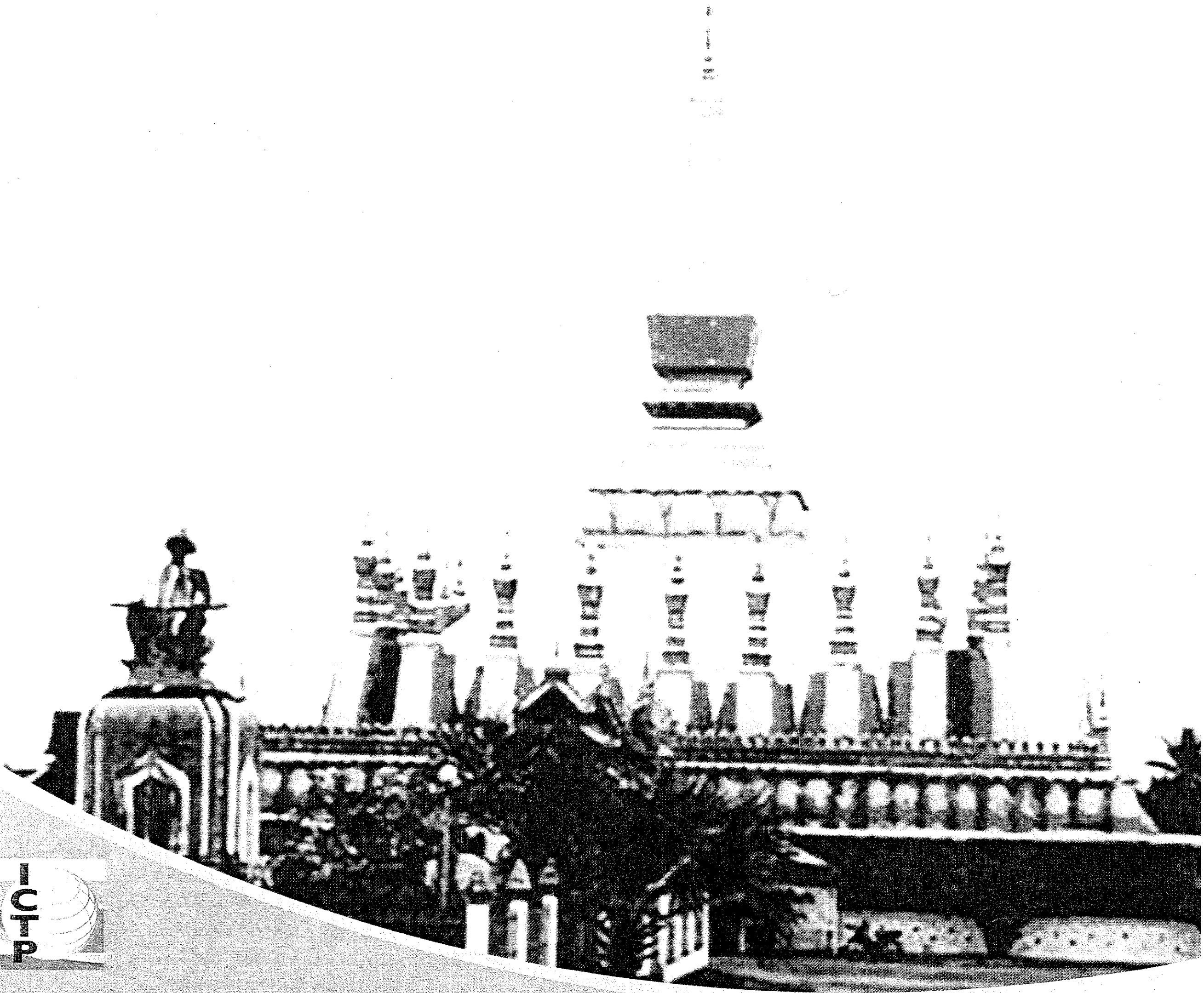
ISSN 1905-0702

Laos Journal on Applied Science

Vol.1 No.1 November 2006 SPECIAL ISSUE



Co-Sponsored by : ICTP

1st International Conference On Applied Science (ICAS 2006)
5-7 November 2006
Vientiane, Laos

Organizers : Faculty of Science
King Mongkut's Institute of Technology Ladkrabang
Faculty of Science, National University of Laos

A GENERALIZATION OF CHAO'S INEQUALITY FOR POPULATION SIZE ESTIMATION

Dankmar Böhning[1], Heinz Holling[2] and Walailuck Böhning[2],
Chukiat Viwatwongkasem[3]

[1]Applied Statistics, School of Biological Sciences, University of Reading
Reading, RG6 6FN, UK
[2]Unit for Methodology and Statistics Institute for Psychology IV
University of Münster, Münster, Germany
[3]Department of Biostatistics Faculty of Public HealthMahidol University,
Bangkok, Thailand

ABSTRACT

The paper considers estimating the population size on the basis of a continuous capture-recapture experiment. As a result of this experiment counts are observed indicating how often a unit has been identified in the study. Zero counts remain unobserved and need to be estimated. Chao (1987) provided a lower bound for this frequency under the assumption of a potential heterogeneous Poisson process. We show here that there is an inequality chain between all ratios of consecutive mixed Poisson probabilities which can be utilized as a device for the diagnosis of population heterogeneity. A further generalization to a heterogeneous Power series distribution is considered and an illustration from a drug user monitoring study in Bangkok metropolis is provided.

KEYWORDS: Cauchy-Schwarz inequality, generalized Chao estimator, diagnosis of population heterogeneity, population size estimation

1. INTRODUCTION

We consider a capture-recapture experiment in continuous time. Let $f_1, f_2, f_3, \ldots, f_m$ denote the frequencies of units identified 1,2,,...,m times within the time span of the study period. Applications include wildlife studies on biodiversity, life-science applications, or medical applications. We will discuss an application on illicit drug use in Bangkok metropolis further below. The problem is that the frequency f_0 of units identified 0 times is *unobserved* – as is the population $N = f_0 + f_1 + \ldots + f_m = f_0 + n$. It is the purpose of capture-recapture inference to provide an estimate of N or f_0.

Let us look at the probabilities $p_0, p_1, p_2, p_3, \ldots, p_m$ for observing a unit 0, 1, 2,...,m times. To find an estimate for N a model such as the Poisson is chosen for the $p_0, p_1, p_2, p_3, \ldots, p_m$

$$p_0 = e^{-\theta}, p_1 = e^{-\theta}\theta, p_2 = e^{-\theta}\theta^2 / 2 ,\ldots,$$

the parameter θ estimated with some $\hat{\theta}$ and the population size estimated with the Horvitz-Thompson estimator

$$\hat{N} = n/(1-\hat{p}_0).$$

However, it is not very realistic to assume a parametric model to hold for the entire population. It is far more realistic to assume a model allowing for population heterogeneity such as

$$p_j = \int_0^\infty e^{-\theta}\theta^j / j!\, f(\theta)d\theta$$

where $f(\theta)$ represents the heterogeneity distribution of the model parameter in the population. Instead of modeling $f(\theta)$ and estimating it in any parametric or non-parametric way, Chao (1987) proceeds in providing an important lower bound for the population size based upon the Cauchy-Schwarz inequality. The lower bound for the unknown probability of a zero is

$$p_1^2 \le 2p_0p_2 \tag{1}$$

leading to Chao's lower bound estimate for the population size

$$f_0 \ge f_1^2/(2f_2).$$

The purpose of the paper is threefold. For one, we provide a generalization of the inequality to all mixed Poisson probabilities. For two, the inequality is generalized to hold for the Power series distribution which is a general distribution for counts. For three, we show how this inequality chain can be used a diagnostic tool for detection of population heterogeneity.

2. METHOD

We will use at various stages the inequality of Cauchy-Schwarz

$$\left(\int_0^\infty u(\theta)v(\theta)f(\theta)d\theta\right)^2 \le \int_0^\infty u(\theta)^2 f(\theta)d\theta \int_0^\infty v(\theta)^2 f(\theta)d\theta \tag{2}$$

with equality if $f(\theta)$ reduces to a one-point distribution (homogeneity). We apply this inequality in the following way. Let $1, \theta, \theta^2, \theta^3, \ldots$ be the basic polynomial functions. Then

$$\left(\int_0^\infty \mu(\theta)\theta^j f(\theta)d\theta\right)^2 \le \int_0^\infty \mu(\theta)\theta^{j-1} f(\theta)d\theta \int_0^\infty \mu(\theta)\theta^{j+1} f(\theta)d\theta \tag{3}$$

if one chooses $u(\theta) = [\mu(\theta)\theta^{j-1}]^{1/2}$ and $v(\theta) = [\mu(\theta)\theta^{j+1}]^{1/2}$. Now, the results follow straightforward.

3. RESULTS

Theorem 1: Let $p_j = \int_0^\infty e^{-\theta}\theta^j / j!\, f(\theta)d\theta$ for *j=0,1,2,...*. Then:

$$\left(\int_0^\infty \exp(-\theta)\theta\ f(\theta)d\theta\right)^2 \le \int_0^\infty \exp(-\theta)\theta^{j-1} f(\theta)d\theta \int_0^\infty \exp(-\theta)\theta^{j+1} f(\theta)d\theta \tag{4a}$$

$$jp_j / p_{j-1} \le (j+1)p_{j+1} / p_j \tag{4b}$$

Proof. (4a) follows directly from (3) by choosing $\mu(\theta) = \exp(-\theta)$. (4b) follows from (4a) by incorporating the associated factorials into the Poisson probabilities:

$$(j!p_j)^2 = \left(\int_0^\infty \exp(-\theta)\theta^j f(\theta)d\theta \right)^2 \le \int_0^\infty \exp(-\theta)\theta^{j-1} f(\theta)d\theta \int_0^\infty \exp(-\theta)\theta^{j+1} f(\theta)d\theta$$
$$= (j-1)!p_{j-1})(j+1)!p_{j+1}$$

from where the result follows.

This inequality chain with respect to the ratios of consecutive mixed Poisson probabilities can easily be generalized to the mixed *power series distribution*. Suppose that the probability of observing a count of j is given by

$$p_j = a_j \int_0^\infty \mu(\theta)\theta^j f(\theta)d\theta$$

for j=0,1,2,...and for known coefficients $a_j \ge 0$ and known function $\mu(\theta)$. The power series distribution covers a variety of known distributions including the Poisson for which $a_j = 1/j!$ and $\mu(\theta) = \exp(-\theta)$. Other members are the binomial, negative binomial or logarithmic distribution. For further details see Johnson, Kotz, and Kemp (1992).

Theorem 2: Let $p_j = a_j \int_0^\infty \mu(\theta)\theta^j f(\theta)d\theta$ for j=0,1,2,.... Then:

$$\left(\int_0^\infty \mu(\theta)\theta^j f(\theta)d\theta \right)^2 \le \int_0^\infty \mu(\theta)\theta^{j-1} f(\theta)d\theta \int_0^\infty \mu(\theta)\theta^{j+1} f(\theta)d\theta \qquad (5a)$$

$$\frac{p_j / a_j}{p_{j-1} / a_{j-1}} \le \frac{p_{j+1} / a_{j+1}}{p_j / a_j} \qquad (5b)$$

Proof. (5a) follows directly from (3). (5b) follows from (5a) by incorporating the associated Power series coefficients into the Poisson probabilities:

$$p_j^2 / a_j^2 = \left(\int_0^\infty \mu(\theta)\theta^j f(\theta)d\theta \right)^2 \le \int_0^\infty \mu(\theta)\theta^{j-1} f(\theta)d\theta \int_0^\infty \mu(\theta)\theta^{j+1} f(\theta)d\theta$$
$$= (p_{j-1} / a_{j-1})(p_{j+1} / a_{j+1})$$

from where (5b) follows.

4. CONCLUSIONS

It is clear that a general lower bound estimate of the population size can be constructed from (5b) by replacing the probabilities by their associated observed frequencies:

$$\frac{f_1^2 / a_1^2}{f_2 / a_2} \le f_0 . \qquad (6)$$

We call the left-hand side of (6) the *generalized Chao-estimator* which reduces to the simple Chao-estimator for the Poisson with $a_1 = 1$ and $a_2 = 1/2$.

One of the more interesting consequences of this result can be seen in turning (4b) and (5b) into a device for detecting the presence of *population heterogeneity*. Note for clarity that in the case of homogeneity the bounds (5b) and (6) become sharp and $\frac{f_1^2 / a_1^2}{f_2 / a_2}$ will be an unbiased estimator for the population size. If there is population heterogeneity (5b) will hold and this will be reflected in the data. Thus we can expect to see that

$$\frac{f_j / a_j}{f_{j-1} / a_{j-1}} \le \frac{f_{j+1} / a_{j+1}}{f_j / a_j} \tag{7}$$

holds in the sample.

To demonstrate how this device can be used we look at a surveillance study on illicit drug use in Bangkok metropolis in the last quarter of 2001 [1]. The count distribution contains the frequency of contacts to treatment' institutions in Bangkok metropolis. See Table 1. There were $f_1 = 2955$ drug users (here heroin) that had exactly 1 contact with a treatment hospital, $f_2 = 1186$ had exactly 2 contacts, and so forth. Users with more than 11 contacts were truncated for this table, since the frequencies became very small.

Table 1 Frequency distribution of heroin users in Bangkok metropolis with exactly *j* contacts

number of contacts j	1	2	3	4	5	6	7	8	9	10
f_j	2955	1186	803	611	416	338	278	180	125	74

From Table 1 a graph can be constructed that plots $\frac{j \times f_j}{f_{j-1}}$ versus j for *j*=1,2,3,... and is shown in Figure 1. Note the clear *monotone* pattern reflecting the fulfillment of inequality (7). For lower counts the inequality is strict indicating a strong heterogeneity in Poisson distribution. For large counts the inequality becomes sharp so that here a more homogenous Poisson can be expected.

The Chao-bound is in this case $\frac{f_1^2}{2f_2} = 2955^2/(2 \times 1186) = 3861$, indicating a hidden number of heroin users at least as large as the number that has been observed.

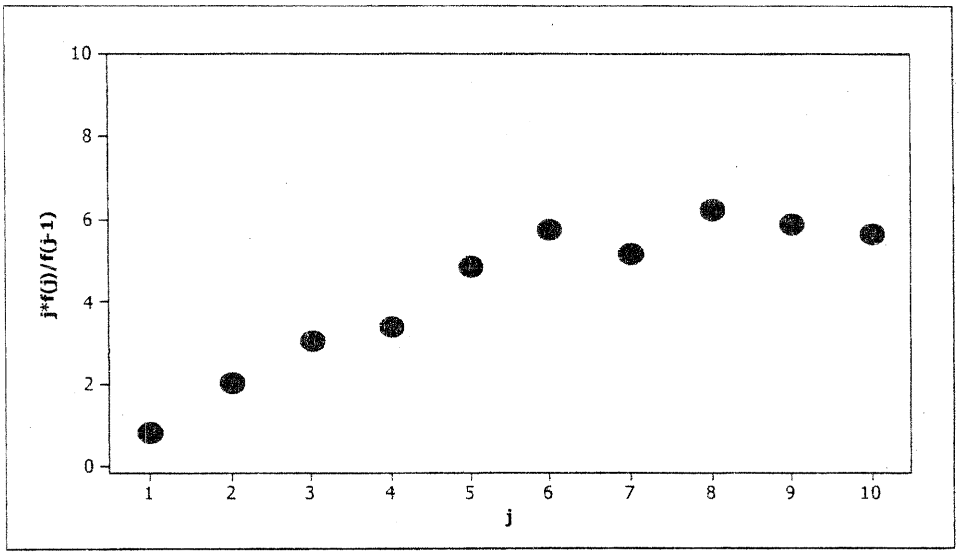


Figure 1 Graph of $\frac{j \times f_j}{f_{j-1}}$ versus j

5. ACKNOWLEDGEMENTS

This research was done with the support of the *German Research Foundation.*

REFERENCES

[1] Böhning, D., Suppawattanabodee, B., Kusolvisitkul, W, and Viwatwongkasem, C. 2004 Estimating the Number of Drug Users in Bangkok 2001: A Capture-Recapture Approach Using Repeated Entries in One List, *European Journal of Epidemiology 19*, 1075-1083.

[2] Chao, A. 1987 Estimating the Population Size for Capture-Recapture Data with Unequal Catchability, *Biometrics 43*, 783-791.

[3] Johnson, N.L., Kotz, S., Kemp, A. W. 1992 *Univariate Discrete Distributions.* New York, Wiley & Sons.