Journal of the Korean Statistical Society (2001), **30**: *1, pp 1–20*

Bootstrap Inference on the Poisson Rates for Grouped Data

Kee-Won Lee[1] and Woo-Chul Kim[2]

ABSTRACT

We present how bootstrap methods can be used to conduct inference on the rates of Poisson distributions when only the grouped data are available. A theoretical justification for the validity of bootstrap is given with an illustration of proposed method using a data set obtained from a pathology laboratory test. Traditional asymptotic methods are compared with bootstrap methods in computing the estimated standard errors and achieved significance levels for one sample and two sample tests. Bootstrap methods are shown to possess a nice property that the small sample distribution of the relevant statistic can be readily obtained from the bootstrap copies.

Keywords: maximum likelihood method; minimum chi-square method; parametric bootstrap; achieved significance level

1. Introduction

Haitovsky (1982-1988) states that there are four types of grouped data from the reason why the grouping is needed. The first type is from the purely descriptive reason to summarize the data in a condensed form. The second reason for grouping data arises in studies where the data sets are obtained from confidential sources where it is necessary to maintain the privacy of the individual's record. The third reason arises when we deal with a large data sets, where grouping is desirable to reduce the cost of data handling. The fourth reason comes from incomplete measurement.

Now, we present another type of grouped data where the reason for grouping is mainly due to practitioner's own tradition to simplify recording. We observe a number of occurrences of a random phenomenon which we believe follows a

[1]Department of Statistics, Hallym University, Chunchon, 200-702, S. Korea.
[2]Department of Statistics, Seoul National University, Seoul, 151-742, S. Korea.

Poisson distribution. The number of occurrences are grouped into a few ordered categories, and only the frequencies in each categories instead of the original countings are recorded. This practice often occurs in the analysis of laboratory test data, where one seeks a relationship between the number of occurrences of a certain positive cell type and the status of the corresponding disease in experimental pathology.

We show how bootstrap can be used to conduct estimation and hypothesis testing for this unique type of gruped data. A formal theoretical justification is given to show that the bootstrap approximation is valid, that is, at least as good as traditional asymptotic approach.

TABLE 1.1 gives a typical data set of this type. The data set was obtained from Lee *et al.* (1989), where they studied the relationship between the occurrences of positive Langerhans cells and the neoplastic transformation of the uterine cervix. The cells were turned up through S-100 protein on the sample tissues sent out from the Department of Pathology in the College of Medicine at Hallym University.

After the positive Langerhans cells were identified, the results were classified and recorded in the following manner. If there were no positive Langerhans cells, then the case was categorized as "$-$". If there was only 1 positive Langerhans cell, then the case was categorized as "$+$". If there were 2, 3 or 4 positive Langerhans cells, the case was categorized as "$++$". Finally, if there were 5 or more positive Langerhans cells, the case was categorized as "$+++$".

TABLE 1.1: Data Set

Groups	"$-$"	"$+$"	"$++$"	"$+++$"
Chronic Cervicitis	8	8	0	0
Dysplasia	0	6	7	0
Carcinoma *in situ*	0	2	6	6
Invasive Carcinoma	0	1	3	6

In the data set, Chronic Cervicitis patient group plays the role of control. Main purpose of the study was to find out whether there is a tendency of increase in the rate of occurrences as the status of cancer aggravates. Statistical issues relevant with the problem will be considered in what follows.

In section 2, maximum likelihood and minimum chi-square methods are used

to estimate the unknown rate of occurrences. Both are known to give asymptotically equivalent estimated rates. The standard error of the estimated rate is computed using both asymptotic and bootstrap methods. The example confirms that the bootstrap offers estimated standard errors close to those obtained from the asymptotic method. In this problem, the bootstrap method has a definite advantage over the asymptotic method in that the derivatives of each cell probabilities are not necessary to estimate standard errors.

In section 3, hypothesis testing procedures are developed for one and two sample cases. Asymptotic and bootstrap methods are used to compute the achieved significance levels for various test statistics. Again, the bootstrap method has an advantage over the asymptotic method for the same reason as above when we use the naive observed difference as the test statistic. Besides, the sampling distribution of the test statistic can be obtained from the bootstrap copies. We draw the plots of sampling distributions for each test statistic.

2. Estimation of the Rates

2.1. *Methods of Estimation*

Suppose that $X_1, \ldots, X_n$ are independent and identically distributed random sample from a Poisson distribution with rate of occurrences λ. Upon observation, the numbers of occurrences are classified into a few, say K, ordered categories in a well-defined manner as described in the previous section. Let $O_1, \ldots, O_K$ be the observed frequencies of each categories, and let $\pi_1, \ldots, \pi_K$ be the corresponding cell probabilities of categories. Then $E_k = n \times \pi_k$ for $k = 1, \ldots, K$ are the expected frequencies of each categories. We use the notations O, π_λ and E_λ for simplicity.

Either the maximum likelihood method or minimum chi-square method can be used to estimate λ for this problem. A Pearson type minimum chi-square estimator can be obtained as follows:

$$\hat{\lambda}_P = \arg\min_{\lambda>0} \sum \frac{(O - E_\lambda)^2}{E_\lambda}, \tag{2.1}$$

where the sum is taken over all the categories. The maximum likelihood estimator is essentially equivalent to the minimizer of Kullback-Leibler measure, and can be obtained as follows:

$$\hat{\lambda}_{K-L} = \arg\min_{\lambda>0} \left\{ -2 \sum O \log(O/E_\lambda) \right\}. \tag{2.2}$$

In the actual computation, we would recommend a search method rather than Newton-Raphson type recursion, which need first and second derivatives of π_λ. We started searching from 0.01, and increased the steps by 0.01 until we hit the point which gives the minimum value in equation (2.1) or in equation (2.2). The fact that the quantity we want to minimize tends to a convex function as the sample size increases was crucial to find the minimizer. It turned out that the search method actually gave an answer faster than Newton-Raphson recursion.

2.2. *Standard Error of Estimated Rate*

Once the estimator of λ is obtained, we need to know the likely size of random variation. Either traditional asymptotic methods or computer intensive bootstrap methods can be used to approximate the standard error of estimated rate.

A. Asymptotic Approach

The estimated rates obtained from equations (2.1) and (2.2) have the following asymptotic properties. See Rao (1957) for the case of minimum chi-square.

Proposition 1. Let π'_λ be the derivative of π_λ with respect to λ. Then, the sampling distribution of $\sqrt{n}(\hat{\lambda} - \lambda)$ tends to a normal distribution with mean 0 and variance $(\sum \pi'^2_\lambda/\pi_\lambda)^{-1}$, and $(\sum \pi'^2_{\hat{\lambda}}/\pi_{\hat{\lambda}})^{-1}$ tends to $(\sum \pi'^2_\lambda/\pi_\lambda)^{-1}$ weakly as the sample size increases.

Therefore, the standard error of the estimated rate can be approximated as

$$\mathrm{SE}(\hat{\lambda}) \approx \sqrt{(\sum \pi'^2_\lambda/\pi_\lambda)^{-1}/n}. \tag{2.3}$$

The quantity $(\sum \pi'^2_\lambda/\pi_\lambda)^{-1}$ will be denoted by $\sigma^2(\lambda)$ for various choices of λ throughout the paper. Further simplified notations without parentheses, such as σ^2 and $\hat{\sigma}^2$, will be used in place of $\sigma^2(\lambda)$ and $\sigma^2(\hat{\lambda})$ respectively unless there is any chance of misunderstanding.

An estimated standard error of $\hat{\lambda}$ based on an asymptotic method can be obtained by plugging in $\hat{\lambda}$ in place of λ in the formula (2.3), that is,

$$\widehat{\mathrm{SE}}_A(\hat{\lambda}) = \hat{\sigma}/\sqrt{n}. \tag{2.4}$$

B. Bootstrap Approach

The bootstrap, initiated by Efron (1979), can also be used to approximate the standard error of the estimated rate. A Monte Carlo approximation to the bootstrap estimated standard error of $\hat{\lambda}$ can be obtained as follows.

Step 1. Choose a random sample of the same size from the fitted Poisson distribution with rate $\hat{\lambda}$.

Step 2. Mark the sample values according to the given rule, and set up the bootstrap observed marking frequencies, say O^*.

Step 3. Compute the bootstrap minimum chi-square estimator and the bootstrap maximum likelihood estimator. That is,

$$\hat{\lambda}^*_P = \arg\min_{\lambda>0} \sum \frac{(O^* - E_\lambda)^2}{E_\lambda}, \tag{2.5}$$

for minimum chi-square method, and

$$\hat{\lambda}^*_{K-L} = \arg\min_{\lambda>0} \left\{ -2 \sum O^* \log(O^*/E_\lambda) \right\}, \tag{2.6}$$

for maximum likelihood method.

At this stage, we can check that the following propostion holds from Proposition 1 and triangular array convergence theory.

Proposition 2. The sampling distribution of $\sqrt{n}(\hat{\lambda}^* - \hat{\lambda})$ tends to a normal distribution with mean 0 and variance σ^2, and that $\sigma^2(\hat{\lambda}^*)$ tends to σ^2 weakly as the sample size increases, for almost all sample paths.

Therefore we can be sure that the bootstrap estimated standard error is at least as good as the one obtained from the asymptotic method. Even better, we can completely avoid computing the derivative of π_λ through Monte Carlo approximation.

Step 4. Repeat *Step 1*–*Step 3* for a sufficiently large number of times, say B times, to obtain $\hat{\lambda}^*_1, \ldots, \hat{\lambda}^*_B$.

Step 5. A Monte Carlo bootstrap estimated standard error of $\hat{\lambda}$ is given by

$$\widehat{\mathrm{SE}}_B(\hat{\lambda}) = \sqrt{\sum_{b=1}^{B} (\hat{\lambda}^*_b - \overline{\hat{\lambda}^*})^2/(B-1)}, \tag{2.7}$$

where $\overline{\hat{\lambda}^*} = \sum_{b=1}^{B} \hat{\lambda}^*_b/B$.

For this problem, the Monte Carlo bootstrap method has a definite advantage over the asymptotic method in that we do not need derivative of π_λ to obtain the estimated standard errors, and in that we can approximate the sampling distribution of estimated rates. We can also approximate the sampling distribution of chi-square statistic computed during the estimation procedure.

2.3. *Summary of Estimated Results*

TABLE 2.1 summarizes the values of estimated rates and their estimated standard errors. Each patient group is given sequential numbers in order to simplify the notation and make them easy to identify. Standard errors calculated from asymptotic approach in the equation (2.4) are denoted by SE_A, while the Monte Carlo bootstrap standard errors obtained from (2.7) are denoted by $\mathrm{SE}_{B,nboot}$, where *nboot* is the number of bootstrap replications. In our case *nboot* is set at 50, 100, and 200. These numbers were chosen according to the suggestions given in Efron and Tibshirani (1993).

TABLE 2.1: Estimated Rates with SE's and Goodness of Fit Measure (denoted by GOF) Computed from the Asymptotic Method

Method	Estimated Rate	SE_A	$\mathrm{SE}_{B,50}$	$\mathrm{SE}_{B,100}$	$\mathrm{SE}_{B,200}$	GOF
Min χ^2	$\hat{\lambda}_1 = 0.62$	.20	.21	.19	.22	3.43
	$\hat{\lambda}_2 = 1.73$	.40	.36	.41	.37	3.81
	$\hat{\lambda}_3 = 3.95$	.61	.64	.61	.64	1.60
	$\hat{\lambda}_4 = 4.73$	.82	.81	.85	.87	1.49
ML	$\hat{\lambda}_1 = 0.50$	.18	.19	.18	.18	4.91
	$\hat{\lambda}_2 = 1.88$	.41	.46	.35	.42	6.25
	$\hat{\lambda}_3 = 4.11$	.62	.64	.66	.63	1.65
	$\hat{\lambda}_4 = 5.01$	.87	.85	.89	.86	1.29

As expected from the data, it is obvious that there is an increasing trend in the rate of occurrences as the status of cancer aggravates. We can also observe that estimated standard errors computed using bootstrap methods are very close to those obtained from asymptotic methods, even when the sample sizes are not very large as in the case of the Invasive Carcinoma patient group. Furthermore, the minimum chi-square method and maximum likelihood method give similar

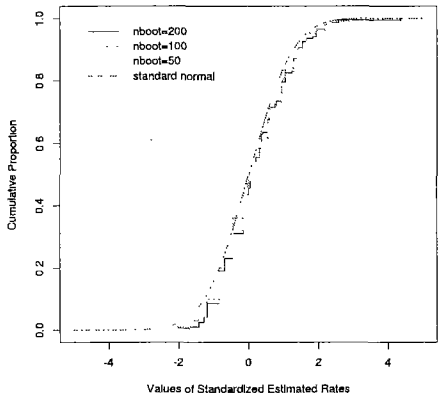
results.

We may also notice that the rate for the Dysplasia patient group is far less than that of either the Carcinoma *in situ* patient group or the Invasive Carcinoma patient group, while those two rates are not far apart. Therefore, a natural question arises as to whether the difference between the rate for the Invasive Carcinoma patient group and that for the Carcinoma *in situ* patient group is real, or if it is just due to chance variation. If the difference turns out to be just due to chance, the next question is how far the rate for the Dysplasia patient group is from the combined rate for the Invasive Carcinoma patient group and Carcinoma *in situ* patient group in terms of ASL. These questions will be answered in the following section, through hypothesis testing.

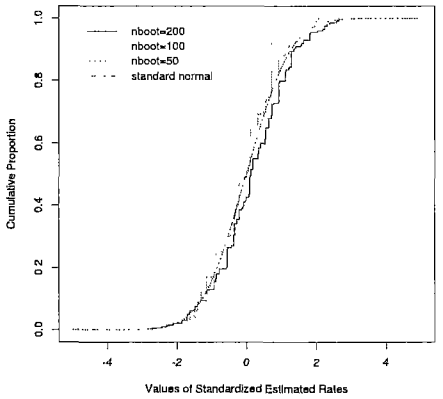
One of the big advantages in using the bootstrap method is that we can approximate the sampling distribution of the estimated quantity under study by the empirical distribution of bootstrap copies of the estimator. FIGURE 2.1 and FIGURE 2.2 depict the approximate sampling distributions of standardized estimated rates, $(\hat{\lambda}^* - \hat{\lambda})/\widehat{\mathrm{SE}}$. Estimated rates for FIGURE 2.1 and FIGURE 2.2 are obtained from the minimum chi-square and the maximum likelihood method respectively. Cumulative distribution function (cdf) of a standard normal distribution is drawn behind for comparison. We plotted the empirical cdfs to accommodate four plots all at once.

We may observe that as the number of bootstrap replications increases, the empirical cdfs of the bootstrap standardized estimated rates get closer to the cdf of standard normal distribution.

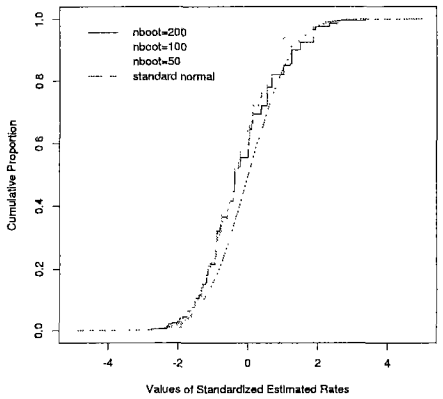
In addition to the sampling distributions of estimated rates themselves, we can also obtain the sampling distributions of goodness of fit measures either from minimum chi-square or maximum likelihood method. These values are obtained as the minimum values of either (2.1) or (2.2). It is well known that the sampling distributions of these goodness of fit measures tend to chi-square distributions with suitable degrees of freedom as the number of observations increases. We plotted the cdf of chi-square distribution in the background for comparison. Note that we used different number of degrees of freedom for each groups of patients out of trial and error. We can also check that the empirical cdfs get closer to the theoretical distribution as the number of bootstrap replications increases.



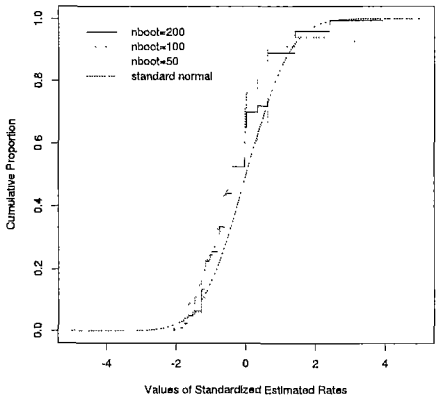
(a) Chronic Cervicitis Patient Group



(b) Dysplasia Patient Group

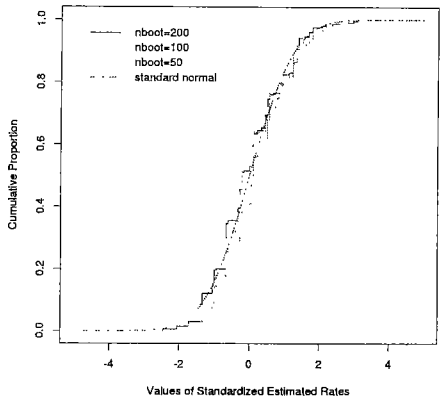


(c) Carcinoma *in situ* Patient Group

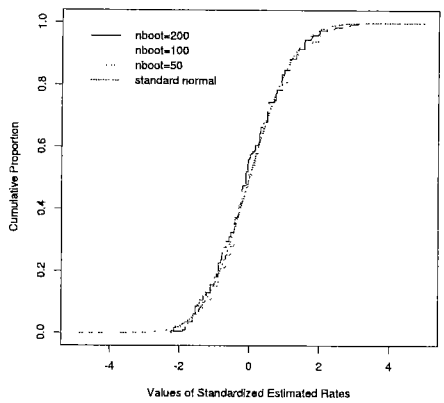


(d) Invasive Carcinoma Patient Group

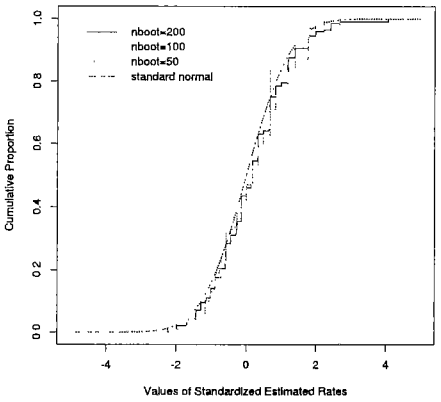
FIGURE 2.1: Bootstrap Sampling Distribution of Standardized Estimated Rates, $(\hat{\lambda}^* - \hat{\lambda})/\widehat{\mathrm{SE}}$ from Minimum Chi-square Method. Cumulative distribution function of standard normal distribution is plotted for comparison.



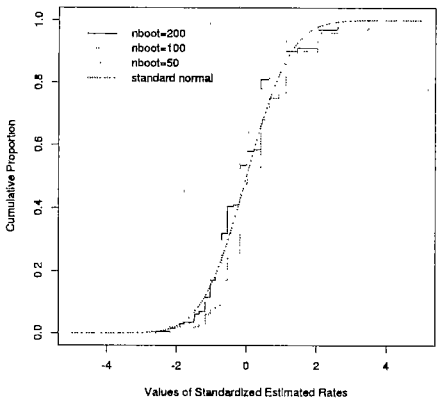
(a) Chronic Cervicitis Patient Group



(b) Dysplasia Patient Group

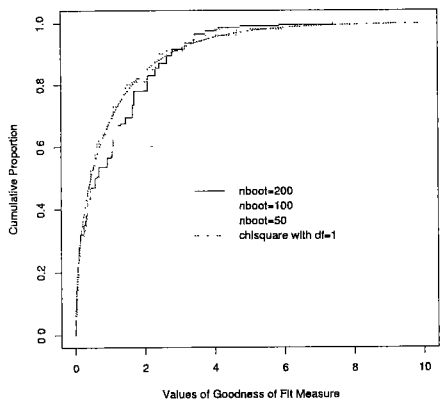


(c) Carcinoma *in situ* Patient Group

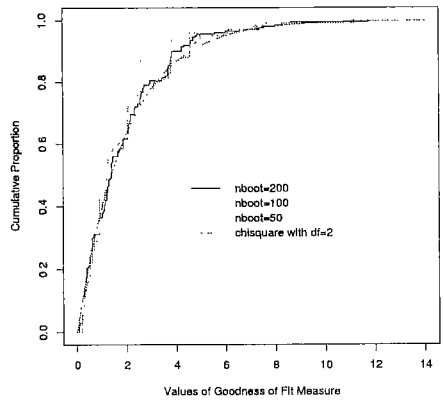


(d) Invasive Carcinoma Patient Group

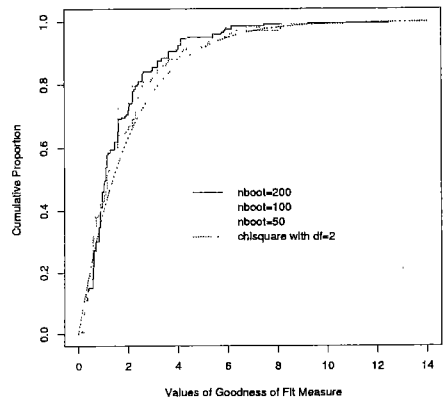
FIGURE 2.2: Bootstrap Sampling Distribution of Standardized Estimated Rates, $(\hat{\lambda}^* - \hat{\lambda})/\widehat{\mathrm{SE}}$ from Maximum Likelihod Method. Cumulative distribution function of standard normal distribution is plotted for comparison.



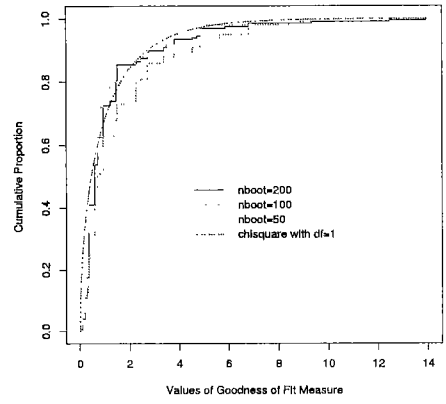
(a) Chronic Cervicitis Patient Group



(b) Dysplasia Patient Group

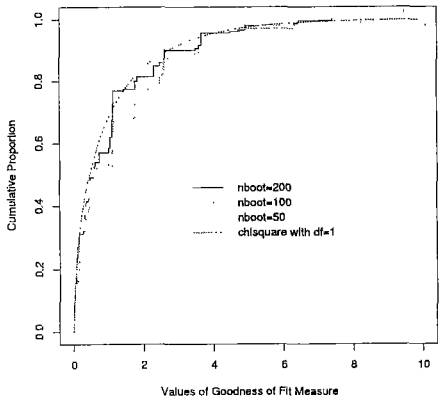


(c) Carcinoma *in situ* Patient Group

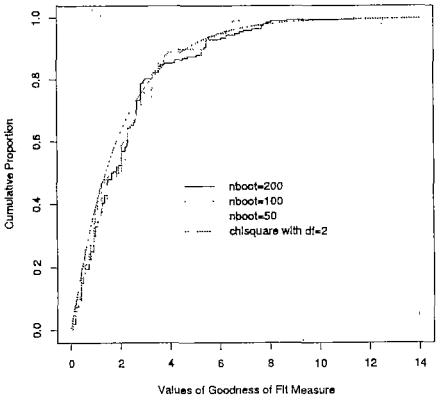


(d) Invasive Carcinoma Patient Group

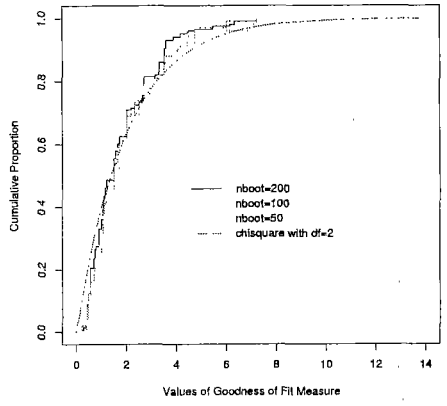
FIGURE 2.3: Bootstrap Sampling Distribution of $\sum(O^* - E_{\hat{\lambda}})^2/E_{\hat{\lambda}}$. Cumulative distribution function of chi-square distribution is plotted for comparison. Note the varying degrees of freedom.



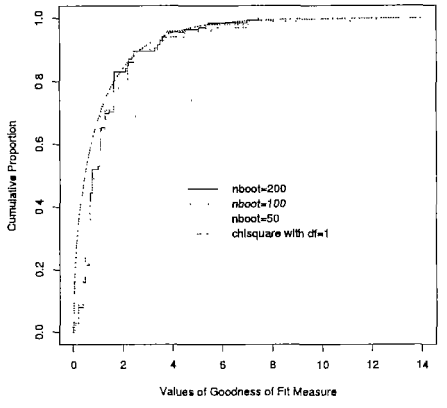
(a) Chronic Cervicitis Patient Group



(b) Dysplasia Patient Group



(c) Carcinoma *in situ* Patient Group



(d) Invasive Carcinoma Patient Group

FIGURE 2.4: Bootstrap Sampling Distribution of $-2\sum O^* \log O^*/E_{\hat{\lambda}}$. Cumulative distribution function of chi-square distribution is plotted for comparison. Note the varying degrees of freedom.

3. One Sample Test

3.1. *Theory*

Consider a one-sided hypothesis that the unknown rate λ is greter than a previously known value of λ_0. We choose a naive difference, unstudentized difference, and studentized difference as our test statistics, that is,

$$\begin{aligned} T_1 &= (\hat{\lambda} - \lambda_0), \\ T_2 &= (\hat{\lambda} - \lambda_0)/\mathrm{SE}_0, \\ T_3 &= (\hat{\lambda} - \lambda_0)/\widehat{\mathrm{SE}}, \end{aligned}$$

where $\mathrm{SE}_0 = \sigma_0/\sqrt{n}$ and $\widehat{\mathrm{SE}} = \hat{\sigma}/\sqrt{n}$. Suppose that $T_1 = t_1$, $T_2 = t_2$, and $T_3 = t_3$ are observed. Two-sided test procedure can be implemented in a similar way.

A. Asymptotic Approach

From the distributional properties of the estimated rates, the Achieved Significance Levels, denoted by ASL throughout the paper as suggested in Efron and Tibshirani (1993), for T_1, T_2, and T_3 are given by

$$\begin{aligned} \mathrm{ASL}_1 &= \Pr\left\{(\hat{\lambda}_{New} - \lambda_0) > t_1\right\} \approx \Pr\left\{Z > \frac{t_1}{\mathrm{SE}_0}\right\}, \\ \mathrm{ASL}_2 &= \Pr\left\{\frac{(\hat{\lambda}_{New} - \lambda_0)}{\mathrm{SE}_0} > t_2\right\} \approx \Pr\left\{Z > t_2\right\}, \\ \mathrm{ASL}_3 &= \Pr\left\{\frac{(\hat{\lambda}_{New} - \lambda_0)}{\widehat{\mathrm{SE}}_{New}} > t_3\right\} \approx \Pr\left\{Z > t_3\right\}, \end{aligned}$$

respectively, where Z denotes a standard normal distribution. $\hat{\lambda}_{New}$ is a hypothetical random variable with the same distribution as $\hat{\lambda}$ generated under the null hypothesis, and $\widehat{\mathrm{SE}}_{New} = \hat{\sigma}_{New}/\sqrt{n}$.

B. Bootstrap Approach

The bootstrap can also be used to approximate the ASL. Repeat *Step 1–Step 4* described in section 2, and compute $\widehat{\mathrm{SE}}_b^* = \hat{\sigma}_b^*/\sqrt{n}$ for $b = 1, \ldots, B$. Then, the bootstrap ASL's and their Monte Carlo approximations are given by

$$\mathrm{ASL}_1^* = \Pr\left\{(\hat{\lambda}^* - \hat{\lambda}) > t_1\right\} \approx \#\left\{(\hat{\lambda}_b^* - \hat{\lambda}) > t_1\right\} \Big/ B,$$

$$\text{ASL}_2^* = \Pr\left\{\frac{(\hat{\lambda}^* - \hat{\lambda})}{\widehat{\text{SE}}} > t_2\right\} \approx \#\left\{\frac{(\hat{\lambda}_b^* - \hat{\lambda})}{\widehat{\text{SE}}} > t_2\right\}\Big/B,$$

$$\text{ASL}_3^* = \Pr\left\{\frac{(\hat{\lambda}^* - \hat{\lambda})}{\widehat{\text{SE}}^*} > t_3\right\} \approx \#\left\{\frac{(\hat{\lambda}_b^* - \hat{\lambda})}{\widehat{\text{SE}}_b^*} > t_3\right\}\Big/B,$$

where $\widehat{\text{SE}}^* = \hat{\sigma}^*/\sqrt{n}$.

We can check that the ASL's obtained from the bootstrap method are asymptotically correct from the distributional properties of bootstrap estimated rates. Furthermore, in case of T_1, the bootstrap method has an advantage over the asymptotic method for the same reason as has already been noted in the previous section.

Note that we used $(\hat{\lambda}^* - \hat{\lambda})$ instead of $(\hat{\lambda}^* - \lambda_0)$ for the bootstrap to work properly, following the guidelines in Hall and Wilson (1993). In order to make an heuristic justification, recall that we drew the bootstrap data set from the fitted Poisson distribution with rate $\hat{\lambda}$, which means that $\hat{\lambda}$ plays the role of λ_0 in the bootstrap world. Therefore, we should use $\hat{\lambda}$ in place of λ_0 in order to reflect the null hypothesis. Refer to Efron and Tibshirani (1993), and Hall and Wilson (1991) for discussion on bootstrap hypothesis testing.

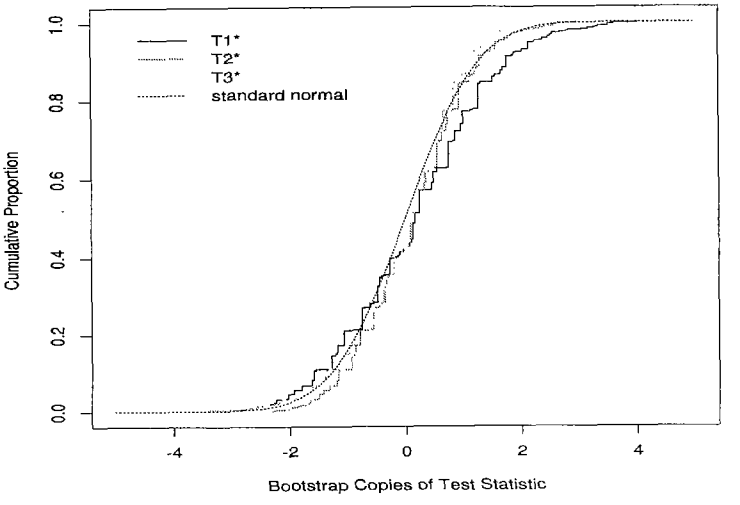
3.2. *Test Results*

In this section we will conduct a one-sided test to find out whether the rate of occurrence of positive Langerhans cells for the Dysplasia patient group is significantly greater than a known value 1.0. TABLE 3.1 summarizes the computed results.

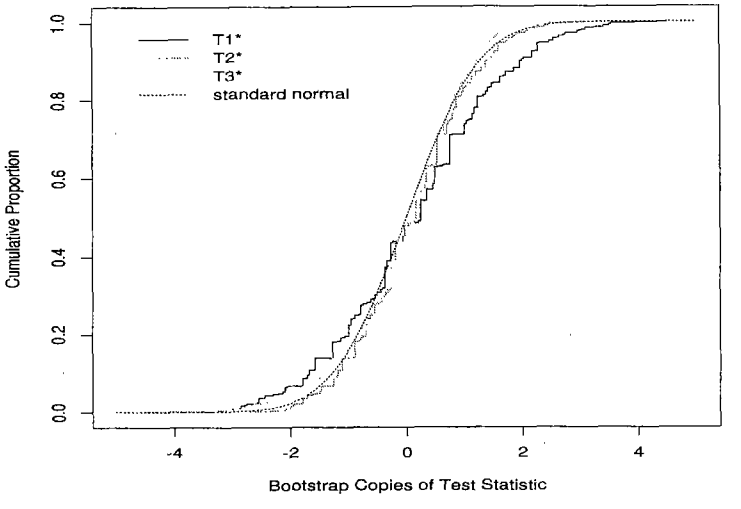
TABLE 3.1: ASL's for One Sample Test ($B = 1,000$)

Method	$\text{ASL}_1 \approx \text{ASL}_2$	ASL_3	ASL_1^*	ASL_2^*	ASL_3^*
Min χ^2	.0060	.0336	.0320	.0100	.0170
ML	.0012	.0158	.0210	.0020	.0030

We set the number of bootstrap replications at $B = 1,000$ to compute Monte Carlo bootstrap ASL's. Note that ASL_1^*, bootstrap ASL for the simplest test statistic, is close to ASL_3, the asymptotic ASL for the studentized test statistic. We would recommend bootstrap approach with T_1 as test statistic in view of computing efforts and easy implementation. From the results in the TABLE 3.1,



(a) Minimum Chi-Square Method



(b) Maximum Likelihood Method

FIGURE 3.1: Bootstrap Sampling Distribution of Various Test Statistic. The values of T_1^* are divided by $\widehat{\text{SE}}$ for proper comparison.

we may conclude that the rate for the Dysplasia patient group is significantly greater than 1.0.

Now another advantage of using Monte-Carlo bootstrap comes in. We can approximate the sampling distribution of the test statistics from the bootstrap copies. FIGURE 3.1 shows the empirical cdfs of the above three test statistics for each method of estimation. The values of T_1^* are divided by $\widehat{\text{SE}}$ for proper comparison. We can check that the distribution of T_1^* has a heavier tail than that of standard normal for both methods of estimation, but the distributions of the other two test statistics are well close to that of standard normal.

4. Two Sample Test

4.1. *Theory*

Suppose that we have two random data sets of grouped frequencies obtained from two Poisson random samples of size n_1 and n_2 with unknown rates λ_1 and λ_2 respectively. Consider a one-sided alternative hypothesis that one of the rate is greater than the other vs the null hypothesis that the observed difference may be due to chance variation.

We can compute the estimated rates $\hat{\lambda}_1$ from the first grouped data, and $\hat{\lambda}_2$ from the second grouped data in the usual way. In addition, denote the pooled estimated rate by $\hat{\lambda}_{Pooled}$, which is computed from the pooled grouped data. We choose a naive difference and a studentized difference as our test statistics, that is,

$$\begin{aligned} T_1 &= (\hat{\lambda}_1 - \hat{\lambda}_2), \\ T_2 &= \frac{(\hat{\lambda}_1 - \hat{\lambda}_2)}{\hat{\sigma}_{Pooled}(1/n_1 + 1/n_2)^{1/2}}. \end{aligned}$$

A. Asymptotic Approach

Let λ_{Pooled} be the hypothetical common unknown rate under the null hypothesis. Then, we can check that the distribution of $(\hat{\lambda}_1 - \hat{\lambda}_2)/\{\hat{\sigma}_{Pooled}(1/n_1 + 1/n_2)^{1/2}\}$ tends to a standard normal distribution, and that $\hat{\sigma}_{Pooled}$ tends to σ_{Pooled} under the null hypothesis. Suppose that $T_1 = t_1$ and $T_2 = t_2$ are observed. Then the ASL's are given by

$$\text{ASL}_1 = \Pr\left\{(\hat{\lambda}_{1,New} - \hat{\lambda}_{2,New}) > t_1\right\}$$

$$\approx \Pr\left\{Z > \frac{t_1}{\hat{\sigma}_{Pooled}(1/n_1 + 1/n_2)^{1/2}}\right\},$$

$$\mathrm{ASL}_2 = \Pr\left\{\frac{\hat{\lambda}_{1,New} - \hat{\lambda}_{2,New}}{\hat{\sigma}_{Pooled,New}(1/n_1 + 1/n_2)^{1/2}} > t_2\right\} \approx \Pr\left\{Z > t_2\right\},$$

where $\hat{\lambda}_{1,New}$, $\hat{\lambda}_{2,New}$, and $\hat{\lambda}_{Pooled,New}$ are the hypothetical random variables generated under the null hypothesis. Two-sided test can be conducted in a similar way.

B. Bootstrap Approach

Bootstrap ASL's for two sample case can be implemented as follows:

Step 1. Choose a bootstrap data set of size $n_1 + n_2$, which consists of categories obtained from the fitted Poisson distribution with rate $\hat{\lambda}_{Pooled}$.

Step 2. Compute $\hat{\lambda}_1^*$ from the first n_1 marking frequencies of the bootstrap data set, $\hat{\lambda}_2^*$ from the rest of the bootstrap data set, $\hat{\lambda}_{Pooled}^*$ from the pooled bootstrap data set, and $\hat{\sigma}_{Pooled}^*$ from $\hat{\lambda}_{Pooled}^*$.

We can check that the sampling distribution of $(\hat{\lambda}_1^* - \hat{\lambda}_2^*)/\{\hat{\sigma}_{Pooled}^*(1/n_1 + 1/n_2)^{1/2}\}$ tends to a standard normal distribution, and that $\hat{\sigma}_{Pooled}^*$ tends to σ_{Pooled} for almost all sample paths under the null hypothesis. This fact can be used to check that the bootstrap ASL's are asymptotically correct.

Step 3. Repeat *Step 1* and *Step 2* for a sufficiently large number of times, say B times, to obtain $\hat{\lambda}_{1,b}^*$, $\hat{\lambda}_{2,b}^*$, $\hat{\lambda}_{b,Pooled}^*$, and $\hat{\sigma}_{b,Pooled}^*$ for $b = 1, \ldots, B$.

Step 4. The bootstrap ASL's and their Monte Carlo approximations are given by

$$\mathrm{ASL}_1^* = \Pr\left\{(\hat{\lambda}_1^* - \hat{\lambda}_2^*) > t_1\right\} \approx \#\left\{(\hat{\lambda}_{1,b}^* - \hat{\lambda}_{2,b}^*) > t_1\right\} \Big/ B,$$

$$\mathrm{ASL}_2^* = \Pr\left\{\frac{(\hat{\lambda}_1^* - \hat{\lambda}_2^*)}{\hat{\sigma}_{Pooled}^*(1/n_1 + 1/n_2)^{1/2}} > t_2\right\}$$

$$\approx \#\left\{\frac{(\hat{\lambda}_{1,b}^* - \hat{\lambda}_{2,b}^*)}{\hat{\sigma}_{b,Pooled}^*(1/n_1 + 1/n_2)^{1/2}} > t_2\right\} \Big/ B.$$

Note again that, in case of T_1, the Monte Carlo bootstrap has a definite advantage over the asymptotic method.

4.2. *Test Results*

We give a two sample hypothesis test of whether the rate for the Invasive Carcinoma patient group is significantly greater than that for the Carcinoma *in situ* patient group. The results are summarized in the upper half of TABLE 4.1.

TABLE 4.1: ASL's for Two Sample Tests

Test	Method	$\hat{\lambda}_{Pooled}$	$\hat{\sigma}_{Pooled}$	$\text{ASL}_1 \approx \text{ASL}_2$	ASL_1^*	ASL_2^*
$\text{H}_1 : \lambda_4 > \lambda_3$	Min χ^2	4.26	2.40	0.21600	0.21000	0.21900
$(B = 1,000)$	ML	4.46	2.49	0.19100	0.20000	0.22900
$\text{H}_1 : \lambda_{34} > \lambda_2$	Min χ^2	3.26	2.02	0.00014	0.00020	0.00020
$(B = 10,000)$	ML	3.45	2.09	0.00016	0.00020	0.00010

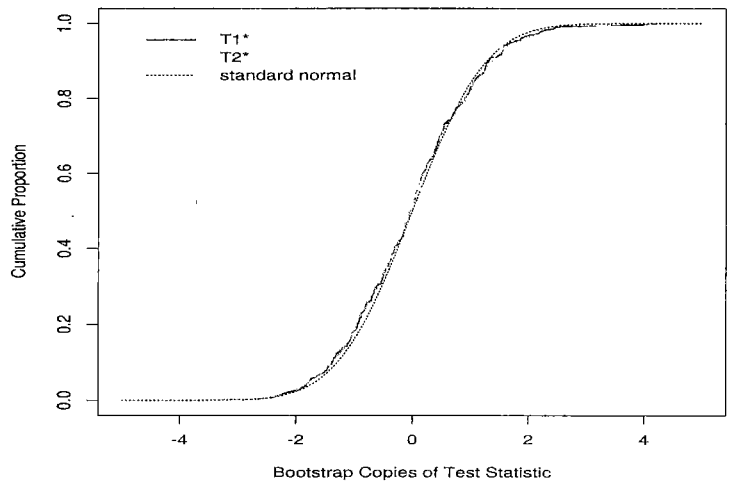
From the computed results, we may conclude that the observed differences between two estimated rates can be explained by chance variation. Again, we would recommend using the bootstrap with the naive difference as the test statistic. We may observe that all the approximate ASL's are very close.

Again, we plotted the empirical cdfs of the bootstrap copies of test statistics against the cdf of standard normal distribution in FIGURE 4.1. The values of T_1^* are divided by $\hat{\sigma}_{Pooled}(1/n_1 + 1/n_2)^{1/2}$ for proper comparison. Note that the bootstrap sampling distributions of both test statistics have heavier tails than that of standard normal.

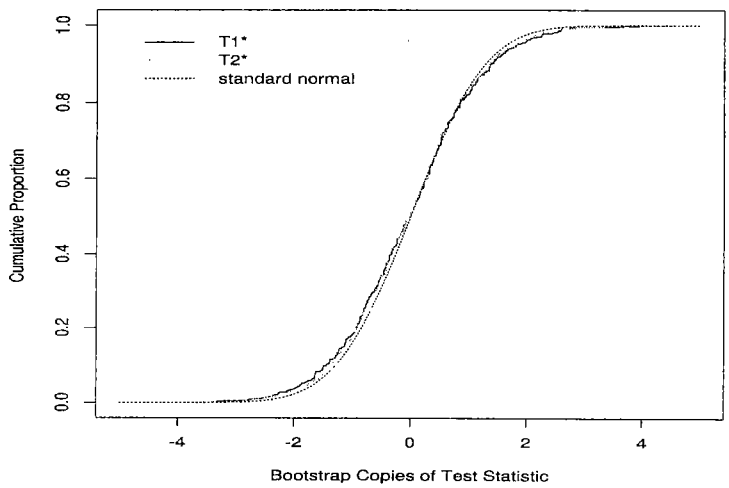
Finally, we test how far the rate for the Dysplasia patient group is from the combined rate of the Carcinoma *in situ* patient group and the Invasive Carcinoma patient group in terms of ASL. Denote the combined rate by λ_{34}. The computed results are summarized in the lower half of TABLE 4.1. We needed more bootstrap replications to compute ASL in a proper order of magnitude, and used $B = 10,000$.

From the computed results, we may conclude that the observed difference is highly statistically significant. We may note that all the ASL's are remarkably close for this moderate size of sample.

FIGURE 4.2 gives empirical cdfs of bootstrap copies of each test statistics for each methods of estimation. Again, the bootstrap sampling distributions of both test statistics have heavier tails than that of standard normal. Especially, the bootstrap sampling distributions from the minimum chi-square method look even skewed.

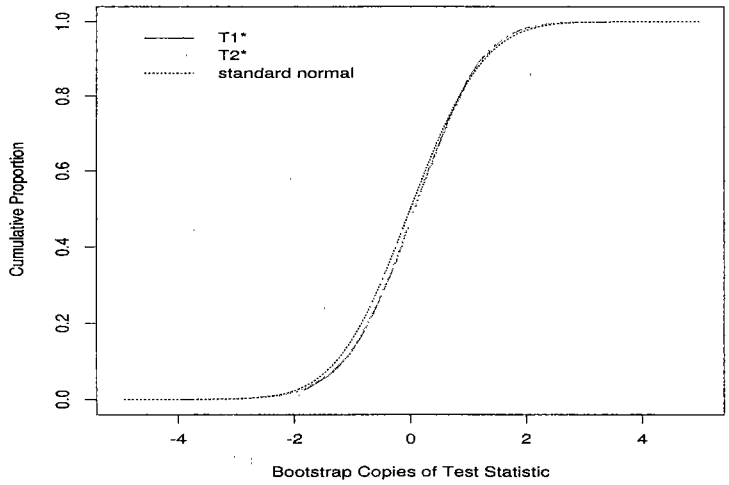


(a) Minimum Chi-Square Method

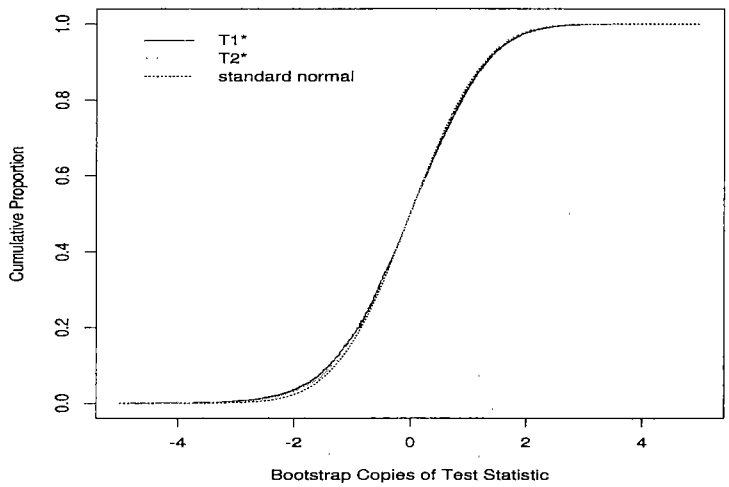


(b) Maximum Likelihood Method

FIGURE 4.1: Bootstrap sampling distribution of test statistics for testing whether the rate of the Invasive Carcinoma patient group is significantly higher than that of the Carcinoma *in situ* patient group. The values of T_1^* are divided by $\hat{\sigma}_{Pooled}(1/n_1 + 1/n_2)^{1/2}$ for proper comparison.



(a) Minimum Chi-Square Method



(b) Maximum Likelihood Method

FIGURE 4.2: Bootstrap sampling distribution of test statistics for testing whether the combined rate of the Carcinoma *in situ* patient group and the Invasive Carcinoma patient group is significantly higher than that of the Dysplasia patient group. The values of T_1^* are divided by $\hat{\sigma}_{Pooled}(1/n_1 + 1/n_2)^{1/2}$ for proper comparison.

5. Concluding Remarks

We developed statistical methods that can be used for conducting an inference on the rate of a Poisson distribution in a situation where only the grouped data are available. The Monte Carlo bootstrap method is shown to possess a number of desirable properties over the traditional asymptotic methods. We also illustrated how typical bootstrap hypothesis testing procedure can be implemented. As a by-product of Monte-Carlo bootstrap method, the approximate sampling distributions of estimated rates, goodness of fit measure, and various test statistics we obtained and plotted. S-Plus is used for numerical works.

ACKNOWLEDGEMENTS

This work was supported by a NON-DIRECTED research fund from Korea Research Foundation 1996.

References

Efron, B. (1979). Bootstrap methods. *Annals of Statistics* **7**, 1-26.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York, London.

Haitovsky, Y. (1982-1988) in the *Encyclopedia of Statistical Sciences*, ed. S. Kotz, Johnson, N. L., and Read, C., Wiley, New York.

Hall, P. and Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757-762.

Lee, M. C., Park, K. M., Kang, G., Lee, D. Y., Shin, H. S., and Park, Y. E. (1989). An immunohistochemical study on the subpopulation of Langerhans cells in cervical carcinoma using S-100 protein. *Technical Report, Dept. of Pathology, Hallym University.*

Rao, C. R. (1957). Theory of method of estimation by minimum chi-square. *Bulletin of the International Statistical Institute* **35(2)**, 25-32.