formula:

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

Here μ and δ are unknown parameters. 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. So we assume that (p,q) has equal chance to fall at any possible points $p = 1,\ldots,n-2$ and $q = p+1,\ldots,n-1$. To reduce the negative influence by extreme values in the data set, we set the $\min(q-p) \geq 6$ which covers at least range of data possible inside the epidemic duration. For the research convenience, only the one-sided alternative is considered for simplifying the discussion.

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent random variables, and let i and j be positive integers satisfying $m_0 \leq i < j < m_1$. And $X_{(1)} < X_{(2)} < ... < X_{(n)}$ are the corresponding order statistics.

The test statistic for detecting the epidemic alternative can then be defined as

$$T = \max_{m_0 \le j-i \le m_1} \frac{\left(S_j - S_i - \frac{j-i}{n} S_n \right) \Big/ \left\{ (j-i)\left(1 - \frac{j-i}{n} \right) \right\}^{\frac{1}{2}}}{\sum_{k=[n/2]+1}^{n} X_{(k)} - \sum_{k=1}^{[n/2]} X_{(k)}} \tag{7}$$

Here $\left[\frac{n}{2} \right]$ is the integer part of $\frac{n}{2}$.

The statistic T is independent of μ and σ. The independency can be seen from the following proposition.

Proposition 2. Let $X_1, X_2, \cdots, X_n$ be an iid sample from a population distribution with mean μ and variance σ^2, and let $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be the corresponding order statistics. Let T be defined as above. Then the distribution of T does not depend on μ and σ^2.

Proof. Let $t_1, t_2, \cdots, t_n$ be iid random variables from $N(0,1)$ distribution.

$$X_k = \mu + t_k \sigma \quad (k = 1, 2, \cdots, n)$$

Then

$$T = \max_{m_0 \le j-i \le m_1} \frac{\left[\sum_{k=i+1}^{j} t_k \left(1 - \frac{j-i}{n} \right) - \frac{j-i}{n} \left(\sum_{k=1}^{i} t_k + \sum_{k=j+1}^{n} t_k \right) \right]}{\left(\sum_{k=[n/2]+1}^{n} t_{(k)} - \sum_{k=1}^{[n/2]} t_{(k)} \right) \left\{ (j-i) \left(1 - \frac{j-i}{n} \right) \right\}^{\frac{1}{2}}}.$$

From (6),

$$S_j - S_i - \frac{j-i}{n} S_n = \sigma \left[\sum_{k=i+1}^{j} t_k \left(1 - \frac{j-i}{n} \right) - \frac{j-i}{n} \left(\sum_{k=1}^{i} t_k + \sum_{k=j+1}^{n} t_k \right) \right]$$

and

$$\sum_{k=n-[n/2]+1}^{n} X_{(k)} - \sum_{k=1}^{[n/2]} X_{(k)} = \left[\frac{n}{2} \right] \mu + \sigma \sum_{k=[n/2]+1}^{n} t_{(k)} - \left(\left[\frac{n}{2} \right] \mu + \sigma \sum_{k=1}^{[n/2]} t_{(k)} \right)$$

$$= \sigma \left(\sum_{k=n-[n/2]+1}^{n} t_{(k)} - \sum_{k=1}^{[n/2]} t_{(k)} \right). \qquad (8)$$

Then we have

$$T = \max_{m_0 \le j-i \le m_1} \frac{\sigma \left[\sum_{k=i+1}^{j} t_k \left(1 - \frac{j-i}{n} \right) - \frac{j-i}{n} \left(\sum_{k=1}^{i} t_k + \sum_{k=j+1}^{n} t_k \right) \right] \Big/ \left\{ (j-i) \left(1 - \frac{j-i}{n} \right) \right\}^{\frac{1}{2}}}{\sigma \left(\sum_{k=[n/2]+1}^{n} t_{(k)} - \sum_{k=1}^{[n/2]} t_{(k)} \right)}$$

$$= \max_{m_0 \le j-i \le m_1} \frac{\left[\sum_{k=i+1}^{j} t_k \left(1 - \frac{j-i}{n} \right) - \frac{j-i}{n} \left(\sum_{k=1}^{i} t_k + \sum_{k=j+1}^{n} t_k \right) \right]}{\left(\sum_{k=[n/2]+1}^{n} t_{(k)} - \sum_{k=1}^{[n/2]} t_{(k)} \right) \left\{ (j-i) \left(1 - \frac{j-i}{n} \right) \right\}^{\frac{1}{2}}}.$$

End of the proof. □

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

Therefore we need to find critical values at significance levels for this proposed test statistic and conduct power comparison. In the present paper, we only consider the one sided test and set the significance level $\alpha = 0.05$. Also we will compare point estimator of the change points obtained by the proposed method with existing methods in the literature.

5. Critical values and power comparison

5.1 Critical values of test statistic T

To obtain the critical values of test statistic T , this research used the Monte Carlo simulation in SAS/IML.

The procedure for the critical value calculation of T is summarized as follows:

1) For n=20, 21, $\cdots$, 100, generate $X_1, X_2, \cdots, X_n$ pseudo random samples from the standard normal distribution $N(0,1)$.

2) Let $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ be the ordered data set of the generated data $X_1, X_2, \cdots, X_n$.

3) Calculate max $\dfrac{\left(S_j - S_i - \dfrac{j-i}{n} S_n\right) \Big/ \left\{(j-i)\left(1-\dfrac{j-i}{n}\right)\right\}^{\frac{1}{2}}}{\sum_{k=[n/2]+1}^{n} X_{(k)} - \sum_{k=1}^{[n/2]} X_{(k)}}$.

4) Repeat steps (1) to (3) 100,000 times.

5) Finally, the critical values are obtained at 90, 95, 97.5, 99, and 99.5 percentiles.

The critical values are listed in Table 1 at page 25. The first column in the table is for the sample size (from 20 to 100). The critical values corresponding to significance levels α = 0.10, 0.05, 0.025, 0.01, and 0.005 are listed through column 2 to column 6. Table 1 keeps 7 decimal places for the critical values.

5.2 Power comparison at n=60

In my research, Monte Carlo simulation is used to simulate the distribution of the test statistic T described in (8). Simulation study is also used for test statistic Z_3 described in (3) with the unknown variance replaced by its estimate. For the generated pseudo

random standard normal numbers, each number is divided by the sample standard deviation $\hat{X}_k = X_k / \hat{\sigma}$. Then we obtained the new critical values for Z_3. Table 2 compares the power of Z_3 with the assumption of known variance, Z_3 with unknown variance, and T with 100,000 repetition Monte Carlo experiments. To keep the setting used in Yao(1993), sample size n=60 and significance level $\alpha = 0.05$ are considered. To see how the sample size effects the power, sample sizes n=20 is also used with the same significance level. Then the critical value of Z_3 with the assumption of known variance is 3.410, while the one in Yao (1993) is 3.37. The difference is not a major concern because the computers are much more powerful than twenty years ago. However, when the data were standardized with estimated variance, the critical value shifted away. The lower power performance indicates Yao(1993) Monte Carlo simulation did not consider this important step. The estimation process apparently lowers the power. Figure 1 indicates that without the estimation, T outperforms Z_3 with unknown variance when n is 60 and the epidemic duration is 6. That means when epidemic character data moderate, test statistic T is more powerful than test statistic Z_3. Figure 1 to Figure 4 show that T always have the higher power across δ = 0.8, 1.2, 1.6, and 2.4. The great benefit of T is it can stay away from variance estimation. Therefore, test statistic T is able to detect the epidemic in other unknown distributions of data.

The procedure for the power calculation for T is summarized as follows:

1) Generate $X_1, X_2, \cdots, X_n$ pseudo random samples from the standard normal distribution $N(0,1)$.

2) Input the δ values in the given duration of generated data set.

3) Calculate the values of the T statistic. In the Z_3 calculation, the data set is divided by the sample standard deviation before calculating the values of test statistic.

4) Let $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ be the ordered data set of the generated data $X_1, X_2, \cdots, X_n$.

5) Counting the number of values exceeding the critical value at the given significance level. The percentage of the numbers that are above the critical value is the simulated power.

Table 2

		n=60		
		Z_3 known var $m_0 = 6 \;\; m_1 = 54$	Z_3 unknown var $m_0 = 6 \;\; m_1 = 54$	T unknown var $m_0 = 6 \;\; m_1 = 54$
Δ	q-p			
	6(54)			
0.8		0.178	0.173	0.182
1.2		0.431	0.386	0.404
1.6		0.740	0.677	0.694
2.4		0.991	0.981	0.983
	10(50)			
0.8		0.341	0.323	0.335
1.2		0.721	0.675	0.686
1.6		0.949	0.922	0.926
2.4		1.000	1.000	1.000
	20(40)			
0.8		0.594	0.567	0.580
1.2		0.938	0.915	0.919
1.6		0.998	0.995	0.995
	30			
0.4		0.191	0.208	0.217
0.8		0.655	0.635	0.647
1.2		0.963	0.949	0.951
Critical		3.410	3.322	0.069

Figure 1 Power with n=60 and q-p=6

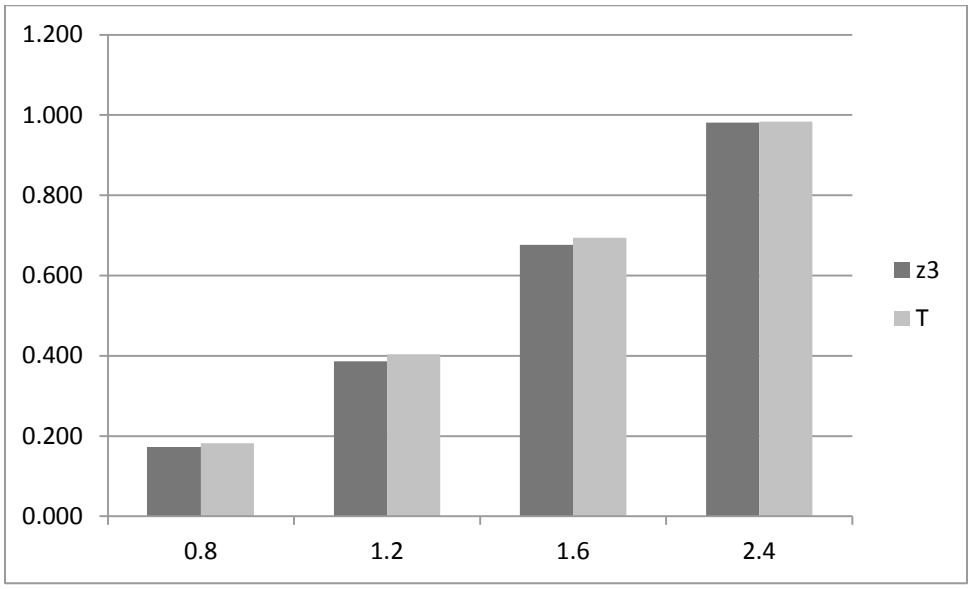


Figure 2 Power with n=60 and q-p=10

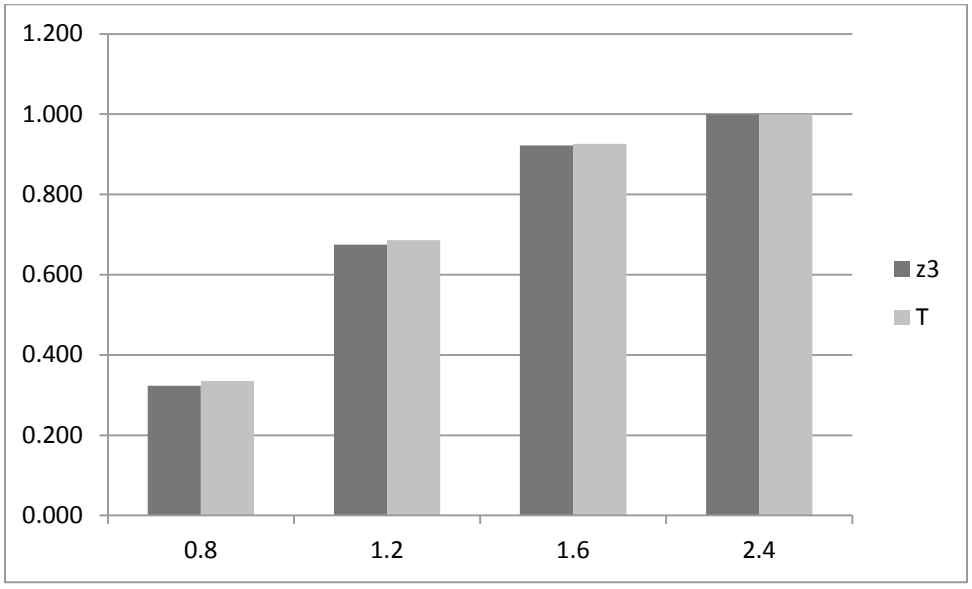


Figure 3 Power with n=60 and q-p=20

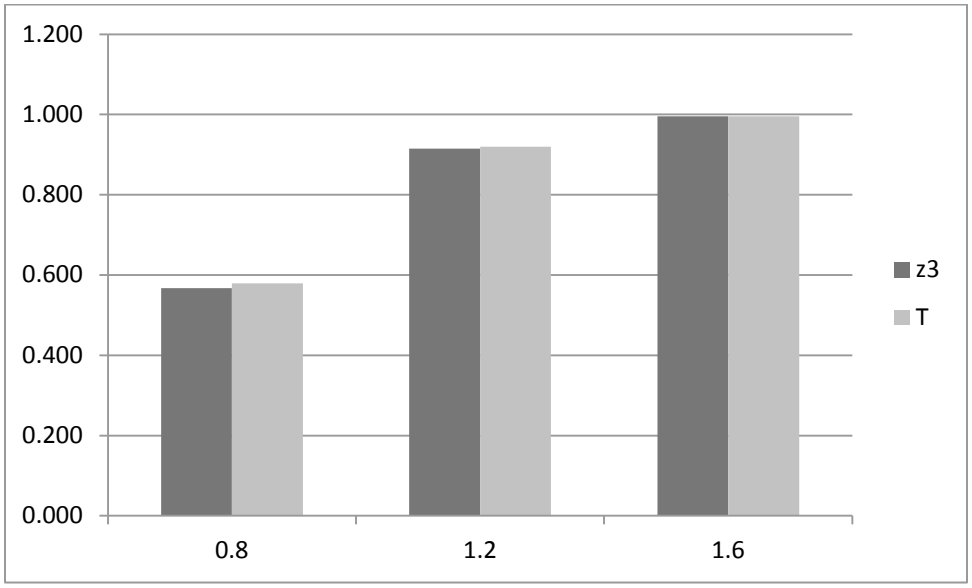
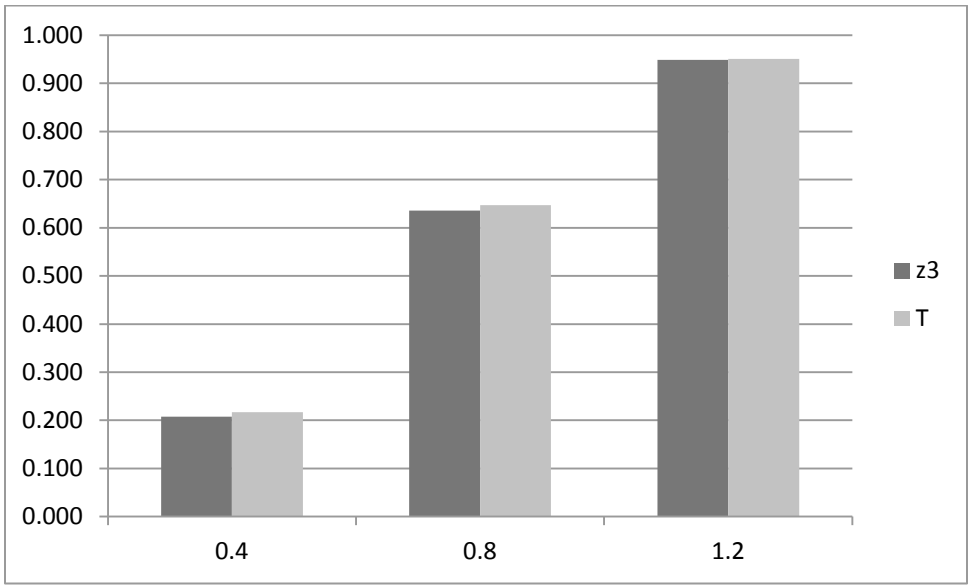


Figure 4 Power with n=60 and q-p=30



When δ=0.8, $q-p$ is near 0, all of the powers are weakened. Also when $q-p$ closed to 54 which is closed to n in this case, the power becomes fading away. The $q-p=20$ case shows the highest power and is followed by $q-p=10$ case. The potential explanation is that when the epidemic duration is too short or too long, the test statistics have has less in detecting epidemic changes.

5.3 Power comparison at n=20

Table 3

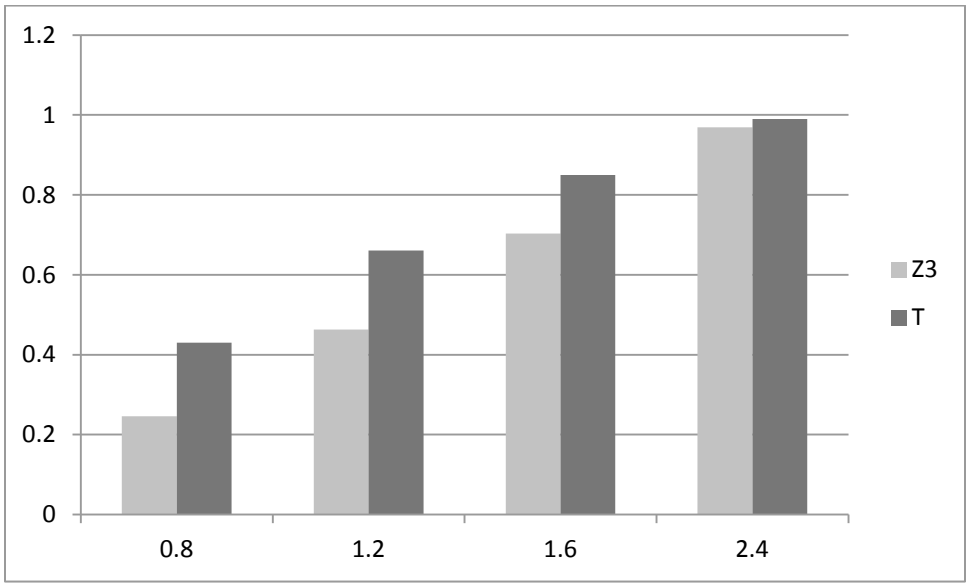
	n=20		
	Z_3 known var	Z_3 unknown var	T unknown var
q-p	min(j-i)=3	min(j-i)=3	min(j-i)=3
10			
0.8	0.296	0.246	0.430
1.2	0.578	0.463	0.661
1.6	0.839	0.703	0.850
2.4	0.996	0.969	0.990
Critical	3.077	2.882	0.169

To see the effect of sample size on the power, n=20 case was used for the three test statistics. Table 3 compares the powers at $q-p$ =10. Under the unknown variance, T shows the steep power over Z_3 to detect the relative long epidemic duration in the short

time period. Critical values with significance level $\alpha = 0.05$ are list at the bottom of Table 3.

In Figure 5, it shows that the test statistic T is more powerful than over Z_3 when δ =0.8, 1.2, 1.6, and 2.4. Especially when the abnormal data is moderate, such as $\delta = 0.8$, the new test statistic T indicates the advantage of parameter free statistic is very outstanding. The relative short period of data collection bearing the mid-range of epidemic characteristic data is an ideal condition for the test statistic T .

Figure 5 Power with n=20 and q-p=10



6. Conclusion

Change-points with epidemic alternatives were formulated 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. Five test statistics for testing epidemic alternatives were summarized and compared in Yao [1].

All the summarized test statistics in Yao [1] are based an assumption that the population variance is known. The power comparison on those test statistics presented in Yao's paper was based on that assumption as well. In practice, the population variance is usually unknown and needs to be estimated. This thesis extended the power comparison result using Monte Carlo simulation. It has been shown that the power of the test statistics mentioned in Yao [1] is lowered when the population variance is assumed to be unknown and needs to be estimated.

To stay away from the known variance assumption, a new method for detecting change-points with epidemic alternative is studied in this research. Since this method is independent of the population mean μ and the population standard deviation σ, 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 existing test statistics summarized in Yao [1], the statistical test presented in this study provides quite decent power. Therefore, the proposed test statistic T should be recommended to practitioners for detecting epidemic change in means.

The newly proposed test statistic is, in fact, a modified version of Z_3, which is one of the test statistics and is the best test statistics out of those five according to Yao's

observation. Monte Carlo simulation result outlines the behaviors of test statistics Z_3 and T under the unknown variance and unknown mean conditions. The test statistic T outperformed the test statistic Z_3 in all of cases. Critical values for the proposed test statistics are obtained using Monte Carlo simulation.

Table 1 Critical Values of Test Statistic T

n	α=0.1	α=0.05	α=0.025	α=0.01	α=0.005
20	0.1532661	0.1688606	0.1815961	0.1951939	0.2036749
21	0.1486433	0.163719	0.1760061	0.1893551	0.1978356
22	0.1436	0.1576624	0.1692808	0.1819009	0.1898116
23	0.1395886	0.1530257	0.1641669	0.1766411	0.1845658
24	0.1352017	0.1479896	0.1584798	0.1703081	0.1779986
25	0.1313733	0.1436647	0.1539657	0.1654042	0.1730334
26	0.1273106	0.139062	0.1488865	0.1598907	0.1672009
27	0.1240343	0.1353173	0.144956	0.1556937	0.1626133
28	0.1204688	0.1313354	0.1404544	0.1508054	0.1575053
29	0.117395	0.1278418	0.1367407	0.1468518	0.153502
30	0.1141603	0.1242228	0.132751	0.1424982	0.1489147
31	0.11144	0.121333	0.1297131	0.1392189	0.1454978
32	0.1085073	0.1179114	0.12601	0.1349894	0.1412021
33	0.1060941	0.1152564	0.1231115	0.1320777	0.1381073
34	0.1035453	0.1123141	0.1199995	0.128639	0.1344773
35	0.1012103	0.1098547	0.117291	0.1257478	0.1315655
36	0.0988655	0.1072324	0.1143707	0.1226992	0.1282133
37	0.0967521	0.1048443	0.1118236	0.1199389	0.125405
38	0.0945904	0.1024628	0.1092844	0.1171646	0.1223346
39	0.092658	0.1003772	0.1070123	0.1147611	0.1199426
40	0.0907363	0.0982262	0.1047477	0.112183	0.1173327
41	0.0889316	0.0962027	0.102564	0.1100055	0.114959
42	0.0871319	0.0942406	0.1004684	0.1076947	0.112462
43	0.0854987	0.092525	0.0986448	0.105662	0.1102375
44	0.0838474	0.0906678	0.0966285	0.1035174	0.1081272
45	0.0823201	0.0889657	0.0947917	0.1014992	0.1061885
46	0.0808164	0.0872768	0.0929381	0.0994785	0.1040623
47	0.079386	0.0857007	0.0912827	0.0977883	0.1021606
48	0.0779583	0.0841454	0.0895448	0.0958764	0.100182
49	0.076659	0.0827643	0.0880865	0.0942737	0.0984957
50	0.0753504	0.0812772	0.0865062	0.0926387	0.0967903
51	0.0741816	0.07999	0.0850953	0.0910755	0.0951482
52	0.0728893	0.0786194	0.0836516	0.0894618	0.0934648
53	0.0717446	0.077344	0.0822736	0.0879371	0.0919361
54	0.070588	0.0760621	0.0809155	0.0866086	0.090482
55	0.0695295	0.0748737	0.0796092	0.085243	0.0890668
56	0.0684147	0.073686	0.0783067	0.0837234	0.0874351
57	0.0673958	0.072593	0.0772241	0.0826052	0.0863144
58	0.0664017	0.0715119	0.0759939	0.0813078	0.0849526
59	0.065438	0.0704479	0.0748695	0.0800886	0.083732

n	α=0.1	α=0.05	α=0.025	α=0.01	α=0.005
60	0.0644937	0.0694155	0.0737757	0.0788786	0.0823889
61	0.0636117	0.0684232	0.0726876	0.0777018	0.0812084
62	0.0626863	0.0674441	0.0716342	0.0766683	0.0800565
63	0.0618954	0.0665836	0.0707145	0.0756277	0.0790058
64	0.0610002	0.0656012	0.0697209	0.0745806	0.0778689
65	0.0602141	0.064736	0.0687764	0.0735302	0.0768127
66	0.0594063	0.0638759	0.0678395	0.0725446	0.0757075
67	0.0586861	0.063085	0.0669783	0.0715431	0.0747156
68	0.0579543	0.0622702	0.0661282	0.0707288	0.0738465
69	0.0571921	0.0614659	0.0652526	0.0696836	0.0727954
70	0.056475	0.0606684	0.0643513	0.0688171	0.0718091
71	0.0557922	0.0599292	0.0635872	0.067996	0.0709522
72	0.0550785	0.0591365	0.0627754	0.0670345	0.0700042
73	0.0544293	0.058464	0.0620465	0.0662155	0.0690837
74	0.053828	0.0577943	0.0613113	0.0655321	0.0684001
75	0.0531799	0.0571118	0.0605593	0.0646665	0.0675556
76	0.0525564	0.0564007	0.0598313	0.0638474	0.0667195
77	0.0519898	0.0558021	0.0591824	0.0631797	0.0659648
78	0.051383	0.0551666	0.0584507	0.0624475	0.0652155
79	0.0508237	0.0545264	0.0578269	0.0616876	0.0644893
80	0.0502454	0.0538859	0.0571635	0.0609614	0.0636071
81	0.0497292	0.0533454	0.0565673	0.0603669	0.0629977
82	0.0491834	0.0527573	0.0559427	0.0596773	0.0622752
83	0.0486995	0.0522011	0.0553312	0.05904	0.0616454
84	0.0481418	0.0515739	0.054662	0.0583653	0.0608823
85	0.0476672	0.0510737	0.0541127	0.0577744	0.060308
86	0.0471643	0.0505249	0.0535292	0.0571333	0.0596431
87	0.0466986	0.0500552	0.0530529	0.0566104	0.0590816
88	0.0462335	0.049529	0.0524598	0.0559853	0.0584285
89	0.0457678	0.0490089	0.0519487	0.0553663	0.0577658
90	0.0453462	0.0485507	0.0514704	0.0549194	0.057346
91	0.0449015	0.0480728	0.0509334	0.0543599	0.0567078
92	0.0444388	0.0475599	0.0503767	0.0537567	0.0561928
93	0.0440412	0.047124	0.0499416	0.0532581	0.0555155
94	0.0436293	0.0466984	0.0494549	0.0527327	0.0550314
95	0.0432234	0.0463091	0.048999	0.0522569	0.0545469
96	0.0428047	0.0458065	0.0485216	0.0517615	0.053981
97	0.042428	0.0453912	0.0480654	0.0512749	0.0534982
98	0.0420359	0.0449814	0.0476233	0.0507524	0.0528879
99	0.0416873	0.0445977	0.0472265	0.0503869	0.0525505
100	0.0412777	0.044173	0.0467718	0.0498619	0.051985

7. References

[1] Yao Q. Test for change-points with epidemic alternatives. Biometrika. 1993; 80:179-191.

[2] Levin B, Kline J. The cusum test of homogeneity with an application in spontaneous abortion epidemiology. Statistics in Medicine. 1985; 4:469-488.

[3] Siegmund,D.(1986). Boundary crossing probabilities and statistical applications. Ann. Statist. 14, 361-404.

[4] A.N. Pettitt, Some results on estimating a change-point using nonparametric type statistics, J. Statist. Comput. Simul. 11 (1980) 261-272.

[5] James, B., James, K.L. & Siegmund,D.(1987). Tests for a change-point. Biomerika 74, 71-83.

[6] Brown, R. T., Durbin, J. & Evans, J.M. (1975). Techniques for testing the constancy of regression relations over time. J. R. Statist. Soc. B 37, 149-192.

[7] Cox, D. R. & Hinkley, D.V. (1974). Theoretical Statistics. London: Chapman and Hall.

[8] Chen Z., Li Z. and Zhou M. Detecting Change-Points in Epidemic Models, in submission.

[9] Page ES.A test for a change in a parameter occurring at an unknown point. Biometrika. 1995; 42: 523-527.