ON ESTIMATING THE TOTAL PROBABILITY OF

THE UNOBSERVED OUTCOMES OF AN EXPERIMENT*

P.J. Bickel and J.A. Yahav

University of California at Berkeley

Robbins (1968) considered the problem of estimating the total probability of the unobserved outcomes of an experiment. In this paper we suggest an estimator, based on n trials, and show that under some regularity conditions one can construct asymptotic confidence intervals for the random quantity we look for.

Consider an experiment with positive outcomes $E_1, E_2, \ldots$ with unknown probabilities $\pi_1, \pi_2, \ldots, \pi_i > 0$, $\Sigma_i \pi_i = 1$. In n independent trials suppose that E_i occurs N_i times $i=1,2,3,\ldots$ with $\Sigma_i N_i = n$. Let $\psi_i = 1$ or 0 accordingly as $N_i = 0$ or $N_i > 0$. Then the random variable $U = \Sigma_i \psi_i \pi_i$ is the sum of the probabilities of the unobserved outcomes. How to estimate U? Robbins (1968) asked this question and suggested the following answer:

Suppose we make one more independent trial of the same experiment and that in the total of n + 1 trials, E_i occurs N_i', $i=1,2,\ldots$ with $\Sigma_i N_i' = n + 1$. Let $V' = \frac{1}{n+1} \Sigma_i I_{\{N_i' = 1\}}$, where I_A is the indicator function of A. In contrast to U, V' is observable, with n + 1 trials, and can be used to predict U (we use the word predict instead of estimate since U is r.v. and not a parameter).

For W' = U − V' Robbins showed:

* Research partially supported by Office of Naval Research Contract N00014-80-C-0163.

Subject classifications: 62C12, 62G99.

Key phrases: Estimation, Asymptotic normality, Unobserved outcomes, Prediction.

$$E[W'] = 0 \quad \text{and} \quad E[W'^2] < \frac{1}{n+1} .$$

Robbins was also interested in the behavior of $E[W'^2]$ for n large. Robbins showed that in the special case in which some k of the π_i are equal to $1/k$ and all the others are 0, letting $\lambda = \frac{n}{k}$ and letting $n \to \infty$, $(n+1)E[W'^2] \to (1+\lambda)e^{-\lambda} - e^{-2\lambda} \leq (1 + \lambda^*)e^{-\lambda^*} - e^{-2\lambda^*} \sim .6080$, where $\lambda^* = .8526$ is the root of $\lambda = 2e^{-\lambda}$. What can we say if we cannot take another observation? We will suggest a predictor depending on the first n trials, and we will construct asymptotic confidence intervals under regularity conditions.

Note first that there is no unbiased predictor for U as a function of the first n trials. However,

$$E[U] = \Sigma_i \pi_i (1-\pi_i)^n .$$

If $V = \frac{1}{n} \Sigma_i I_{\{N_i=1\}}$

$$E[V] = \Sigma_i \pi_i (1-\pi_i)^{n-1}$$

Now,

$$(\Sigma_i \pi_i (1-\pi_i)^{n-1})^{n/n-1} \leq \Sigma_i \pi_i (1-\pi_i)^n \leq \Sigma_i \pi_i (1-\pi_i)^{n-1} . \tag{1}$$

We may conclude that $(V)^{n/n-1}$ tends to underpredict U while V overpredicts. V was suggested by Good (1953) as an estimator of $E[U]$.

If $W = V - U$,

$$E[W] = \Sigma_i \pi_i^2 (1-\pi_i)^{n-1} = 0(\tfrac{1}{n}) \tag{2}$$

To see this we write,

$$\Sigma_i \pi_i^{\,2}(1-\pi_i)^{n-1} \leq \Sigma_i \pi_i(\pi_i e^{-(n-1)\pi_i}) \leq \Sigma_i\ \pi_i \frac{1}{n-1} e^{-1} = e^{-1} \cdot \frac{1}{n-1}\ .$$

A little algebra shows,

$$E[W^2] = \Sigma_i \frac{1}{n}(1-\pi_i)^{n-1} \cdot \pi_i + \Sigma_i \pi_i^2 (1-\pi_i)^n \tag{3}$$

$$- \Sigma_{i\neq j} \pi_i \pi_j (1-\pi_i-\pi_j)^{n-2}(-\frac{1}{2n} + (\pi_i + \pi_j)^2 - \frac{1}{2}) = 0(\frac{1}{n})$$

Assume that if $k \to \infty$ as $n \to \infty$

A: (i) $G_n(x) = \frac{1}{k} \Sigma_{i=1}^k I_{\{n\pi_i \leq x\}} \to G_0(x)$

(ii) $\lim_{x\to 0} G_0(x) = 0$ and $\lim_{x\to\infty} G_0(x) = 1$

(iii) $\sup_n \int_0^\infty x^2 dG_n(x) < \infty$.

We note that under A, $\frac{n}{k} = \frac{1}{k} \Sigma_i n\pi_i \to \int_0^\infty x dG_0(x)$

We get

$$\sqrt{n}\ E[W] \to 0 \tag{4}$$

and

$$\begin{aligned} \sigma_n^2 = nE[W^2] \to (\int_0^\infty x dG_0(x))^{-2}\{\int_0^\infty xe^{-x} dG_0(x)\ \cdot \int_0^\infty x dG_0(x) \\ + \int_0^\infty x^2 e^{-x} dG_0(x) \int_0^\infty x dG_0(x) - (\int_0^\infty xe^{-x} dG_0(x))^2\} \end{aligned} \tag{5}$$

The limiting variance can be estimated consistently by,

(6) $$\hat{\sigma}_n^2 = \frac{1}{n}\Sigma_i\, I_{\{N_i=1\}}\left(1 - \frac{1}{n}\Sigma_i\, I_{\{N_i=1\}}\right) + \frac{2}{n}\Sigma_i\, I_{\{N_i=2\}}.$$

We note that $\hat{\sigma}_n^2 \leq 1$.

As for the limiting variance σ_0^2, we can show that

(7) $$.6080 \lessapprox \sup_{G_0} \sigma_0^2 \lessapprox .6179.$$

To see that we note,

$$\sigma_0^2 = (A/B)(1-A/B) + C/B,$$

where $A = \int_0^\infty xe^{-2}dG_0(x)$, $B = \int_0^\infty x dG_0(x)$, and $C = \int_0^\infty x^2 e^{-2} dG_0(x)$.

For the special case $x \equiv \alpha$ we get Robbin's result, namely $\sigma_0^2 = e^{-\alpha}(1-e^{-\alpha}) + \alpha e^{-\alpha}$ and $\sup_\alpha[(1+\alpha)e^{-\alpha}-e^{-2\alpha}] \approx .6080$. On the other hand we note that $x(1-x) \leq .25$ and that $\sup \frac{C}{B} = \sup_\alpha \alpha e^{-\alpha} = e^{-1} \approx .3679$ and (7) follows.

We conjecture that

(8) $$\frac{\sqrt{n}\,W}{\hat{\sigma}_n} \to N(0,1).$$

Unfortunately W is not of the form studied by Steck (1957), although we believe an extension of Steck's result will prove the conjecture. Under A, Steck's theory yields

(9) $$\frac{(V - E(U))}{\tau_n} \to N(0,1),$$

where

$$\tau_n^2 = \frac{1}{n}\Sigma_i\pi_i(1-\pi_i)^{n-1} - \Sigma_i\pi_i^2(1-\pi_i)^{2n-2} + \Sigma_{i\neq j}\pi_i\pi_j(1-\pi_i-\pi_j)^{n-2}(1-\frac{1}{n})$$

$$- \Sigma_{i\neq j}\pi_i\pi_j(1-\pi_i)^{n-1}(1-\pi_j)^{n-1} + (\Sigma_i\pi_i^2(1-\pi_i)^{n-1})^2.$$

And,

$$n\tau_n^2 \to (\int_0^\infty x dG_0(x))^{-2}\{\int_0^\infty xe^{-x}dG_0(x)\int_0^\infty x dG_0(x)$$

(10)

$$-(\int_0^\infty (xe^{-x} - x^2e^{-x})dG_0(x))^2\}.$$

For a detailed application of Steck's theory to this case, see the appendix in Bickel and Yahav (1985).

The limiting variance can be estimated consistently by

(11) $$n\hat{\tau}_n^2 = \frac{\Sigma_i I_{\{N_i=1\}}}{n} - \frac{(\Sigma I_{\{N_i=1\}} - 2\Sigma I_{\{N_i=2\}})^2}{n^2}.$$

Hence,

(12) $$\frac{\sqrt{n}(V - E[U])}{\sqrt{n\hat{\tau}_n^2}} \to N(0,1).$$

Using (12) one can construct approximate confidence intervals for E[U]. For U itself, use (4) and (6) and the Chebychev inequality to construct conservative intervals, using Chebychev's inequality pending verification of conjecture (8).

REFERENCES

Bickel, P.J. and Yahav, J.A. (1985). On estimating the number of unseen species: How many executions were there? Technical Report 43, Department of Statistics, University of California at Berkeley.

Good, I.H. (1953). On the population frequencies of species and the estimation of population parameters. Biometrika **40** 237-264.

Robbins, H.E. (1968). Estimating the total probability of the unobserved outcomes of an experiment. Ann. Math. Statist. **39** 256-57.

Steck, G.P. (1957). Limit theorems for conditional distributions. University of California Publications in Statistics, Vol. 2, No. 12, 235-284.