On Estimating the Zero Class from a Truncated Poisson Sample

C. J. Park*

ABSTRACT

A procedure for estimating the zero class from a truncated Poisson sample is developed. Asymptotic normality of the estimator is established and a confidence interval for the missing zero class is obtained. This procedure is compared with the method obtained by Dahiya and Gross [2]. Applications are given to illustrate the results obtained.

1. Introduction

Let X_1, $X_2, \ldots X_N$, be a rendom sample from a Poisson distribution

(1) $p(x;\lambda)=\lambda^x e^{-\lambda}/x!$ $x=0, 1, 2, \ldots,$ $\lambda>0.$

Assume that only non-zero x_i's are observed and all x_i's for which $x_i=0$ are missing. Let $T=\sum_{i=1}^{N} X_i$ and let

$S_j=\sum_{i=1}^{N} \delta_j(X_i)$, $j=0,1,2,\ldots,T$, where $\delta_j(X_i)=1$ if $X_i=j$ and $\delta_j(X_i)=0$ if $X_i \neq j$, for $i=1, 2, \ldots, N$. Then it follows that

$$T=\sum_{j=1}^{T} jS_j \quad \text{and } N=\sum_{j=0}^{T} S_j.$$

Let $S=\sum_{j=1}^{T} S_j=N-S_0.$

Recently Dahiya and Gross [2] obtained confidence interval of N utilizing

* San Diego State University

the maximum likelihood estimator. The estimation of λ has been studied in detail by many authors and references related to the estimation of λ are given in Dahiya and Gross [2].

In this paper we are concerned with estimation of N. We propose a simple estimator of N given by

(2) $\hat{N}=ST/(T-S_1)$

Using this estimator $(1-\alpha)\times 100$ percent confidence interval of N is obtained and given by

$$\hat{N}\pm z_{\alpha/2}\sqrt{S_1ST(T+S)/(T-S_1)^3}\,,\quad \text{where}$$

$z_{\alpha/2}$ is such that $P[|z|>z_{\alpha/2}]=\alpha$, z being a standard normal random variable.

2. Confidence Interval of N.

Before obtaining the confidence interval of N, we introduce the following lemmas which may shed some light on how good is the estimator given by (2).

Lemma 1 Let the sample size $N\to\infty$. Then

(a) $S/N \xrightarrow{P} 1-e^{-\lambda}$,

(b) $S_1/T \xrightarrow{P} e^{-\lambda}$, and

(c) $\hat{N}/N \xrightarrow{P} 1$,

(d) $T/\hat{N} \xrightarrow{P} \lambda$, where $\xrightarrow{P}$ is convergence in probability.

Proof Proofs of (a) and (b) are immediate consequence of standard convergence theorem for the averages of independent random variables. Proof of (c) follows by writing

$$\hat{N}/N=(S/N)/(1-S_1/T) \qquad \text{and}$$

using (a), (b), and the result given in [1, p 254]. Proof of (d) follows using (c) and the result given in [1, p 254].

Lemma 2 As the sample size $N\to\infty$,

$[(S_1/\lambda-S_0)/\sqrt{N},\ (T-N\lambda)/\sqrt{N}]$ converges in distribution to bivariate normal with mean zeros and variance-covariance matrix

$$\begin{pmatrix} e^{-\lambda}(1+\frac{1}{\lambda}) & e^{-\lambda} \\ e^{-\lambda} & \lambda \end{pmatrix}$$

Proof Let $\delta_j(X)=1$ if $X=j$ and $\delta_j(X)=0$ if $X\neq j$ Then we can write

$$(S_1/\lambda - s_0) = \sum_{i=1}^{N} [\frac{1}{\lambda}\ \delta_1(X_i) - \delta_0(X_i)] \text{ and}$$

$$T = \sum_{i=1}^{N} X_i, \text{ where } X_i\text{'s are a random sample from (1).}$$

It can be easily seen that for $i=1, 2, \ldots, N$,

$$E\left[\frac{1}{\lambda}\delta_1(X_i) - \delta_0(X_i)\right] = 0,$$

$$E(X_i) = V(X_i) = \lambda,$$

$$V\ (\frac{1}{\lambda}\delta_1(X_i) - \delta_0(X_i)) = e^{-\lambda}(1+\frac{1}{\lambda})$$

$$Cov\left[\ (\frac{1}{\lambda}\ \delta_1(X_i) - \delta_0(X_i),\ \ X_i\right] = e^{-\lambda}$$

Thus from the multivariate Central Limit Theorem, lemma 2 follows.

Now we prove the following theorem concerning distribution of $\hat{N}$, defined in (2), which is used in obtaining the confidence interval for N and S_0.

Theorem As the sample size $N \to \infty$

$$(\hat{N}-N)/\sqrt{N} \xrightarrow{D} N(0, \sigma^2), \text{ where}$$

$N(0, \sigma^2)$ denotes a normal random variable with mean zero and variance given by $\sigma^2 \doteq e^{-\lambda}(1+(1-e^{-\lambda})/\lambda)(1-e^{-\lambda})^{-2}$

Proof $(\hat{N}-N)/\sqrt{N} = (S/(1-\frac{S_1}{T}) - N)/\sqrt{N}.$

Since $(1-\frac{S_1}{T}) \xrightarrow{P} (1-e^{-\lambda})$, the limiting distribution $(\hat{N}-N)/\sqrt{N}$ is the same as the limiting distribution of

$$[S-N(1-S_1)/T)]/(1-e^{-\lambda})\sqrt{N}$$

Since $N-S=S_0$, we can write

(3) $[S-N(1-S_1/T)]/\sqrt{N} = [(S_1/\lambda - S_0) - [S_1/\lambda T]\ \ [T-N\lambda]]/\sqrt{N}$ From (b) of lemma 1, $S_1/\lambda T \xrightarrow{P} e^{-\lambda}/\lambda$, thus it follows using lemma 2 and the result given in [2, p 254] in conjunction with the fact that the left hand side of (3)

is a linear combination of asymptotically jointly normal random variables with the weight on second term converging to $e^{-\lambda}/\lambda$ in probability, $[S-N(1-S_1/T)]/[(1-e^{-\lambda})\sqrt{N}] \xrightarrow{D} N(0, e^{-\lambda}(1+\frac{1}{\lambda}(1-e^{-\lambda}))(1-e^{-\lambda})^{-2})$, which completes the proof of the theorem.

In order to find the confidence interval of N and S_0, let

$$\hat{\sigma}^2 = S_1(T+S)/(T-S_1)^2$$

Then from Lemma 1, it follows that

$$\hat{\sigma}\sqrt{\hat{N}}/\sigma\sqrt{N} \xrightarrow{P} 1, \qquad \text{hence}$$

$$(\hat{N}-N)/\hat{\sigma}\sqrt{\hat{N}} \xrightarrow{D} N(0, 1).$$

Note that $\hat{\sigma}\sqrt{\hat{N}} = \sqrt{STS_1\ (T+S)/(T-S_1)^3}$, thus $(1-\alpha)\times 100$ percent confidence interval of N is given by

$$\hat{N} \pm z_{\alpha/2}\hat{\sigma}\sqrt{\hat{N}}$$

The $(1-\alpha)\times 100$ perent confidence interval of S_o is given by

$$\hat{N}-S \pm z_{\alpha/2}\,\hat{\sigma}\sqrt{\hat{N}}$$

Remark It can be easily seen that the widths of the confidence intervals obtained by Dahiya and Gross [2] divided by the widths of the confidence interval obtained in this paper converges in probability to $R = \dfrac{(1-e^{-\lambda})}{\sqrt{(1-e^{-\lambda}-\lambda e^{-\lambda})(1+\frac{1}{\lambda}(1-e^{-\lambda}))}} \leq 1$. Thus the interval obtained in this paper gives asymptotically longer interval than the one given by Dahiya and Gross [2]. However, the numerical comparisons show that $.92 \leq R \leq 1.0$, $.92$ occurs when $\lambda = 3.8$. We do not know the relative efficiency for small N.

3. Applications

(a) Interval estimation of the number of cells in the classical occupancy problem.

Assume that a random sample of size T has been observed from a multinomial distribution with unknown N of equiprobable cells. The point estimate of N

has been obtained by Harris [4]. We obtain the interval estimate of N. Let S_j denote the number of cells occurring j times in the sample.

Then
$$\sum_{j=0}^{T} jS_j = T,$$

$$\sum_{j=0}^{T} S_j = N, \text{ and}$$

$S=N-S_0$, which denote the number of distinct cells observed in the sample. We propose a simple estimator of N given by

$$\hat{N}=ST/(T-S_1)$$

Let N and $T\to\infty$ such that $T/N\to\lambda$, $\lambda>0$. Then it can be shown that the limiting distribution of $(\hat{N}-N)/\sqrt{N}$ is the same as the limiting distribution of

$(S-N(1-S_1/T))/[(1-e^{-\lambda})\sqrt{N}]$, which converges in distribution $N(0, \sigma^2)$ where σ^2 is defined in the theorem of Section 1. Hence defining $\hat{\sigma}^2$ as was done in section 2, we can obtain the interval estimator of N.

(b) Interval estimation for truncated Poisson sample.

The data used in Dahiya and Gross [2] can be summerized as follows

x	1	2	3	4	Total
S_x	32	16	6	1	55

Form this data we have,

$S=55$, $T=86$, $S_1=32$. Thus

$$\hat{N}=88$$

$$\hat{S}_0=33$$

A 95 percent confidence interval for S_0 is $10 \leq S_0 \leq 56$.

(c) Examples when N is known.

The following exemples are summary data from Feller [3, p.123].

	S_0	S_1	S	T	$\hat{N}$	N	$\hat{\sigma}\sqrt{\hat{N}}$	$\hat{\lambda}=\frac{T}{N}$	$\hat{\lambda}_1=\frac{T}{\hat{N}}$
1	5	19	113	346	118	120	3.12	2.93	2.89
2	26	40	102	195	128	128	7.97	1.52	1.52
3	59	86	185	354	244	224	12.56	1.45	1.45
4	83	134	433	989	516	501	11.43	1.92	1.97
5	8	16	65	168	73	72	3.40	2.30	2.34

6	7	11	48	134	55	52	2.63	2.44	2.56
7	3	7	97	378	100	99	1.55	3.78	3.82
8	60	80	150	254	210	219	15.29	1.21	1.16

The last column is based on part (d) of lemma 1.

REFERENCES

[1] Cramer, H., *Mathematical Methods of Statistics,* Princeton, N.J: University Press, 1945.

[2] Dahiya, R.C. and Gross, A.J., "Estimating the Zero Class from a Truncated Poisson Sample," *Journal of the American Statistical Association,* 68 (September 1973), 731-733.

[3] Feller, W., *Probability Theory and Its Application,* John Wiley and Sons, Inc., 1959.

[4] Harris, B., "Statistical Inference in the Classical Occupancy Problem Unbiased Estimation of the Number of Classes," *Journal of the American Statistical Association,* 63 (September 1969), 837-846.