MODELS FOR CATALOGUING PROBLEMS

By Martin Knott

London School of Economics

1. Introduction and summary. Suppose that the members of a population fall into an unknown number of classes of various sizes. A random sample of N observations is taken and from the information in the sample must be estimated the parameters of the population; roughly speaking these are the number of classes of a given size. This description is deliberately vague because there are two models available in the literature for this problem.

Good (1953), Good and Toulmin (1956), Harris (1959), and Trybula (1959) use an infinite population and investigate estimators for useful functions of the population parameters. Examples of applications are given.

Goodman (1949) uses a finite population and obtains rather more restricted results. Des Raj (1961) also treats a very special case of this model.

Such problems have been called cataloguing problems, see Harris (1959), which is the reason for the title.

It is the object of this paper to show that the second model, that with finite population, is more suitable for these problems, and to extend the results of Goodman to match those available for the model with infinite population.

The unbiassed estimators obtained would be modified for practical use, and the main contribution of the paper is thought to lie in the simplification of the theory connected with the problem.

2. Model with infinite population. Following a remark by Harris (1959) the identifiability of the model will be investigated. It will be supposed that estimates must be made from observed values of the random variables N_i, $i = 1, 2, \cdots, N$, where N_i takes as its value for a particular sample, n_i = the number of classes represented there i times. This means that any known ordering of the population classes is ignored. Suppose that there is a constant probability $p_j > 0$ for an observation from the population to be in class j for the k population classes $j = 1, 2, \cdots, k$, where k is unknown. The population parameters are thus $\{p_1, \cdots, p_k\}$. These are identifiable up to a permutation if they are a reordering of any other system of parameters $\{q_1, \cdots, q_{k'}\}$ giving the same probabilities for every sample as the first.

For the same probability of every sample, $N_i = n_i$, under both systems

$$(2.01)\quad (N!/(1!)^{n_1}(2!)^{n_2}\cdots(N!)^{n_N})(1^{n_1}\cdots N^{n_N})_{\{p_j\}} = (N!/(1!)^{n_1}(2!)^{n_2}\cdots(N!)^{n_N})(1^{n_1}\cdots N^{n_N})_{\{q_j\}},$$

for all possible sets $n_1, \cdots, n_N$ such that $\sum_{i=1}^{N} in_i = N$. The monomial symmetric functions are taken over the sets $\{p_j\}$ and $\{q_j\}$ as indicated, and if any

Received 19 September 1966; Revised 22 November 1966.

$n_i = 0$ that part is understood to be omitted. For instance, $(1^2 2)_{\{p_1,p_2,p_3,p_4\}} = p_1^2 p_2 p_3 + p_1 p_2^2 p_3 + p_1 p_2 p_3^2 +$ similar terms with $p_2 p_3 p_4$, $p_1 p_3 p_4$, $p_1 p_2 p_4$. It follows from (2.01) that either

(a) $k = k' < N$

or

(b) $N \leqq \min [k, k']$.

If case (a) applies put $H = k = k'$, and if case (b) put $H = N$. Then

$$(1^{n_1} \cdots N^{n_N})_{\{p_j\}} = (1^{n_1} \cdots N^{n_N})_{\{q_j\}} \tag{2.02}$$

when $\sum i n_i = N$, and this is equivalent to

$$(1^s)_{\{p_j\}} = (1^s)_{\{q_j\}}, \quad s = 1, 2, \cdots, H. \tag{2.03}$$

In case (a), (2.03) implies that $\{p_j\}$ is a permutation of $\{q_j\}$, and this is as much identifiability as one can hope for. In case (b) it may be seen that only functions of the symmetric functions $(1^s)_{\{p_j\}}$, $s = 1, 2, \cdots, H$ are indentifiable, so that if $k = k' = N$ then the system is identifiable up to a permutation. This is not necessarily true when such information is not given.

To summarise the results: if it is known that $N \geqq$ the number of population classes, then $\{p_j\}$ is identifiable up to a permutation; if this is not known, then many systems of population parameters will give the same probabilities, for all possible samples.

For instance given sample size two, the sets $\{\frac{1}{2}, \frac{1}{2}\}$, $\{\frac{1}{6}, \frac{1}{6}, \frac{2}{3}\}$ are not distinguishable, neither are the sets $\{\frac{1}{3}, \frac{2}{9}, \frac{2}{9}, \frac{2}{9}\}$, $\{\frac{1}{3}, \frac{1}{3}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}\}$, $\{\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{4}{9}\}$. It is thought that in practice the system will rarely be identifiable up to a permutation, and this leads one to think of other possible models.

The most interesting functions to estimate are the number of classes and the population coverage. If $D(M, r)$ is the random variable that takes as its value for each sample of M observations the number of classes represented r times in the sample, then a useful quantity to estimate on the basis of a sample of N observations is the expected value of $D(M, r)$. The other basic function of interest is the expected value of $C(M, r)$, a random variable taking as its value for each sample of M observations the total population probability of classes represented r times in the sample. The value of $C(M, r)$ for a sample is called the population coverage in species represented r times in the sample.

It is easily shown that

$$E[D(M, r)] = \sum_{j=1}^{k} \binom{M}{r} p_j^r (1 - p_j)^{M-r}, \qquad \text{see Good (1953),} \tag{2.04}$$

and

$$E[C(M, r)] = \sum_{j=1}^{k} \binom{M}{r} p_j^{r+1} (1 - p_j)^{M-r}, \qquad \text{see Harris (1959).} \tag{2.05}$$

One can see that neither of these is identifiable on the basis of a sample of N observations unless $M \leqq N$. Good (1953), and Good and Toulmin (1956) were

interested in the estimation of these quantities for $M \geqq N$. This suggests that the basic model might well be changed, in spite of the efforts of Harris (1959) to put bounds round this lack of identifiability.

Before leaving this model it is instructive to look for an estimator of that portion of, say, (2.04) which is identifiable. Now,

$$\begin{aligned}
E\left[\sum_{i=r}^{N}\left(\binom{M}{r}\binom{N-M}{i-r}/\binom{N}{i}\right)N_i\right] &= \sum_{i=r}^{N}\left(\binom{M}{r}\binom{N-M}{i-r}/\binom{N}{i}\right)\sum_{j=1}^{k}\binom{N}{i}p_j^{\,i}(1-p_j)^{N-i}, \\
&= \sum_{j=1}^{k}\sum_{i=r}^{N}\left(\binom{M}{r}\binom{N-M}{i-r}\binom{N}{i}/\binom{N}{i}\right)p_j^{\,i}\sum_{v=0}^{i-N}\binom{N-i}{v}p_j^{\,v}(-1)^v, \\
&= \sum_{j=1}^{k}\sum_{w=r}^{N}\binom{M}{r}p_j^{\,w}\sum_{i=r}^{w}\binom{N-M}{i-r}\binom{-(N-w+1)}{w-i}, \\
&= \sum_{j=1}^{k}\sum_{w=r}^{N}\binom{M}{r}p_j^{\,w}\binom{-(M-w+1)}{w-r}, \\
&= \sum_{j=1}^{k}\binom{M}{r}p_j^{\,r}\sum_{w=0}^{N-r}\binom{M-r}{w}(-1)^w p_j^{\,w},
\end{aligned}$$

which is that part of $E[D(M, r)]$ which is identifiable when this is expanded in powers of p_j. This means that

$$\sum_{i=r}^{N}\left(\binom{M}{r}\binom{N-M}{i-r}/\binom{N}{i}\right)N_i \tag{2.06}$$

is unbiassed for the identifiable part of $E[D(M, r)]$. This estimator is the one suggested by Good and Toulmin, though their derivation was not rigorous and they were forced to assume all $p_j < \frac{1}{2}$. It is Equation (14) of their paper, and was intended as an estimator for $E[D(M, r)]$. The same technique can be used on the identifiable part of $E[C(M, r)]$. The unbiassed estimator for this is

$$\sum_{i=r+1}^{n}\left(\binom{M}{r}\binom{N-M-1}{i-r-1}/\binom{N}{i}\right)N_i \tag{2.07}$$

Notice that if $M \leqq N$ these estimators are exactly unbiassed for $E[D(M, r)]$, $E[C(M, r)]$ respectively. Putting $r = 0$, (2.07) reduces to

$$\sum_{i=1}^{N}\left(\binom{N-M-1}{i-1}/\binom{N}{i}\right)N_i, \tag{2.08}$$

which closely approximates the latter part of Good and Toulmin equation (22), this being the estimator used there for $E[C(M, 0)]$. If $M = N$ in (2.08), one has

$$\sum_{i=1}^{N}(-1)^{i-1}/\binom{N}{i}N_i \tag{2.09}$$

as an exactly unbiassed estimator of the average population probability of all classes not represented in a sample of N observations. In Good (1953) the estimator (N_1/N) was suggested for this.

3. Model with finite population. It is assumed that the population is of a known size L, and that there are K_j classes with j elements, $j = 1, 2, \cdots, N$ so that

$$\sum_{j=1}^{N} jK_j = L.$$

This means that it is known that no class has more than N elements. Samples will be taken without replacement.

If one samples without replacement, then the results of Goodman (1949) show that the parameters K_j are identifiable because there exists a unique unbiassed estimator for each K_j.

When the restriction on the number of elements in a class is not given there is quite possibly no identifiability. The population with $L = 6$, $K_2 = 3$, and that with $L = 6$, $K_1 = 3$, $K_3 = 1$, cannot be distinguished by a sample of two observations. On the other hand, a population of fewer than six elements is always identifiable for a sample of two; even if we do not know L, the only case of indistinguishability for a sample of two is the pair $L = 3$, $K_1 = 1$, $K_2 = 1$; $L = 4$, $K_2 = 2$.

Goodman obtained an explicit unbiassed estimator for $\sum_{j=1}^{N} K_j$, but did not give an explicit form for an unbiassed estimator of K_j, or for the interesting parameters corresponding to $E[D(M, r)]$, $E[C(M, r)]$, which have the same meaning as before except that $C(M, r)$ takes now the value of the total frequency of classes represented r times in the sample.

First the estimator for $E[D(M, r)]$ is established, and to do this a lemma is required.

LEMMA. *For* $j \leqq N$

$$\sum_{i=s}^{j} \binom{M}{s}\binom{N-M}{i-s}/\binom{N}{i} \times \binom{j}{i}\binom{L-j}{N-i}/\binom{L}{N} = \binom{j}{s}\binom{L-j}{M-s}/\binom{L}{M}, \tag{3.01}$$

for all M. (*As usual* $\binom{a}{b} = 0$ *for* $b < 0$.)

PROOF. For $M = 0, 1, 2, \cdots, N$, the left hand side is

$$\sum_{i=s}^{j} \binom{i}{s}\binom{N-i}{M-s}/\binom{N}{M} \times \binom{j}{i}\binom{L-j}{N-i}/\binom{L}{N}.$$

The second part of the term summed gives the probability of obtaining i successes in a random sample of size N drawn from a population of size L. The first part is the probability of obtaining j successes in a random sample of size M drawn from the size N sample just mentioned. The summation is thus equal to $\binom{j}{s}\binom{L-j}{M-s}/\binom{L}{M}$, which is the right hand side of the lemma. This is equal to

$$\binom{M}{s}\binom{L-M}{j-s}/\binom{L}{j}.$$

The relation is thus basically between polynomials in M of degree not greater than N, and so the result is true generally, for all M.

It is easily established that $E[D(M, r)] = \sum_{j=r}^{N}[\binom{j}{r}\binom{L-j}{M-r}/\binom{L}{M}]K_j$, and so the expected value of

$$\sum_{i=r}^{N} [\binom{M}{r}\binom{N-M}{i-r}/\binom{N}{i}]N_i \tag{3.02}$$

which is

$$\begin{aligned}
&\sum_{i=r}^{N} [\binom{M}{r}\binom{N-M}{i-r}/\binom{N}{i}] \sum_{j=i}^{N} [\binom{j}{i}\binom{L-j}{N-i}/\binom{L}{N}]K_j \\
&= \sum_{j=r}^{N} \sum_{i=r}^{j} \binom{M}{r}\binom{N-M}{i-r}/\binom{N}{i} \times [\binom{j}{i}\binom{L-j}{N-i}/\binom{L}{N}]K_j \\
&= \sum_{j=r}^{N} [\binom{j}{r}\binom{L-j}{M-r}/\binom{L}{M}]K_j,
\end{aligned}$$

by the Lemma, which is equal to $E[D(M, r)]$. It has thus been proved that (3.02) is an unbiassed estimator for $E[D(M, r)]$, and this result holds for all M. Putting

$M = L$, the explicit unbiassed estimator for K_r is

$$\sum_{i=r}^{N} [\binom{L}{r}\binom{N-L}{i-r}/\binom{N}{i}]N_i . \tag{3.03}$$

An unbiassed estimator of the average number of classes in a sample of M observations is, from (3.02)

$$\sum_{r=1}^{N}\sum_{i=r}^{N} \binom{M}{r}\binom{N-M}{i-r}/\binom{N}{i}N_i = \sum_{i=1}^{N} [1 - \binom{N-M}{i}/\binom{N}{i}]N_i . \tag{3.04}$$

Once more putting $M = L$ one obtains Goodman's unbiassed estimator for $\sum_{j=1}^{N} K_j$,

$$\sum_{i=1}^{N} [1 - \binom{N-L}{i}/\binom{N}{i}]N_i . \tag{3.05}$$

Using the same approach it is fairly easy to prove that for an unbiassed estimator of $E[C(M, r)]$ one has

$$(L - M)(r + 1)/(M + 1) \sum_{i=r+1}^{N} [\binom{M+1}{r+1}\binom{N-M-1}{i-r-1}/\binom{N}{i}]N_i + r \sum_{i=r}^{N} [\binom{M}{r}\binom{N-M}{i-r}/\binom{N}{i}]N_i . \tag{3.06}$$

Putting $r = 0$, the unbiassed estimator of the average population frequency of all classes not represented in a sample of size M, is

$$(L - M) \sum_{i=1}^{N} [\binom{N-M-1}{i-1}/\binom{N}{i}]N_i . \tag{3.07}$$

If $M = N$ in (3.06) one has an unbiassed estimator of the average total population frequency of classes represented r times in the sample of size N,

$$(L - N) \sum_{i=r+1}^{N} (-1)^{i-r-1}[\binom{N}{r}/\binom{N}{i}]N_i + rN_r . \tag{3.08}$$

Putting $r = 0$ once more, one has the unbiassed estimator for the average total population frequency of all classes not represented in the sample of size N,

$$(L - N) \sum_{i=1}^{N} [(-1)^{i-1}/\binom{N}{i}]N_i . \tag{3.09}$$

Des Raj (1961) obtained the unbiassed estimator for K_r when it is known that $r \leqq N$, and $K_{r+1} = K_{r+2} = \cdots = K_N = 0$. From (3.03) this is clearly

$$\binom{L}{r}/\binom{N}{r}N_r , \tag{3.10}$$

and (3.06), (3.08) also simplify a lot in this special case.

4. Comments. Section 2 has shown that there is a lack of identifiability in practice for parameters estimated in this model by the several authors mentioned in the introduction. The estimators they suggested are in many cases closely approximated by unbiassed estimators for the identifiable parts of the population parameters in question. Section 3 shows that for the alternative finite population model it is easy to find exactly unbiassed estimators for all useful functions of population parameters, under far less stringent conditions for identifiability.

Although the estimators in Section 3 may give absurd results, this was also a fault of those suggested for the infinite population model. Good and Toulmin

(1956) suggested smoothing techniques to overcome this difficulty, and they also could be used for the estimators of Section 3.

In conclusion, the finite population is neater in its theory, and has exactly unbiassed estimators instead of approximately unbiassed ones.

The author is grateful to Professor A. Stuart for supervising this work, which partially fulfilled the requirements for a Ph.D. degree at the University of London.

REFERENCES

GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika.* **40** 237–264.

GOOD, I. J. and TOULMIN, G. H. (1956). The number of new species and the increase in population coverage when a sample is increased. *Biometrika.* **43** 45–63.

GOODMAN, L. A. (1949). On the estimation of the number of classes in a population. *Ann. Math. Statist.* **20** 572–579.

HARRIS, B. (1959). Determining bounds on integrals with applications to cataloguing problems. *Ann. Math. Statist.* **30** 521–548.

RAJ, D. (1961). On matching lists by samples. *J. Amer. Statist. Assoc.* **56** 151–155.

TRYBULA, S. (1959). The estimation of frequency in a population of elements belonging to classes not represented in the sample. *Zastos. Mat.* **4** 244–248.