Capture–Recapture

In a typical capture–recapture experiment in biological and ecologic sciences, we place traps or nets in the study area and sample the population several times. At the first trapping sample a number of animals are captured; the animals are uniquely tagged or marked and released into the population. Then at each subsequent trapping sample we record and attach a unique tag to every unmarked animal, record the capture of any animal that has been previously tagged, and return all animals to the population. At the end of the experiment the complete capture history for each animal is known. Such experiments are also called mark-recapture, tag-recapture, and multiple record systems in the literature. The simplest type only includes two samples; one is the capture sample and the other the recapture sample. This special two-sample case is often referred to as a "dual system" or a "dual-record system" in the context of census undercount estimation.

The capture–recapture technique has been used to estimate population sizes and related parameters such as survival rates, birth rates, and migration rates. Biologists and ecologists have long recognized that it would be unnecessary and almost impossible to count every animal in order to obtain an accurate estimate of population size. The recapture information (or the proportion of repeated captures) by marking or tagging plays an important role because it can be used to estimate the number missing in the samples under proper assumptions. Intuitively, when recaptures in subsequent samples are few, we know that the size is much higher than the number of distinct captures. However, if the recapture rate is quite high, then we are likely to have caught most of the animals.

According to Seber [15], the first use of the capture–recapture technique can be traced back to Laplace, who used it to estimate the population size of France in 1786. The earliest applications to ecology include Petersen's and Dahl's work on fish populations in 1896 and 1917, respectively, and Lincoln's use of band returns to estimate waterfowl in 1930. More sophisticated statistical theory and inference procedures have been proposed since the paper by Darroch [5], who founded the mathematical framework of this topic. See [15]–[17] and references therein for the historical developments, methodologies, and applications.

The models are generally classified as either closed or open. In a closed model the size of a population, which is the main interest, is assumed to be constant over the trapping times. The closure assumption is usually valid for data collected in a relatively short time during a nonbreeding season. In an open model, recruitment (birth or immigration) and losses (death or emigration) are allowed. It is usually used to model the data from long-term investigations of animals or migrating birds. In addition to the population size at each sampling time, the parameters of interest also include the survival rates and number of births between sampling times. Here we concentrate on closed models because of their applications to epidemiology and health science.

Applications to Epidemiology

The capture–recapture model originally developed for animal populations has been applied to human populations under the term "multiple-record systems". A pioneering paper is that of Sekar & Deming [18], who used two samples to estimate the birth and death rates in India. Wittes & Sidel [19] were the first to use three-sample records to estimate the number of hospital patients. Related subsequent applications were given in an earlier overview by El-Khorazaty et al. [7].

Epidemiologists recently have shown renewed and growing interest in the use of the capture–recapture technique. As LaPorte et al. [13] indicated, the traditional public-health approaches to counting the number of occurrences of diseases are too inaccurate (surveillance), too costly (population-based registries; *see* **Disease Registers**), or too late (death certificates) for broad monitoring. They felt that it was time to start counting the incidences of diseases in the same way as biologists count animals. Two recent review articles [11, 12] by the International Society for Disease Monitoring and Forecasting proposed that the capture–recapture method would provide a technique for enhancing our ability to monitor disease. Reference [12] also reviewed its applications to the following categories: birth defects, cancers, drug use, infectious diseases, injuries, and diabetes as well as other areas of epidemiology.

The purpose of most applications to epidemiology is to estimate the size of a certain target population by merging several existing but incomplete lists

of names of the target population. If each list is regarded as a trapping sample and identification numbers and/or names are used as "tags", then it is similar to a closed capture–recapture setup for wildlife estimation. Now the "capture in a sample" corresponds to "being recorded or identified in a list", and "capture probability" becomes "ascertainment probability". Two major differences between wildlife and human applications are (i) there are more trapping samples in wildlife studies, whereas in human studies only two to four lists are available; and (ii) in animal studies there is a natural temporal or sequential time order in the trapping samples, whereas for epidemiologic data such order does not exist in the lists, or the order may be different for some individuals. Researchers in wildlife and human applications have respectively developed models and methodologies along separate lines. Three of these approaches are discussed after the data structure and assumptions are explained.

Data Structure and Assumptions

Ascertainment data for all identified individuals are usually aggregated into a categorical data form. We give in Table 1 a three-list hepatitis A virus example for illustration. The purpose of this study was to estimate the number of people who were infected by hepatitis in an outbreak that occurred in and around a college in northern Taiwan from April to July 1995. Our data are restricted to those records from students of that college. A total of 271 cases were reported from the following three sources: (i) P-list (135 cases): records based on a serum test conducted by the Institute of Preventive Medicine of Taiwan. (ii) Q-list (122 cases): records reported by the National Quarantine Service based on cases reported by the doctors of local hospitals. (iii) E-list (126 cases): records based on questionnaires collected by epidemiologists.

In Table 1, for simplicity, the presence or absence in any list is denoted by 1 and 0, respectively. There are seven observed cells Z_{100}, Z_{010}, Z_{001}, Z_{110}, Z_{011}, Z_{101}, and Z_{111}. Here $Z_{111} = 28$ means that there were 28 people recorded on all three lists; $Z_{100} = 69$ means that 69 people were recorded on list P only. A similar interpretation pertains to other records. Let n_1, n_2, and n_3 be the number of cases in P, Q and E, respectively. Then $n_1 = Z_{111} + Z_{110} + Z_{101} + Z_{100} = 135$. Similar expressions hold for n_2 and n_3. There is one missing cell, Z_{000}, the number of uncounted. The purpose is to predict Z_{000} or to estimate the total population size.

Table 1 Data on hepatitis A virus

Hepatitis A virus list			
P	Q	E	Data
1	1	1	$Z_{111} = 28$
1	1	0	$Z_{110} = 21$
1	0	1	$Z_{101} = 17$
1	0	0	$Z_{100} = 69$
0	1	1	$Z_{011} = 18$
0	1	0	$Z_{010} = 55$
0	0	1	$Z_{001} = 63$
0	0	0	$Z_{000} = ??$

A crucial assumption in the traditional approach is that the samples are independent. Since individuals can be cross classified according to their presence or absence in each list, the independence for two samples is usually interpreted from a 2×2 categorical data analysis in human applications. This assumption in animal studies is expressed in terms of the "equal-catchability assumption": all animals have the same probability of capture in each sample. However, this assumption is rarely valid in most applications. Lack of independence among samples leads to a bias for the usual estimators derived under the independence assumption. The bias may be caused by the following two sources:

1. List dependence within each individual (or substratum): that is, inclusion in one sample has a direct causal effect on any individual's inclusion in other samples. For example, an individual with a positive for the serum test of hepatitis is more likely to go to the hospital for treatment and thus the probability of being identified in local hospital records is larger than that of the same individual given as negative by the serum test. Therefore, the "capture" of the serum test and the "capture" of hospital records become positively dependent. This type of dependence is usually referred to as "list dependence" in the literature.
2. Heterogeneity between individuals (or substrata): even if the two lists are independent within individuals, the ascertainment of the two lists may become dependent if the capture probabilities

are heterogeneous among individuals. This phenomenon is similar to **Simpson's paradox** in categorical data analysis. That is to say, aggregating two independent **2 × 2 tables** might result in a dependent table. Hook & Regal [10] provided an example.

The above two types of dependences are usually confounded and cannot be easily disentangled in a data analysis without further assumptions. We discuss three approaches (the ecologic model, the **loglinear model**, and the sample coverage approach) that allow for the above two types of dependences.

Ecologic Models

Pollock, in his 1974 Ph.D. thesis and subsequent papers (e.g. Pollock [14]), proposed a sequence of models mainly for wildlife studies to relax the equal-catchability assumption. This approach aims to model the dependences by specifying various forms of "capture" probability. The basic models include (i) model M_t, which allows capture probabilities to vary with time; (ii) model M_b, which allows the capture of behavioral responses; and (iii) model M_h, which allows heterogeneous animal capture probabilities. Various combinations of these three types of unequal capture probabilities (i.e. models M_{tb}, M_{th}, M_{bh}, and M_{tbh}) are also proposed.

Only for model M_t are the samples independent. List dependence is present for models M_b and M_{tb}; heterogeneity arises for model M_h; and both types of dependences exist for models M_{bh} and M_{tbh}. For any model involving behavioral response, the capture probability of any animal depends on its "previous" capture history. However, there is usually no sequential order in the lists, so those models have limited use in epidemiology. Models M_h and M_{th} might be useful for epidemiological studies. Various estimation procedures have been proposed. See [15]–[17] for reviews.

Loglinear Models

The loglinear model approach is a commonly used technique for epidemiological data. Loglinear models that incorporate list dependence were first proposed by Fienberg [8] for dealing with human populations. Cormack [4] proposed the use of this technique for several ecologic models.

In this approach the data are regarded as a form of an incomplete 2^t **contingency table** (t is the number of lists) for which the cell corresponding to those individuals uncounted by all lists is missing. A basic assumption is that there is no t-sample **interaction**. This assumption implies an extrapolation formula for the number of uncounted. For three lists, the most general model is a model with main effects and three two-sample interaction terms. Various loglinear models are fitted to the observed cells and a proper model is selected using deviance statistics and the **Akaike information criterion**. The chosen model is then projected onto the unobserved cell.

List dependences correspond to some specific interaction terms in the model. As for heterogeneity, **quasi-symmetric** and partial quasi-symmetric models of **loglinear models** can be used to model some types of heterogeneity, i.e. **Rasch models** and their generalizations; see [1] and [6]. Since the quasi-symmetric or partial quasi-symmetric models are equivalent to assuming that some two-factor interaction terms are identical, the heterogeneity corresponds to some common interaction effects in loglinear models. Details of the theory and development are fully discussed in [11] and [12].

Sample Coverage Approach

The idea of sample coverage, originally from I.J. Good and A.M. Turing (Good [9]), has been used in species and animal population size estimation; see Chao & Lee [2]. The same approach was also applied to epidemiologic data in [3].

This approach aims to model dependences by some parameters, which are called "coefficients of variation", defined for two or more samples. The magnitude of the parameters measures the degree of dependence of samples. The two types of dependences are confounded in these measures. In the independent case, all dependence measures are zero. This general model encompasses the Rasch model and the ecologic models as special cases.

A common definition for the sample coverage of a given sample is the probability-weighted fraction of the population that is discovered in that sample. For multiple-sample type of data the sample coverage is

modified to be the probability-weighted fraction of the population that is jointly covered by the available samples. See [3] for a formal definition. The basic motivation here is that sample coverage can be well estimated even in the presence of two sources of dependences. Thus an estimate of population size can be obtained via the relation between the population size and sample coverage. Chao et al. [3] have shown that an estimator of C is $\hat{C} = 1 - (Z_{100}/n_1 + Z_{010}/n_2 + Z_{001}/n_3)/3$, which is one minus the average of the proportion of individuals listed in only one sample (i.e. singletons). Let D be the average of the distinct cases for three pairs of samples. In this approach, when all three samples are independent, a valid estimator is $\hat{N}_0 = D/\hat{C}$. If any type of dependence arises, then Chao et al. [3] attempt to account for the dependences by adjusting $D/\hat{C}$ based on a function of the estimates of the coefficients of variation. In the same reference, estimators of population size are proposed separately for high sample coverages (e.g. if $\hat{C}$ is over 55%) and low sample coverages.

Analysis of the Hepatitis Example (Low Sample Coverage)

Several loglinear models were fitted to the hepatitis data given in Table 1. Except for the saturated model, the loglinear models that do not take heterogeneity into account (e.g. models with one or two interaction terms) do not fit the data well, whereas all other models that take heterogeneity into account (quasi-symmetric and partial quasi-symmetric models) fit well. All those adequate models yielded very similar estimates – 1300 with an approximate estimated standard error of 520.

The coverage estimate is $\hat{C} = 51.27\%$, which is considered to be low. The average of the distinct cases for three pairs of samples is $D = 208.667$. If the incorrect independence is assumed, then an estimate would be $\hat{N}_0 = D/\hat{C} = 407$. (The loglinear independent model yields a similar estimate of 388.) It follows from Chao et al. [3] that the estimates for dependence measures are relatively large, which indicates that the three samples are pairwise positively dependent and $\hat{N}_0$ would generally underestimate. However, one cannot distinguish which type of dependence is the main cause of the bias. Incorporating the bias due to dependences along the sample coverage approach results in an estimate of 508. An estimated standard error of 40 is calculated by using a **bootstrap method** based on 1000 replications. The resulting 95% **confidence interval** is (407, 591) based on the same bootstrap replications.

This example shows that the loglinear and sample coverage approaches may give widely different estimates. Moreover, the example in [11] further shows that several loglinear models that fit the data equally well might also result in quite different estimates. **Simulation** comparisons of the two approaches and other examples with high sample coverages are provided in Chao et al. [3].

References

[1] Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort, *Biometrics* **50**, 494–500.

[2] Chao, A. & Lee S.-M. (1992). Estimating the number of classes via sample coverage, *Journal of the American Statistical Association* **87**, 210–217.

[3] Chao, A., Tsay, P.K., Shau, W.-Y. & Chao, D.-Y. (1996). Population size estimation for capture–recapture models with applications to epidemiological data, *Proceedings of Biometrics Section, American Statistical Association*, pp. 108–117.

[4] Cormack, R.M. (1989). Loglinear models for capture-recapture, *Biometrics* **45**, 395–413.

[5] Darroch, J.M. (1958). The multiple recapture census I. Estimation of a closed population, *Biometrika* **45**, 343–359.

[6] Darroch, J.N., Fienberg, S.E., Glonek, G.F.V. & Junker, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability, *Journal of the American Statistical Association* **88**, 1137–1148.

[7] El-Khorazaty, M.N., Imery, P.B., Koch, G.G. & Wells, H.B. (1977). A review of methodological strategies for estimating the total number of events with data from multiple-record systems, *International Statistical Review* **45**, 129–157.

[8] Fienberg, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables, *Biometrika* **59**, 591–603.

[9] Good, I.J. (1953). On the population frequencies of species and the estimation of population parameters, *Biometrika* **40**, 237–264.

[10] Hook, E.B. & Regal, R.R. (1993). Effects of variation in probability of ascertainment by sources ("variable catchability") upon capture–recapture estimates of prevalence, *American Journal of Epidemiology* **137**, 1148–1166.

[11] International Society for Disease Monitoring and Forecasting (1995). Capture–recapture and multiple-record systems estimation I: history and theoretical

development, *American Journal of Epidemiology* **142**, 1047–1058.

[12] International Society for Disease Monitoring and Forecasting (1995). Capture–recapture and multiple-record systems estimation II: Applications in human diseases, *American Journal of Epidemiology* **142**, 1059–1068.

[13] LaPorte, R.E., McCarty, D.J., Tull, E.S. & Tajima, N. (1992). Counting birds, bees, and NCDs, *Lancet* **339**, 494–495.

[14] Pollock, K.H. (1991). Modelling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future, *Journal of the American Statistical Association* **86**, 225–238.

[15] Seber, G.A.F. (1982). *The Estimation of Animal Abundance*, 2nd Ed. Griffin, London.

[16] Seber, G.A.F. (1986). A review of estimating animal abundance, *Biometrics* **42**, 267–292.

[17] Seber, G.A.F. (1992). A review of estimating animal abundance II, *International Statistical Review* **60**, 129–166.

[18] Sekar, C. & Deming W.E. (1949). On a method of estimating birth and death rates and the extent of registration, *Journal of the American Statistical Association* **44**, 101–115.

[19] Wittes, J.T. & Sidel, V.W. (1968). A generalization of the simple capture–recapture model with applications to epidemiological research, *Journal of Chronic Diseases* **21**, 287–301.

(*See also* **Structural and Sampling Zeros**)

ANNE CHAO