



## The Effect of Dependence in a Binary Sequence on Tests for a Change point or a Changed Segment

Peter J. Avery

*Applied Statistics*, Vol. 50, No. 2. (2001), pp. 243-246.

Stable URL:

<http://links.jstor.org/sici?sici=0035-9254%282001%2950%3A2%3C243%3ATEODIA%3E2.0.CO%3B2-1>

*Applied Statistics* is currently published by Royal Statistical Society.

---

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/rss.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

---

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).

# The effect of dependence in a binary sequence on tests for a changepoint or a changed segment

Peter J. Avery

University of Newcastle, UK

[Received September 1999. Revised November 2000]

**Summary.** The effect of partial dependence in a binary sequence on tests for the presence of a changepoint or changed segment are investigated and exemplified in the context of modelling non-coding deoxyribonucleic acid (DNA). For the levels of dependence that are commonly seen in such DNA, the null distributions of the test statistics are approximately correct and so conclusions based on them are still valid. A strong dependence would, however, invalidate the use of such procedures.

**Keywords:** Binary sequence; Changed segment; Changepoint; Deoxyribonucleic acid; Dependence

## 1. Introduction

Pettitt (1979) developed a test statistic for the existence of a changepoint  $\tau$  in a sequence of  $n$  Bernoulli random variables  $R_i$ , where

$$\Pr(R_i = 1) = \begin{cases} \theta_1, & i = 1, 2, \dots, \tau, \\ \theta_2, & i = \tau + 1, \tau + 2, \dots, n. \end{cases}$$

Avery and Henderson (1999) discussed this approach in the context of the analysis of non-coding deoxyribonucleic acid (DNA) sequence data and extended Pettitt's method to test for a changed segment, i.e. where

$$\Pr(R_i = 1) = \begin{cases} \theta_1, & i = 1, 2, \dots, \tau_1, \\ \theta_2, & i = \tau_1 + 1, \tau_1 + 2, \dots, \tau_2, \\ \theta_1, & i = \tau_2 + 1, \tau_2 + 2, \dots, n. \end{cases}$$

As pointed out by Avery and Henderson (1999), non-coding DNA can often be modelled by Bernoulli random variables if we code one of the four bases, e.g. T, as 1 and the other bases (e.g. A, C and G) as 0. Avery and Henderson (1999) looked at the four possible binary sequences from each of three introns in the human preproglucagon gene (Bell *et al.*, 1983). In nine cases, there was no evidence of a departure from successive bases being independently determined. However, in the remaining three cases there was evidence of dependence. The null distributions of the test statistics developed by Pettitt (1979) and Avery and Henderson (1999) for the existence of a changepoint or a changed segment assume that successive bases are independent. Avery and Henderson (1999) stated that

*Address for correspondence:* Peter J. Avery, Department of Statistics, University of Newcastle, Newcastle upon Tyne, NE1 7RU, UK.  
E-mail: P.J.Avery@ncl.ac.uk

‘Any departures from the independence assumption do not seem to affect greatly the ability of the test procedure to find changed segments but it may be that it will tend to inflate our test statistic slightly and so it would be sensible to be cautious over marginally significant results’.

In this paper we present the results of some simulations which broadly substantiate this conclusion for parameter values that are appropriate to the data in Avery and Henderson (1999) but highlight the gradual breakdown of the null distributions in more extreme cases of dependence.

## 2. Methods

Sequences were simulated from a first-order Markov chain model where

$$\Pr(R_i = 1) = \theta, \quad i = 1, 2, \dots, n,$$

and

$$\Pr(R_{i+1} = 1 | R_i = 1) = \theta + \delta, \quad i = 1, 2, \dots, n - 1.$$

Given  $\theta$  and  $\delta$ , all the other transition probabilities are defined. For example,

$$\Pr(R_{i+1} = 1 | R_i = 0) = \theta\{1 - \delta/(1 - \theta)\}, \quad i = 1, 2, \dots, n - 1.$$

If we estimate  $(\theta, \delta)$  for the three cases in Avery and Henderson (1999) where there was a significant departure from independence, we obtain (0.272, 0.060), (0.199, 0.064) and (0.213, 0.047) for T, C and G respectively in intron 2. An alternative way of converting a DNA sequence into a binary sequence is to code two bases as 1 and two as 0. The most obvious way to do this is to code A and G (purines) as 1 and C and T (pyrimidines) as 0. However, this very often leads to a significant dependence of successive bases and for the human preproglucagon gene of Bell *et al.* (1983) all three introns give significant  $\chi^2$ -statistics and  $(\theta, \delta)$  estimates of (0.472, 0.032), (0.529, 0.072) and (0.523, 0.041) respectively. (This is why we have not advocated this way of converting to a binary sequence previously.) We have thus carried out simulation studies for the cases  $\theta = 0.25, 0.5$  and  $\delta = 0, 0.05, 0.1$ .

The test statistic developed by Pettitt (1979) for the detection of a changepoint is

$$K_n = \max |nS_t - tS_n|,$$

where  $S_t = \sum_{i=1}^t R_i$ , whereas Avery and Henderson (1999) used

$$K_n^* = \max |nS_t - tS_n| - \min |nS_t - tS_n|$$

to detect a changed segment, where the maximization and minimization are over all possible values of  $t$ .

For large values of  $n$ ,

$$\Pr[K_n/\sqrt{\{nS_n(n - S_n)\}} \geq z] \simeq 2 \sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2 z^2) \tag{1}$$

(Gnedenko and Korolyuk, 1961) and

$$\Pr[K_n^*/\sqrt{\{nS_n(n - S_n)\}} \geq z] \simeq 2 \sum_{k=1}^{\infty} (4k^2 z^2 - 1) \exp(-2k^2 z^2) \tag{2}$$

(Avery and Henderson, 1999). For significance testing, we need the values of  $z$  which give certain specified values for these probabilities, i.e. significance levels,  $\alpha$  say. If  $\alpha$  is small we

**Table 1.** Percentage of simulations exceeding the cut-off values when testing for a changepoint (CP) or a changed segment (CS)†

$\alpha$ (%)	Percentages for the following values of $\delta$ :					
	0		0.05		0.1	
	CP	CS	CP	CS	CP	CS
$\theta = 0.25$						
5	4.4	3.7	6.3	6.5	10.3	11.9
1	0.7	0.7	1.4	1.4	3.3	3.2
0.1	0.1	0.1	0.3	0.4	0.5	0.5
$\theta = 0.5$						
5	5.0	4.0	8.1	8.7	14.1	18.4
1	0.9	0.8	2.1	2.1	4.5	6.6
0.1	0.0	0.0	0.4	0.3	0.9	1.2

†Each result is based on 3500 simulations.

can ignore terms with  $k > 1$  and solve for  $z$ . In the former case for testing for a changepoint, we find, using approximation (1),  $z_\alpha = 1.35810, 1.62762, 1.94947$  and in the latter case, for testing for a changed segment, we find, using approximation (2),  $z_\alpha = 1.74726, 2.00092, 2.30297$ , for  $\alpha = 0.05, 0.01, 0.001$  respectively. Table 1 gives the percentage of simulations exceeding these cut-off values. The results were very similar for  $n = 200, 300, 500, 1500$  and so they have been amalgamated in Table 1.

From the simulations with  $\delta = 0$ , we see that using the asymptotic null distribution is slightly conservative, particularly in the changed segment case, as discussed in Avery and Henderson (1999). For  $\theta = 0.25$  and  $\delta = 0.05$ , which are close to the values in the three cases in Avery and Henderson (1999) where there was a significant dependence, there are modest increases in the significance levels. As the test statistics for the existence of a changed segment were all significant at 0.1% in these three cases, it is clear that the dependence does not change the conclusions from these tests. If  $\theta = 0.5$  and  $\delta = 0.05$ , the increase in the significance level is larger but the conclusions from tests should still generally be reliable as long as at least a significance level of 1% is used. However, if  $\delta$  is close to 0.1, the tests must be used very cautiously, particularly if  $\theta$  is close to 0.5.

### 3. Discussion

If we take the three introns of the human preproglucagon gene of Bell *et al.* (1983) and code A and G (purines) as 1 as discussed above, we obtain values for  $K_n^*/\sqrt{\{nS_n(n - S_n)\}}$  of 2.07906, 3.67981 and 1.87181 respectively. Using equation (2) above gives significance levels of 0.6%, 0.0% and 2.4% respectively. However, using the simulation results for  $\theta = 0.5$  and  $\delta = 0.05$  gives empirical significance levels of 1.3%, 0.0% and 4.6%. It is clear that there is strong evidence of a changed segment in intron 2. The changed segment is estimated to be 723 bases long starting at base 401, which are close to the estimates based on C, A or G given in Avery and Henderson (1999). The evidence for a changed segment in intron 3 is very slight but there does now seem to be some evidence of a changed segment in intron 1. The changed segment in intron 1 is estimated to be 1018 bases long starting at base 257. Thus it goes approximately from window 12 to window 64 in the plot of intron 1 given in Avery and Henderson (1999).

In conclusion, the tests of Pettitt (1979) and Avery and Henderson (1999) can be used in the presence of some dependence between successive bases as long as the significance levels

are used judiciously. However, any strong dependence would make their use much more problematical.

### Acknowledgements

I am grateful to D. A. Henderson for assistance with the simulations and the preparation of this paper.

### References

- Avery, P. J. and Henderson, D. A. (1999) Detecting a changed segment in DNA sequences. *Appl. Statist.*, **48**, 489–503.
- Bell, G. I., Sanchez-Pescador, R., Laybourn, P. J. and Najarian, R. C. (1983) Exon duplication and divergence in the human preproglucagon gene. *Nature*, **304**, 368–371.
- Gnedenko, B. V. and Korolyuk, V. S. (1961) On the maximum discrepancy between two empirical distributions. In *Selected Translations in Mathematical Statistics and Probability*, vol. 1, pp. 13–16. Providence: American Mathematical Society.
- Pettitt, A. N. (1979) A non-parametric approach to the change-point problem. *Appl. Statist.*, **28**, 126–135.