Liet. matem. rink., **44**, *spec. nr.*, 2004, 589–593

Power analysis of dyadic increment (DI) statistic

Danas ZUOKAS (VU)
e-mail: danaz78@one.lt

1. Introduction

Let $X_1, \ldots, X_n$, be a sequence of independent binomial random variables with

$$\mathrm{P}(X_i = 1) = \mu_i, \quad \mathrm{P}(X_i = 0) = 1 - \mu_i, \quad i = 1, \ldots, n.$$

We want to test the null hypothesis of a constant occurrence probability,

$$\mathrm{H}_0\colon \mu_1 = \cdots = \mu_n := \mu_0,$$

under the following so called changed segment (or epidemic) alternative H_A: there exist integers k^* and m^*, $0 \leqslant k^* < m^* \leqslant n$, and $\mu_\mathrm{A} \neq \mu_0$ ($\mu_0, \mu_\mathrm{A} \in (0,1)$) such that

$$\mathrm{P}(X_i = 1) = \begin{cases} \mu_\mathrm{A}, & i \in \{k^* + 1, \ldots, m^*\}, \\ \mu_0, & i \in \{1, \ldots, n\} \setminus \{k^* + 1, \ldots, m^*\}. \end{cases}$$

Here k^* stands for the beginning, m^* the end, $l^* = m^* - k^*$ the length of epidemic. The quantity $|\mu_\mathrm{A} - \mu_0|$ is referred to the size of epidemic. If the null hypothesis is rejected, the next step is to estimate the parameters of the epidemic: l^*, k^*, m^*, μ_0 and μ_A. We deal only with the hypothesis testing problem. Note that the problem of epidemic change in an occurrence probability can also be reformulated in the terms of epidemic change in the means of observations, because $\mathrm{E}X_i = \mathrm{P}(X_i = 1) = \mu_i$. And so all the methods for testing epidemic change in the means can be applied. The most common ones are the maximum likelihood and those based on cumulative sums.

For a short survey of epidemic change problem we refer to Csörgő and Horváth [2], where basically the cumulative sum type test statistics to test for epidemic change in the mean of random variables are discussed. We also refer to Yao [6], where several statistics of different types are analyzed in the case of normally distributed observations. The problem of a changed segment in a binomial sequence was considered by Curnow and Fu [3] and Avery and Henderson [1]. Several cumulative sum type statistics were introduced by Račkauskas and Suquet [5] for the sequences of random elements with values in abstract measurable spaces. The distinct feature of these statistics is certain weight functions ρ from some class $\mathcal{R}$ (see [5] for the definition), which, when changing the parameters of the weight function, allow to detect epidemics of various lengths.

In the next section we present DI statistic, formulate the theorems of convergence under H_0 and consistency for this statistic and give the critical values. Then we investigate the influence of the parameters of epidemic and weight function to the power of DI statistic. We end up with conclusions.

2. Dyadic increment statistic

We refer to [5] for details. Denote by D_j the set of dyadic numbers in $[0, 1]$ of level j:

$$D_0 = \{0, 1\}, \quad D_j = \left\{(2l-1)2^{-j};\ 1 \leqslant l \leqslant 2^{j-1}\right\}, \quad j \geqslant 1. \tag{1}$$

For $r \in D_j$, $j \geqslant 0$, denote $r^- = r - 2^{-j}$ and $r^+ = r + 2^{-j}$. If we set $S(0) = 0$ and $S(t) = \sum_{i \leqslant t} X_i$, $0 < t \leqslant n$, then $\overline{X} = S(n)/n$ and the dyadic increment statistic is

$$\mathrm{DI}_n(\rho) = \frac{1}{(n\overline{X}(1-\overline{X}))^{1/2}} \max_{1<2^j \leqslant n} \frac{1}{\rho(2^{-j})} \max_{r \in D_j} \left| S(nr) - \frac{S(nr^-)}{2} - \frac{S(nr^+)}{2} \right|. \tag{2}$$

In our analysis we use the following expression of the weight function

$$\rho(h) = \rho(h, \alpha, \beta, \gamma) = h^{\alpha} \log^{\beta}(\gamma/h), \quad 0 < h \leqslant 1. \tag{3}$$

This is a proper weight function if either $\alpha \in (0, 1/2)$ and $\beta \in \mathbb{R}$ or $\alpha = 1/2$ and $\beta > 1/2$. We also analyze the case where $\alpha = 0$ and $\beta = 0$ (no weight function used). The problem with using the parametric weight functions is that there is no strict rule for assigning certain values to parameters. It therefore remains interesting and open theoretical question of data driven choice of parameters.

Let $(W(t),\ t \in [0, 1])$ be a standard Wiener process. The distribution of

$$\mathrm{DI}(\rho) = \sup_{j \geqslant 1} \frac{1}{\rho(2^{-j})} \max_{r \in D_j} \left| W(r) - \frac{1}{2} W\left(r^-\right) - \frac{1}{2} W\left(r^+\right) \right| \tag{4}$$

serves as limiting distribution for DI statistic under H_0. This is formulated as a theorem.

THEOREM 1. *When* H_0 *holds and* ρ *is as in* (3), *then* $\mathrm{DI}_n(\rho) \xrightarrow[n\to\infty]{\mathcal{D}} \mathrm{DI}(\rho)$.

This theorem is a special case of a more general result proved in [5] for any sequence of independent identically distributed random variables and weight function ρ from a broader class $\mathcal{R}$.

Proof. Since in the special case of binomial random variables $\overline{X}(1-\overline{X})$ is an estimate of the variance of X_1, the result follows from Theorem 3 and Remark 2 in [5].

Next assume that l^* and $n - l^*$ tend to infinity with $n \to \infty$. For the consistency of statistic $\mathrm{DI}_n(\rho)$ we formulate Theorem 2.

Table 1. The critical values

	$\alpha=0$ $\beta=0$	$\alpha=1/8$ $\beta=0$	$\alpha=1/4$ $\beta=0$	$\alpha=3/8$ $\beta=0$	$\alpha=1/2$ $\beta=1$	$\alpha=1/2$ $\beta=0.6$
$\alpha_s=0.10$	0.8864	1.0124	1.1930	1.5310	0.7460	1.0400
$\alpha_s=0.05$	1.0163	1.1441	1.3210	1.6430	0.8510	1.1410
$\alpha_s=0.01$	1.2965	1.4316	1.6070	1.9010	1.0830	1.3810

THEOREM 2. *Suppose that* $\mathrm{H_A}$ *holds and* ρ *is as in* (3). *Moreover, assume*

$$\lim_{n\to\infty} \frac{n^{1/2}h_n|\mu_\mathrm{A}-\mu_0|}{\rho(h_n)} = \infty, \tag{5}$$

where $h_n=(l^*/n)(1-l^*/n)$. *Then* $\mathrm{DI}_n(\rho) \underset{n\to\infty}{\xrightarrow{\mathrm{P}}} \infty$.

The proof of this result goes along the lines of the proofs of Theorem 4 in [5] and Theorem 2 in [7] and is omitted due to length restriction for the paper.

We can now give the argument pro the weight function. Assume for a moment that $l^*/n\to 0$. If $\alpha=0$ and $\beta=0$, condition (5) reduces to $l^*/n^{1/2}\to\infty$, i.e. the length of epidemic should tend to infinity faster than $n^{1/2}$ for consistency of the statistic. Similarly, when $\alpha<1/2$, $\beta=0$, l^* should significantly exceed $n^{(1-2\alpha)/(2-2\alpha)}$. For example, taking $\alpha=1/4$, the length of epidemic should be such that $n^{1/3}=o(l^*)$. Thus when using the weight function we have the possibility to detect shorter epidemics.

The major advantage of DI statistic is that the exact analytical form for the distribution of $\mathrm{DI}(\rho)$ is known. We can therefore find precise critical values associated with the certain significance level α_s. In Table 1 critical values are given for various α and β (furher we fix $\gamma=\exp(1)$).

3. The power analysis

First we investigate how the power of DI statistic depends on l^*, n, μ_0, μ_A and k^*. We choose $\rho(h)=h^\alpha$ ($\beta=0$) and $\alpha=1/4$ in (3). The convenient way for power analysis is the so called size-power curves on a correct size-adjusted (not nominal size) basis (see Davidson and MacKinnon [4]). For every set of parameter values we compute 10000 replications of the statistics and the corresponding p-values first for the sample with no changed segment then for the same sample but now with the changed segment at $\{k^*+1,\dots,m^*\}$. We plot the empirical cumulative distribution function for p-values under $\mathrm{H_A}$ (which is the empirical power function) but on x-axis we have the values of empirical distribution function for p-values under $\mathrm{H_0}$ instead of nominal size α_s. That is we adjust the power to size. The results are presented in Fig. 1.

First let l^* increase all other parameters keeping fixed. From Fig. 1(a) we see that the power increases quite rapidly. Next, we take $k^*=n/2$. From Fig. 1(b) we conclude that the power decreases with n increasing because the length of epidemic relatively to the number of observations decreases. Increase $|\mu_\mathrm{A}-\mu_0|$. We see in Fig. 1(c) that the power increases and again very quickly. We observe rather interesting effect, which

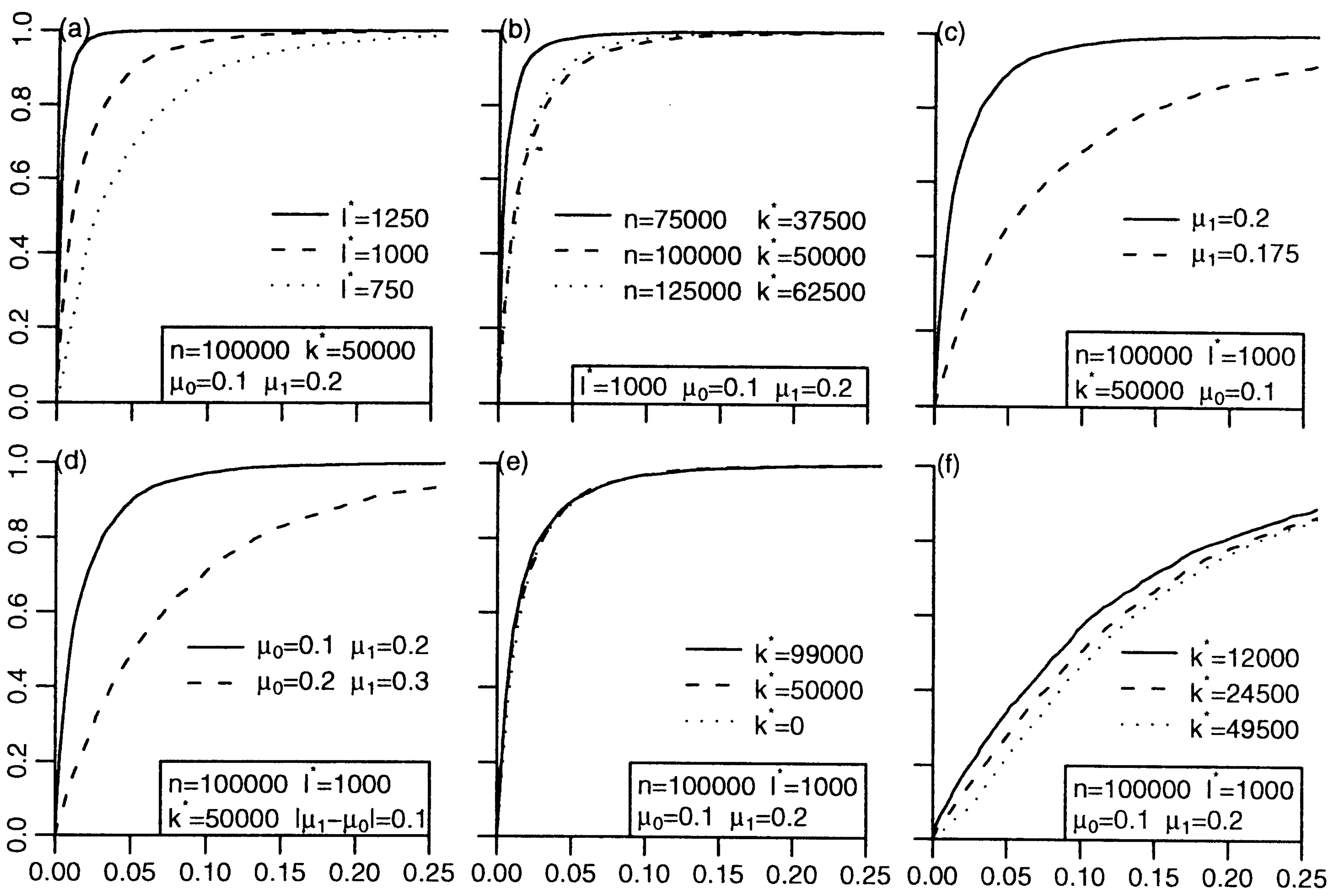

Fig. 1. The adjusted size-power curve plots.

was mentioned in Avery and Henderson [1]. Namely, that shifting both μ_0 and μ_A without changing $|\mu_A - \mu_0|$ (keeping other parameters fixed) decreases the power. Fig. 1(d) illustrates this effect. Figs. 1(e) and 1(f) demonstrate that the beginning of the epidemic has no impact on the power of the statistic as long as there is some level j of dyadic splitting and some r from D_j (see (1)) when whole epidemic is located between nr and either nr^- or nr^+ and the length of the epidemic is comparable with $n2^{-j}$. Indeed, from (2) we see that, when testing for the epidemic, we take differences of partial sums of $n2^{-j}$ observations to the left and to the right of nr. So if, in the worst case, the middle of the epidemic is right on nr and the length is $n2^{-j}$ then statistic may not detect it. Of course, at some higher level of the dyadic splitting it will nonetheless detect the epidemic but it will be shorter and hence harder to detect.

We also analyze the impact of the weight function ρ (as in 3) to the power of the DI statistic. In our next numerical simulations we fixed $n = 100000$, $l^* = 1000$, $k^* = 50000$ and $\mu_0 = 0.1$, $\mu_A = 0.2$. For various α and β we have computed the values of the empirical power function for significance levels $\alpha_s = 0.1, 0.05$ and 0.01. In Table 2 we present the results of simulations.

From Table 2 it is clear that numerical simulations confirm the theoretical reasoning. Take for example $\alpha_s = 0.05$, which is the typical significance level in most practical applications. If we do not use weight function, we detect only about a quarter of epidemics (2560 out of 10000 in our calculations) compared to almost all cases when taking $\alpha = 3/8$ and $\beta = 0$. Even more impressive results are for smaller significance

Table 2. The power of the DI statistic for various α and β values

	$\alpha=0$ $\beta=0$	$\alpha=1/8$ $\beta=0$	$\alpha=1/4$ $\beta=0$	$\alpha=3/8$ $\beta=0$	$\alpha=1/2$ $\beta=1$	$\alpha=1/2$ $\beta=0.6$
$\alpha_s=0.10$	0.4259	0.6745	0.9685	0.9991	0.8823	0.9972
$\alpha_s=0.05$	0.2560	0.4035	0.8962	0.9965	0.6349	0.9870
$\alpha_s=0.01$	0.0781	0.1008	0.5033	0.9708	0.1231	0.8595

level $\alpha_s = 0.01$. The case $\alpha = 0$ and $\beta = 0$ demonstrates almost no power while the case $\alpha = 3/8$ and $\beta = 0$ detects about 97% of the epidemics.

4. Conclusions

Even for large n the computation of $\mathrm{DI}_n(\rho)$ (defined in (2)) is very quick. The distribution of $\mathrm{DI}(\rho)$ is known therefore the precise critical values can be found. DI statistic is a powerful tool and numerical simulations support this statement. These are the main advantages. As a drawback we can point that this statistic is too rough to estimate the length or the beginning of the epidemic. Also the choice of the parameters of the weight function is not determined. Moreover the location of the epidemic can be very important for the power of statistic. Concluding, DI statistic can be effectively used when for the large number of observations the presence of a changed segment is tested and estimating the parameters of the epidemic is not required. The computations usually are very quick.

References

1. P.J. Avery, D.A. Henderson, Detecting a changed segment in DNA sequences, *J. R. Stat. Soc., Ser. C – Appl. Stat.*, **48**, 489–503 (1999).
2. M. Csörgő, L. Horváth, *Limit Theorems in Change-Point Analysis*, John Wiley & Sons, Baffins Lane, Chichester (1997).
3. R.N. Curnow, Y.-X. Fu, Locating a changed segment in a sequence of Bernoulli variables, *Biometrika*, **77**, 295–304 (1990).
4. R. Davidson, J.G. MacKinnon, Graphical methods for investigating the size and power of hypothesis tests, *Manchester School*, **66**, 1–26 (1998).
5. A. Račkauskas, Ch. Suquet, Hölder norm test statistics for epidemic change, *Pub. IRMA Lille 59-III* (2003).
6. Q. Yao, Tests for change-points with epidemic alternatives, *Biometrika*, **80**, 179–191 (1993).
7. D. Zuokas, Detecting and locating a changed segment in a binomial sequence: comparison of tests (submitted).

REZIUMĖ

D. Zuokas. Diadinių prieauglių (DI) statistikos galios analizė

Šiame darbe nagrinėjamas epideminio pasikeitimo binominėje sekoje uždavinys. Šiam uždaviniui suformuluotos diadinių prieauglių (DI) statistikos konvergavimo ir suderinamumo teoremos. Monte-Karlo metodu tiriama DI statistikos galia. Rezultatai rodo, kad ši statistika tinka tirti ilgas binomines sekas, kai reikalinga (greita) išvada apie tai, ar yra epideminis pasikeitimas, toliau nevertinant epidemijos parametrų.