INFERENCES FOR BINOMIAL CHANGE-POINT DATA

James M Freeman

Manchester Business School
University of Manchester
Booth Street West
Manchester, M15 6PB, UK
(e-mail: jim.freeman@mbs.ac.uk)

Abstract: In this paper we describe a procedure for detecting a systematic change in parameter for a sequence of binomial variables. The procedure is based on a goodness of fit argument. Tests for an unknown change point are given. The procedure is found to be appropriate to problems in which the data series has been subject to a single discrete change in binomial parameter or where there have been cumulative changes in binomial parameter, before or after an unknown point.

Keywords: Binomial parameter, Change point, Goodness of fit

1 Introduction

A sequence of independent binomial variables is subject to a change in distribution after an unknown point. Formally, we can describe this situation as follows: \( R_1, R_2, R_3, \ldots, R_k \) are independent random variables, such that, for a value \( \tau \),

\[
\begin{align*}
R_i \text{ is distributed as } & B(n_i, \theta_0) & (1 \leq i \leq \tau) \\
B(n_i, \theta_1) & (i = \tau+1, \tau+2, \ldots, k)
\end{align*}
\]

Previous work, with this type of model, has been directed toward (i) estimating the change-point, \( \tau \) and (ii) on testing the hypothesis that no change in distribution has occurred.

Most analytical approaches, developed for dealing with binomial change-point data, assume the \( \theta_1 \) and \( \theta_2 \) parameters, like \( \tau \), to be unknown. Particular attention has been devoted to the case of the \( R_i \) being (Bernouilli) zero-one variables i.e. with \( n_i = 1 \) for all \( i \).

The problem has been analysed from a variety of perspectives: Hinkley and Hinkley's [1970] likelihood work has been extended by Pettitt [1980] and Worsley [1983] - the latter authors also considering alternative CUSUM-based procedures (see, as well, Page [1955]). In contrast, Smith [1975] and Pettitt [1979] offer, respectively, Bayesian and non-parametric methodologies. More general techniques, such as those of Worsley [1986], and Kander and Zacks [1966] provide further scope for analysis.

The paper introduces a new procedure for analysing binomial change point data. The procedure is more general than many of the techniques developed in this area: it is not only capable of monitoring problems involving a single change in parameter level but also those where the change in parameter level has been cumulative after some unknown point. The technique, based on an unusual 'goodness of fit' argument is compared with the maximum likelihood estimation approach of Hinkley and Hinkley [1970]. Finally, the technique is illustrated on a number of relevant data sets mostly from the literature.

2 Analysis

Referring to model (1), we distinguish between the rival sets of assumptions:
H_1 : \theta_0 \text{ and } \theta_1 \text{ are fixed with } \theta_0 \neq \theta_1 1

H'_1 : \theta_0 \text{ is fixed and } \theta_1 \text{ is a linear function of pre-chosen scores} 2

s_i \text{ which we write as } \theta_1 = \theta_1(s_i) = \theta_0 + \beta(s_i - \bar{s}) 3

where \beta \text{ is fixed and } \bar{s} = \sum_{i=1}^{k} n_i s_i / n 4

The choice of scores s_i can be somewhat arbitrary. Options that have been considered [Williams, 1989] include in the case of assay experiments, the dose or logarithm of the dose being tested. Alternatively, s_i may be taken to equate with its index i or defined for example as

s_i = n_i + ... + n_{i-1} + .5(n_i + 1)

(the latter choice providing a measure of rank correlation between doses and binary responses.)

Under the null hypothesis H_0 we assume no change in \theta_0 5 has taken place and can therefore write

H_0: \tau = k \text{ or } H_0: \theta_1 = \theta_0 6

Operationally, we assume the data, for which model (1) is being considered, arises from an experiment involving the comparison of k groups. In the i'th group, there are n_i independent binary responses, comprising r_i 'successes' and n_i - r_i 'failures' (i = 1,2,... k).

Under H_0, the probability of a successful experimental trial is \theta_0 7.

Correspondingly, the probability of a trial resulting in failure is 1 - \theta_0 8.

The r_i can be regarded as realisations of the random variables R_i, described by model (1). These data are conveniently arranged in the form of a 2 x k contingency table [Cochran, 1954] as follows:

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Success</td>
<td>r_1</td>
<td>r_2</td>
</tr>
<tr>
<td>Failure</td>
<td>n_1-r_1</td>
<td>n_2-r_2</td>
</tr>
<tr>
<td>Total Trials</td>
<td>n_1</td>
<td>n_2</td>
</tr>
</tbody>
</table>

Table 1. Contingency table formulation

where \( R = \sum_{i=1}^{k} r_i \) and \( N = \sum_{i=1}^{k} n_i \)

13

Further, we write \( p_i = r_i/n_i \) (i = 1,2,3 ... k)
When either H₁ or H' holds, it is often found from a plot of pᵢ against i - particularly in the case where not all nᵢ = 1 - that there appears to be a linear association between the two variables. Such an association would normally be investigated using the procedures, for example, of Cochran [1954] and Armitage [1955]. The Cochran-Armitage statistic [Williams, 1988] provides a measure of the strength of this apparent relationship. An alternative, which has received much less attention, is the R square ratio, computed directly from Cochran's [1954] analysis of variance summary.

As we now show, the latter R square statistic can be adapted to suit the specific circumstances of the change-point problem:

\[
R_i^2 = \frac{(S_i / v_{1i} + S_{2i} / v_{2i})}{(T_{1i} / v_{1i} + T_{2i} / v_{2i})} \quad \text{(14) for } t = 2,3, \ldots, k-2
\]

where \( S_{1i} \) and \( S_{2i} \) correspond with the sums of squares from the regression of the \( p_i \) ratios on their index \( i \) for the first \( t \) and last \( (k-t) \) observations respectively and \( T_{1i} \) and \( T_{2i} \) are the corresponding corrected sums of squares on \( p_i \). The quantities \( v_{1i} \) and \( v_{2i} \), given respectively by :

\[
v_{1i} = N_{1i} p_{1i} q_{1i} \left/ \left( N_{1i} - 1 \right) \right., \quad v_{2i} = N_{2i} p_{2i} q_{2i} \left/ \left( N_{2i} - 1 \right) \right.
\]

are used to convert sums of squares quantities here to corresponding \( \chi^2 \) values.

Note that

\[
N_{1i} = \sum_{i=1}^{t} n_i, \quad R_{1i} = \sum_{i=1}^{t} r_i, \quad p_{1i} = R_{1i} / N_{1i}, \quad q_{1i} = 1 - p_{1i}.
\]

Similarly,

\[
N_{2i} = \sum_{i=t+1}^{k} n_i, \quad R_{2i} = \sum_{i=t+1}^{k} r_i, \quad p_{2i} = R_{2i} / N_{2i}, \quad q_{2i} = 1 - p_{2i}.
\]

In addition, it can be shown

\[
S_{1i} = \left( \sum_{i=1}^{t} r_i (i - \tilde{i}_{1i}) \right)^2 / \left( \sum_{i=1}^{t} n_i (i - \tilde{i}_{1i})^2 \right) \quad \text{(6)19}
\]

\[
S_{2i} = \left( \sum_{i=t+1}^{k} r_i (i - \tilde{i}_{2i}) \right)^2 / \left( \sum_{i=t+1}^{k} n_i (i - \tilde{i}_{2i})^2 \right) \quad \text{(7)20}
\]

where

\[
T_{1i} = \sum_{i=1}^{t} n_i \left( p_i - \tilde{p}_{1i} \right)^2 21 \text{ and } \tilde{i}_{1i} = (\sum_{i=1}^{t} i n_i) / (\sum_{i=1}^{t} n_i)
\]

\[
T_{2i} = \sum_{i=t+1}^{k} n_i \left( p_i - \tilde{p}_{2i} \right)^2 23 \tilde{i}_{2i} = (\sum_{i=t+1}^{k} i n_i) / (\sum_{i=t+1}^{k} n_i)
\]

24

The \( R_i^2 \) 25 statistic is analogous to that used by Freeman [1986] in an analysis of normal change-point data. Straightforward application of Freeman's methodology to model (1) confirms the following estimation procedure to be appropriate:
Under hypothesis $H_1, 26$, estimate the change-point $\tau$ 27 as the value of $t$ at which $R_t^2 28$ is minimised.

Under $H_1, 29$ estimate $\tau$ 30 as the value of $t$ at which $R_t^2 31$ is maximised.

The distribution of $R_t^2 32$, which is discrete, can be determined in relation to the multivariate hypergeometric distribution for the $R_t 33$, conditioned on $R$, the number of successes across all experiments. (Note that $R$ is sufficient for $\theta_0 34$, under the no-change hypothesis [Pettitt, 1979].)

Under $H_0, 35$, the latter distribution can be shown to tend asymptotically to:

$$B(R_t^2 | l, (k - 4) / 2) = \frac{\Gamma((k - 2)/2)}{\Gamma((k - 4)/2)} (I - R_t^2)^{(k - 6)/2}$$

but this result is known only to be valid if expected frequencies $n_i p_{37}$ and $n_i (1 - p) 38$ are sufficiently large e.g. at least 5 (though this may be conservative [Copas, 1989]). In many applications, expected frequencies are often too small to allow computation of accurate critical values from the asymptotic distribution: zero-one data is an obvious case in hand. Unfortunately, the method usually adopted for overcoming this kind of technical difficulty, that of amalgamating groups, has little to recommend it in the context of a change-point analysis.

In addition, the $R_t^2 39$ variates themselves are highly correlated.

Notwithstanding this fact, a Type 1 extreme value distribution can be used as an approximation to the distribution of the maximum value of $R_t^2 40$. See Freeman [1986] for details and corresponding critical values – since confirmed using computer simulation methods by So [1998]. By default, the distribution of the minimum value of $R_t^2 41$ can also be determined.

3 Applications

3.1 Page's data

Forty observations were simulated by Page [1955] the first twenty arising from the N(5,1) distribution and the remainder from the N(6,1) distribution. The data were subsequently converted to Bernoulli observations by subtracting 5 from each normal variate and coding the resultant value as 1 if greater than zero, 0 otherwise.

The minimum value of the $R_t^2 42$ statistic for these data is 1.713E-3 which occurs for $t = 18$ is not significant.

The corresponding Cochran-Armitage statistic for the entire data set takes the value 2.471. This is a significant result under $H_{43}$ and the sign here points to an abrupt increase in probability.

In contrast to the estimated change-point of 18 found here, Page [1955] and Pettitt [1979] independently suggest the value 17 for the change-point. For a one-sided test of the hypothesis $H_0$ of no change against change, Page obtained results significant at the 1% level. Unlike our procedure, however, Page's relied
on the initial mean value 0 being known.

Pettitt's one-sided non-parametric test, which assumes $\theta$, $\delta$ unknown, indicates significance just short of the 1% level.

Taking a quite different viewpoint, Smith [1975] deduces the odds on a change having occurred, are about 100 to 1. From a table of posterior probabilities for $T$, he derives the two estimates of $T$ of 18 (posterior median) and 19.24 (posterior mean).

3.2 Lindisfarne Scribes' data

The Lindisfarne Scribes' data [Pettitt, 1979] refer to the number of occurrences of present indicative third person singular endings "-s" and "-δ 45" for different sections of Lindisfarne. It is believed different scribes used the endings "-s" and "-δ 46" in different proportions. A plot of the $R^2$ statistic against $t$ is shown in Figure 1. From this, it can be seen the $R^2$ statistic assumes its minimum value of 6.175E-4 at $t = 7$. We therefore deduce an abrupt change in binomial probability occurred after the seventh section. The p-value associated with this result is close to (and slightly greater than) 5% from a corresponding simulation analysis. Pettitt's results suggest the change occurred after the sixth section with a (conservative) significance probability of 0.25%. Note that the maximum value of $R^2$ (of 3.657E-1) occurs for $t = 5$, indicating that an incremental change in probability may have occurred beforehand. This view that the data were subject to two change-points is one shared by Smith [Pettitt, 1979] who believes changes occurred after the sixth and seventh sections.

The Cochran-Armitage statistic for these data, using the index $t$ as correlate, yields the significant value of 3.091. From the sign of the coefficient here, we deduce there was an abrupt increase in the proportion of "-s" endings after the seventh section, preceded by a cumulative increase before the fifth.

![Fig. 1. $R^2$ plot Lindisfarne Scribes data](image)

3.3 Club foot data

Worsley [1983] presents data on the number of cases of birth deformity talipes or club foot in the first
month of gestation for the years 1960 to 1976 in a region of northern New Zealand. It is believed that a change in probability occurred after 1965. Worsley's procedure showed that this was indeed the case, confirming the no-change hypothesis should be rejected at the 5% level.

The $R_t^2$ plot for these data is shown in Figure 2. The minimum value of $R_t^2 = 1.029E-2$ occurs at $t = 6$. This coincides with Worsley's estimate of the year, an abrupt change occurred.

The value of the appropriate Cochran-Armitage statistic for the full data set is $2.603$. This significant result leads to the conclusion that there was an abrupt increase in the rate of club foot incidence after 1965.

![Fig. 2. $R_t^2$ plot](image)

3.4 Simulated data

Data were simulated by the author as follows: (k =) twenty pairs of random digits were drawn from simple random number tables. These were adopted as the $n_i$ values. Assuming a change-point to hold at $T = 7$, $\theta_0 = 0.5$ was taken as

- $0.5$ for $i = 1, 2, ..., 7$
- $0.5 + 0.02i$ for $i = 8, 9, ..., 20$

Applying these probabilities in turn to each of the binary trials within each experimental group, we obtained the $r_i$ data, shown in Table 2 below:
The $R^2$ for these data is shown in Figure 3. The maximum value of $R^2_t (= 0.556)$ occurs at $t = 5$ which is our estimate of $\tau$ under model $H_1$. The $p$-value for this result from simulation < 2.5%. The associated Cochran-Armitage statistic for the set is calculated as 4.361. This is highly significant and we deduce a cumulative increase in 'success' probability took place after the fifth group.

3 Conclusions

A novel approach to the identification and testing of an unknown change point for a series of Binomial variates has been introduced. The approach has been demonstrated to have the particular advantage of distinguishing between situations where a Binomial parameter undergoes an abrupt as opposed to a
cumulative value change. Results obtained using the procedure, have been found to compare well with
established alternatives across a range of archetypal data sets. However power characteristics may
contrast less favourably – particularly in the special case of Bernouilli observations - and this is an area
for future research.

References


D.V. Hinkley and E.A. Hinkley. Inference about the change point in a sequence of binomial variables. *Biometrika* 57

Z. Kander and S. Zacks. Test procedures for possible changes in parameters of statistical distributions occurring at unknown


A.N. Pettitt. A simple cumulative sum type statistic for the change point problem with zero-one observations. *Biometrika* 67
1:79-84, 1980.


A.F.M Smith. A Bayesian approach to inference about a change point in a sequence of random variables. *Biometrika* 62 2:407-
416, 1975.


K.J. Worsley. The power of the likelihood ratio and cumulative sum tests for a change in binomial probability. *Biometrika*
70:455-64. 1983.

K.J. Worsley. Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*