한국통계학회논문집 제3권 2호
1996년 8월 pp. 175-185

Change-Point Problems in a Sequence of Binomial Variables[1)]

Kwang Mo Jeong[2)]

Abstract

For the Change-point problem in a sequence of binomial variables we consider the maximum likelihood estimator(MLE) of unknown change-point. Its asymptotic distribution is quite limited in the case of binomial variables with different number of trials at each time point. Hinkley and Hinkley (1970) gives an asymptotic distribution of the MLE for a sequence of Bernoulli random variables. To find the asymptotic distribution a numerical method such as bootstrap can be used. Another concern of our interest is the inference on the change-point and we derive confidence sets based on the likelihood ratio test(LRT). We find approximate confidence sets from the bootstrap distribution and compare the two results through an example.

1. Introduction

In this paper we are concerned with the estimation and confidence sets for the change point in a sequence of binomial variables. Let $X_1, X_2, \cdots, X_T$ be independent binomial random variables with $X_i \sim B(n_i, p_i)$. The change-point problem in binomial variables can be represented as $p_1 = p_2 = \cdots = p_\tau \neq p_{\tau+1} = \cdots = p_T$ for some unknown time point τ. The MLE and other nonparametric estimation procedures using cumulative sum type statistics are popular in the estimation of τ. Similarly LRT and cumulative sum test statistics are widely used for testing the existence of change-points.

Change-point models in a sequence of binomial variables have been treated by many authors. Hinkley and Hinkley (1970) suggested a MLE and an LRT in Bernoulli case. They derived exact and asymptotic distributions using random walks and also discussed the relative efficiency of LRT with respect to MLE. On the other hand Pettitt (1979, 1980) proposed a nonparametric estmation procedure maximizing a cumulative sum statistic which is equivalent to the Kolmogorov-Smirnov type statistic in Bernoulli case. The power of LRT and

1) The present studies were supported (in part) by the Matching Fund Programs of Research Institute for Basic Sciences, Pusan National University, Korea, 1994, Project No. RIBS-PNU-94-102.

2) Professor, Department of Statistics, Research Institute of Information and Communication, Pusan National University, 609-735.

cumulative sum test was compared by Worsley (1983). As we have shown in the references above the change-point problem of binomial variables is limited only to the Bernoulli case. In a sequence of exponential family random variables Worsley (1986) discussed about confidence regions and tests. He also obtained critical values of LRT in particular in the exponentially distributed random variables. But it is intractable for general distributions and we need an alternative method such as bootstrap to solve this problem.

Confidence set or region was not so well treated until now for binomial variables. The asymptotic distribution of $\hat{\tau}$ can be used to find confidence sets for Bernoulli case but it is not possible in a general binomial case where its asymptotic distribution has not been known. As commented by Hinkley and Hinkley (1970) the distribution of MLE is not degenerate and hence it is not consistent. This unhappy state of affairs is a characteristic of chage-point problems. We follow the procedure of Worsley (1986) to find a confidence set for change-point in a sequence of binomial variables. We first consider LRT for testing

$$H_0^{(k)}: \tau = k \text{ against } H_1^{(k)}: \tau \neq k. \tag{1.1}$$

A confidence set can be found by reversing the testing procedure and hence we choose the time points such that the null hypothesis is not rejected. We are confronted with the problem of determining critical point of LRT. For general discussions on the confidence sets using LRT we refer to Siegmund (1988). As a numerical technique we use the bootstrap method to find the critical points.

Two types of confidence set, one using $\hat{\tau}$ directly and the other using LRT procedure, can be compared. It is known that MLE is not sufficient so that the inference based on $\hat{\tau}$ is not efficient. It seems to be reasonable to consider any inference based on LRT. According to Hinkley and Hinkley (1970) the power of LRT is superior to $\hat{\tau}$ and they also give power comparisons through Monte Carlo simulations. In section 2 we consider MLE and its asymptotic distributions as discussed by Hinkley (1970), Hinkley and Hinkley (1970) in a sequence of normal and Bernoulli random variables, respectively. The likelihood based confidence set will be discussed in Section 3 and we also give the parametric bootstrap algorithm to find critical points of LRT. The proposed method is applied to a real data set. Finally we summarize the results and also comment on further research topic.

2. Maximum Likelihood Estimation and Asymptotic Distribution

Let p_0 and q_0 be the common binomial probability before and after the unknown change-point τ, respectively. Then change-point model can be written as $p_1 = p_2 = \cdots = p_\tau = p_0$ and $p_{\tau+1} = \cdots = p_T = q_0$. Let $\theta = (\tau, p_0, q_0)$ then log-likelihood

function $L(\theta)$ is given by

$$\begin{aligned} L(\theta) &= \sum^{\tau} x_i \log\left(\frac{p_0}{1-p_0}\right) + \sum^{\tau} n_i \log(1-p_0) + \sum_{\tau+1}^{T} x_i \log\left(\frac{q_0}{1-q_0}\right) + \sum_{\tau+1}^{T} n_i \log(1-q_0) \\ &= \sum^{\tau} x_i \left\{\log\left(\frac{p_0}{1-p_0}\right) - \log\left(\frac{q_0}{1-q_0}\right)\right\} + \sum^{\tau} n_i \left\{\log(1-p_0) - \log(1-q_0)\right\} \\ &\quad + \left\{\sum^{T} x_i \log\left(\frac{q_0}{1-q_0}\right) + \sum^{T} n_i \log(1-q_0)\right\} . \end{aligned}$$

We note that $L(\theta)$ can be written simply as

$$L(\theta) = S_\tau \Delta + N_\tau \log\left(\frac{1-p_0}{1-q_0}\right) + S \log\left(\frac{q_0}{1-q_0}\right) + N \log(1-q_0) , \tag{2.1}$$

where

$$S_\tau = \sum^{\tau} X_i, \quad S = \sum^{T} X_i, \qquad N_\tau = \sum^{\tau} n_i, \quad N = \sum^{T} n_i$$

and $\Delta = p_0^* - q_0^*$ with $p_0^* = \log\left(\frac{p_0}{1-p_0}\right)$, $q_0^* = \log\left(\frac{q_0}{1-q_0}\right)$.

Note that when S_τ and S are given the log-likelihood is constant and hence independent of nuisance parameter Δ. Further we note that conditionally on S, S_τ is sufficient for Δ when τ is fixed in which case exact inference about Δ is possible if conditioned on S. Because τ is unknown τ itself becomes an extra parameter of interest and the conditional inference is not so straightforward.

Here we temporarily assume p_0 and q_0 are known. Then the MLE of τ is found to maximize the $L(\tau)$, that is,

$$\hat{\tau} = \arg\max \{L(k) \mid k = 1, 2, \cdots, T-1\} .$$

Maximizing $L(\tau)$ is reduced to maximizing the first two terms in the equation (2.1), i.e.,

$$Y_\tau = S_\tau \Delta + N_\tau \log\left(\frac{1-p_0}{1-q_0}\right) .$$

We note that $\{Y_t\}$ defines two random walks W and W' defined by

$$W = (0, Y_{\tau-1} - Y_\tau, \cdots, Y_1 - Y_\tau), \quad W' = (0, Y_{\tau+1} - Y_\tau, \cdots, Y_{T-1} - Y_\tau) .$$

Let Z_i be defined by

$$\begin{aligned} Z_i &= (Y_{\tau-i}-Y_{\tau-i+1})\,\Delta^{-1} \\ &= -x_{\tau-i+1}-n_{\tau-i+1}\log(\frac{1-p_0}{1-q_0})\,\Delta^{-1} \\ &= -x_{\tau-i+1}-n_{\tau-i+1}\Theta \quad , \end{aligned}$$

where $\Theta=\log(\frac{1-p_0}{1-q_0})\Delta^{-1}$. Similarly

$$Z_i'=(Y_{\tau+i}-Y_{\tau+i-1})\,\Delta^{-1}=x_{\tau+i}+n_{\tau+i}\Theta.$$

Note that W and W' represent the log-likelihood function of t relative to the true value τ for $t\leq\tau$ and $t>\tau$, respectively. Apart from the factor Δ random walks W and W' can be written in terms of Z_i and Z_i' as

$$W=(0,Z_1,Z_1+Z_2,\cdots,\sum^{\tau-1}Z_i),\quad W'=(0,Z_1',Z_1'+Z_2',\cdots,\sum^{T-\tau-1}Z_i').$$

In the following few paragraphs we review the general theory of Hinkley (1970) for the asymptotic distribution of $\hat{\tau}$. The random walks W and W' are independent and all properties of the likelihood are expressed in terms of W and W'. Let M and M' be the respective maxima of W and W'. The events involving $\hat{\tau}$ can be expressed in terms of events involving M and M'. The MLE $\hat{\tau}$ corresponds to the position of the larger of the two random walk maxima. If each maximum is zero then $\hat{\tau}=\tau$, and the event $\hat{\tau}=\tau+k$ is equivalent to $M'=Y_1'+\cdots+Y_k'>0$ and $M'>M$. Similar result holds for the event $\hat{\tau}=\tau-k$.

Let I and I' be the indices of the maxima of W and W' defined respectively by

$$I=\inf\{k\mid M=\sum^{k}Y_i\},\quad I=0 \text{ if } M=0$$

and

$$I'=\inf\{k\mid M'=\sum^{k}Y_i'\},\quad I'=0 \text{ if } M'=0.$$

Then we can express the asymptotic distribution of $\hat{\tau}$ as follows.

$$\begin{aligned} P(\hat{\tau}=\tau) &= P(M=0)P(M'=0) \\ P(\hat{\tau}=\tau+k) &= P(I'=k, M'>M, M'>0), \quad k=0,1,2,\cdots, \\ P(\hat{\tau}=\tau-k) &= P(I=k, M>M', M>0), \quad k=0,1,2,\cdots. \end{aligned} \tag{2.2}$$

Hinkley (1970) obtained the approximate distribution of $\hat{\tau}$ in normal case. In the sequence of Bernoulli variables Hinkley and Hinkley (1970) also obtained the asymptotic distribution of $\hat{\tau}$ but it is not tractable in the general binomial case. So as an alternative to find the asymptotic distribution in (2.2) we propose a bootstrap method. There are many authors using bootstrap in change-point problems. Hinkley and Schechtman (1987) used bootstrap to find conditional distribution of $\hat{\tau}$ given some ancillary statistics. This conditional approach stems from the work of Cobb (1978) who suggested a conditional solution given ancillary statistics in change-point problem of normal variables. On the other hand Dumbgen (1991), Boukai (1993), and Antoch and Huskova (1995) suggested bootstrap in nonparametric procedures. We assumed until now p_0 and q_0 are known and in the unknown case we substitute the MLE $\hat{p}_0$ and $\hat{q}_0$.

We briefly review a parametric bootstrap procedure. Let f_η belong to a parametric family of probability density functions indexed by unknown parameter η. The observed data $\aleph=(X_1,\cdots,X_T)$ is a sample from f_η and let $\theta=g(\eta)$ be a real-valued parameter. When $\hat{\eta}$ is an MLE of η the MLE of θ is $\hat{\theta}=g(\hat{\eta})$. Given $\aleph=(X_1,\cdots,X_T)$ we generate a bootstrap sample from $f_{\hat{\eta}}$ denoted by $\aleph^*=(X_1^*,\cdots,X_T^*)$, which is the so called parametric bootstrap sample. A bootstrap version of $\hat{\eta}$, denoted as $\hat{\eta}^*$, is calculated from the bootstrap sample $\aleph^*=(X_1^*,\cdots,X_T^*)$. The distribution of $\hat{\theta}^*$ is defined by $\hat{G}(s)=P_{\hat{\eta}}(\hat{\theta}^*\le s)$. We give a parametric bootstrap algorithm to find the asymptotic distribution.

Bootstrap Algorithm for the Asymptotic Distribution of MLE

Step 1: Find MLE $\hat{\tau}$ from the given binomial data

Step 2: Estimate the MLE $\hat{p}_0$ and $\hat{q}_0$ using the given observations before and after $\hat{\tau}$, respectively.

Step 3: Generate a bootstrap sample with the fixed sizes of trials n_i and the MLEs $\hat{p}_0$ and $\hat{q}_0$.

Step 4: Find a bootstrap version of MLE $\hat{\tau}$, denoted as $\hat{\tau}^*$, from the bootstrap sample.

Step 5: Repeat Step 3 and Step 4 many times(B).

3. Likelihood Ratio Test and Confidence Set

As before p_0 and q_0 are assumed to be known and τ is the only parameter of interest. The LRT for testing the hypotheses (1.1) is defined by

$$LR^{(k)} = \max L(t) - L(k) \tag{3.1}$$

and we reject $H_0^{(k)}$ when $LR^{(k)}$ is large. The level α critical point $c_\alpha^{(k)}$ is determined to satisfy the equation

$$P\{LR^{(k)} \le c_\alpha^{(k)}\} = 1-\alpha .$$

By reversing the LRT critical region we find a $100(1-\alpha)\%$ confidence set in which $H_0^{(k)}$ is not rejected. The confidence sets directly based on the MLE are demonstrably inferior to those obtained by, for example, that of LRT as commented by Hinkley (1970) and Siegmund (1988).

Given S_k and S, $L(k)$ is constant and the problem is reduced to find a constant $c_\alpha^{'(k)}$ satisfying

$$P\{\max L(k) \le c_\alpha^{'(k)} \mid S_k, S\} = 1-\alpha .$$

The confidence set is most easily determined as the set of k for which

$$P(\max L(t) \le \max L(t)_{\text{obs}} \mid S_k, S) \le 1-\alpha \tag{3.2}$$

where $\max L(t)_{obs}$ denotes the maximum of $L(t)$ for the given data. Equivalently we approximate the probability $P(\max L(t) > \max L(t)_{\text{obs}} \mid S_k, S)$, which is called the observed level at k, and we determine a $100(1-\alpha)\%$ confidence set consisting of all k with observed levels greater than α.

The approximations to (3.2) was done by Smith (1975), and Raferty and Akman (1986) in the Bayesian sense. Some numerical computaion is required in this case. According to the general discussion of Siegmund (1988) for the exponential family random variables the approximation is of the form

$$P(\max L(t) \geq a \mid S_k,\ S) \sim \nu^* \exp\{-(a-L(k))\}\ , \qquad (3.3)$$

where $a-L(k)$ is assumed small compared to k, and ν^* is a distribution dependent quantity. Here we require the conditional distribution of $\max L(t)$ given S_k and S. Worsley (1986) and Siegmund (1988) obtained the approximation (3.3) for the normal and exponentially distributed random variables but it has not been studied in detail for other distributions. In this paper we suggest a bootstrap method to find confidence sets of τ. A bootstrap algorithm finding confidence set can be written as follows.

Bootstrap Algorithm for Confidence Sets

Step 1: For each k we generate a bootstrap sample from the given binomial sample, where the unknown p_0 and q_0 are substituted by the MLEs $\hat{p}_0$ and $\hat{q}_0$, respectively.

Step 2: Compute the bootstrap version of $LR^{(k)}$ defined in (3.1), denoted as $LR^{(k)^*}$ using the generated bootstrap sample.

Step 3: By repeating Step 1 and Step 2 many times(B) we determine the observed levels using the bootstrap distribution of $LR(k)^{*(j)}$, $j=1, 2, \cdots, B$.

Step 4: Select every point $k \in \{1, 2, \cdots, T-1\}$ whose observed level is greater than α. This set is a $100(1-\alpha)\%$ confidence region of τ.

It is well-known that the likelihood of change-point problem is not smooth and so that the asymptotic distribution theory of LRT does not hold. For the regular model with no change-point Cox (1987) comments on confidence regions based on LRT statistic with chi-square approximation as an alternative to bootstrap. Let $\theta=(\psi, \lambda)$, where ψ is a parameter of interest and λ is a nuisance parameter. The $100(1-\alpha)\%$ confidence set of ψ is

$$\{\psi \mid 2\{\max L(u, \hat{\lambda}) - L(\psi, \hat{\lambda}\} \leq \chi^2_{d,\alpha}\}$$

with $\hat{\lambda}_\psi$ denoting the MLE of λ given ψ, where $\chi^2_{d,\alpha}$ is the upper α-percentile point of chisquare distribution with degree of freedom d with d denoting the dimension of ψ. By using the bootstrap distribution of $\hat{\tau}$ we can also derive confidence sets of τ. In this case the percentage point c_α is defined by the largest integer satisfying

$$P(\hat{\tau}-\tau \leq c_\alpha) \leq \alpha \quad (\alpha=0.01, 0.05, 0.10),\ P(\hat{\tau}-\tau < c_\alpha) < \alpha \quad (\alpha=0.90, 0.95, 0.99).$$

For general discussions about bootstrap confidence set we refer to Efron (1985), and also for

the confidence sets based on LRT, see Siegmund (1988).

Example: The Lindisfarne Scribes' data which was given originally by Ross (1950) and obtained from Pettitt (1979). The data refer to the number of occurrences of present indicative third person singular endings "-s" and "- ∂" for different section of Lindisfarne. It is believed different scribes used the endings "-s" and "- ∂" in different proportions. The data are given in Table 3.1.

Table 3.1: Lindisfarne Scribes' data

Section(i)	1	2	3	4	5	6	7	8	9
No. of "-s"	12	26	31	17	7	28	34	10	29
No. of "- ∂"	9	10	13	4	2	24	11	1	8
Section(i)	10	11	12	13	14	15	16	17	18
No. of "-s"	30	16	17	24	14	5	17	17	16
No. of "- ∂"	9	2	0	7	2	1	3	4	4

The plotting of likelihoof funcion is given in Figure 3.1

For this data the MLE of change-point is $\hat{\tau}=6$ and Table 3.2 gives a bootstrap distribution of $\hat{\tau}$ with B=5,000 repetitions.

Table 3.2: Bootstrap Distribution of $\hat{\tau}$ (B=5,000)

$\hat{\tau}^* - \hat{\tau}_{obs}$	-5	-4	-3	-2	-1	0	1	2	3	4	5	≥6
Probability	.0066	.0114	.0302	.0372	.0606	.6790	.0608	.0436	.0212	.0132	.0080	.0080

From Table 3.2 we may determine a equal-tail confidence interval of τ. For example, 90% interval is [3, 9] and 95% confidence interval is [2, 10]. Table 3.3 gives the observed significance levels of LRT. From this table 90% confidence set is {6, 7, 8, 9, 10} which also becomes the 95% confidence set of τ. On the other hand 99% confidence set is {6, 7, 8, 9, 10, 11}. We note that a confidence set using LRT is shorter than equal-tail interval based on MLE. This fact coincides with the fact that the inference based on LRT is more efficient than that of MLE.

Table 3.3: Observed Levels of LRT

k	level(k)	k	level(k)
1	0.0041	10	0.1522
2	0.0081	11	0.0313
3	0.0023	12	0.0022
4	0.0062	13	0.0052
5	0.0012	14	0.0014
6	1.0000	15	0.0027
7	0.4061	16	0.0013
8	0.1532	17	0.0019
9	0.1058		

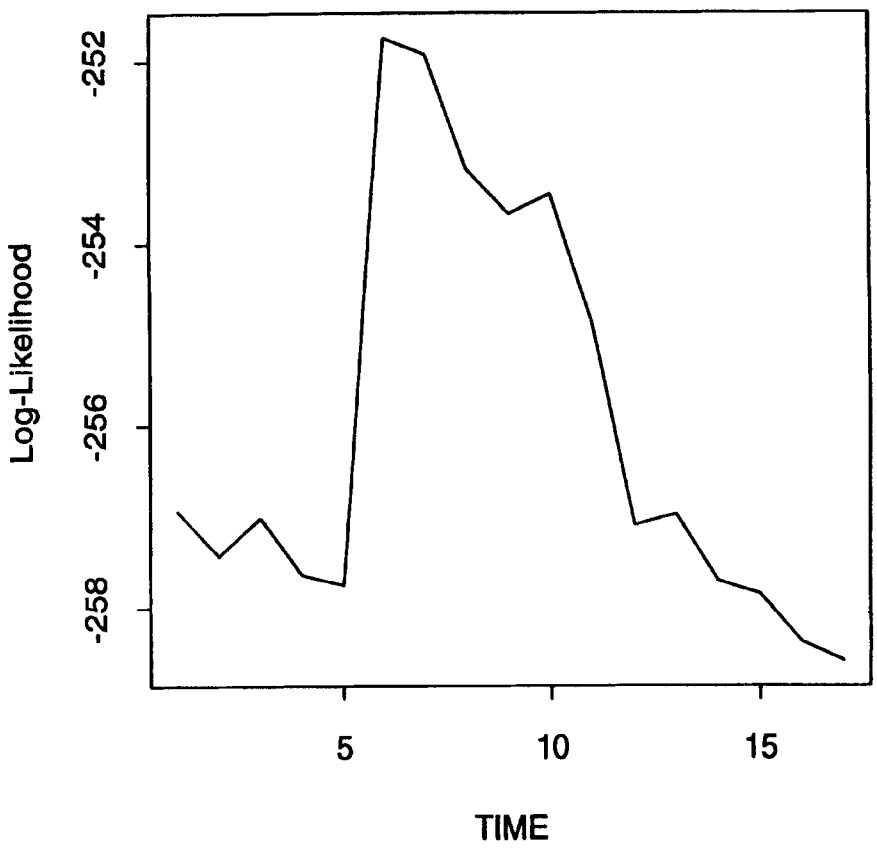

Figure 3.1: Log-Likelihood Function of Lindisfarne Scribe Data

4. Summary and Discussion

In a sequence of binomial variables we are concerned with the constancy of 'success' probability p_i over all time period. We first consider an MLE to estimate change-point τ and discuss general random walks representations to find the asymptotic distribution of $\hat{\tau}$. Except for normal and Bernoulli cases the asymptotic distribution is not plausible. A bootstrap method is suggested to approximate the asymptotic distribution of MLE. Another interest of this paper is to find confidence sets of change-point. Standard asymptotic distribution theory of LRT does not hold because the likelihood of change-point model is not smooth. Similarly we may use the bootstrap distribution of LRT in determing confidence sets. Two types of confidence sets, one using MLE and the other LRT, are considered and explained through an example.

We have not discussed about the coverage error of bootstrap confidence set and remain it as a further research topic.

References

[1] Antoch, J. and Huskova, M (1995). Change-Point Problem and Bootstrap, *Nonparametric Statistics*, Vol. 5, 123-144]

[2] Boukai, B. (1993). A Nonparametric Bootstrapped Estimate of the Change-Point, *Nonparametric Statistics*, Vol. 3, 123-134

[3] Cobb, G. W. (1978). The Problem of the Nile: Conditional Solution to a Change-Point Problem, *Biometrika*, Vol. 65, 243-251

[4] Cox, D. R. (1987), Comment on "Better Bootstrap Confidence Intervals" by Efron, B., *Journal of the American Statistical Association*, Vol. 82, 190

[5] Dumbgen, L. (1991) The Asymptotic behaviour of some nonparametric change-Point Estimators, *The Annals of statistics*, Vol. 19, 1471-1495

[6] Efron, B. (1985), Bootstrap Confidence Intervals for a Class of Parametric Problems, *Biometrika*, Vol. 72, 45-58

[7] Hinkley, D. V. (1970), Inference About the Change-Point in a Sequence of random variables, *Biometrika*, Vol. 57, 1-16

[8] Hinkley, D .V. and Hinkley, E. A. (1970), Inference about the Change-Point in a Sequence of Binomial Variables, *Biometrika*, Vol. 57, 477-488

[9] Hinkley, D. V. and Schechtman, E. (1987), Conditional Bootstrap Methods in the Mean-Shift Model, *Biometrika*, Vol. 74, 85-93

[10] Pettitt, A. N. (1979), A Nonparametric Approach to the Change-Point Problem, *Applied statistics*, Vol. 28, 126-135

[11] Pettitt, A. N. (1980), A Simple Cumulative Sum Type statistic for the Change-Point Problem with Zero-One Observations, *Biometrika*, Vol. 67, 79-84

[12] Raferty, A. E. and Akman, V. E. (1986), Bayesian Analysis of a Poisson Process with a Change-Point, *Biometrika*, Vol. 73, 85-90

[13] Ross, A. S. C. (1950), Philological Probability Problems, *Journal of the Royal Statistical Society, B*, 20, 93-101

[14] Siegmund, D. (1988), Confidence Sets in Change-Point Problems, *International Statistical Review*, Vol. 56, 31-48

[15] Smith, A. F. M. (1975), A Bayesian Approach to Inference about a Change-Point in a Sequence of random variables, *Biometrika*, Vol. 62, 407-416

[16] Worsley, K. J. (1983), The Power of Likelihood ratio and Cumulative Sum Tests for a Change in a Binomial Probability, *Biometrika*, Vol. 70, 455-464

[17] Worsley, K. J. (1986), Confidence regions and Tests for a Change-Point in a Sequence of Exponential Family Random Variables, *Biometrika*, Vol. 73, 91-104