Applications of Capture-Recapture Methods

Thomas N. Herzog
FHA/HUD

ABSTRACT

Applications of capture-recapture procedures go back to at least 1896, to the work of Petersen [1896], who was interested in the size of fish populations. Another early paper by Lincoln [1930] was concerned with waterfowl. The method gets its name from such applications concerned with estimating the number of fish in a pond, or more generally, estimating the population size of various species in the wild.

Another classic paper, by Sekar and Deming [1949] was concerned with estimating birth and death rates in an area near Calcutta, India. Capture-recapture procedures have more recent application to (1) estimating the size of the undercount in censuses and (2) estimating the number of duplicate records on a list or a database. Other possible applications include (1) estimating the number of drug addicts in the United States and (2) estimating the number of children in Massachusetts possessing a specific congenital abnormality. Finally, Ball et al. [1999] have used capture-recapture methods to estimate the number of people killed in Kosovo. We review the general methodology first and then describe a few of these applications.

Capture-Recapture Procedures and Applications

Methods known as capture-recapture procedures have applications to a number of the problems. The simplest version of this methodology, also known as *dual systems estimation*, involves two separately compiled, but incomplete, lists of the members of a population. With three or more such lists, the methodology is called *multiple systems estimation*.

Applications of capture-recapture procedures go back to at least 1896, to the work of Petersen [1896] who was interested in the size of fish populations. Another early paper by Lincoln [1930] was concerned with waterfowl. The method gets its name from such applications concerned with estimating the number of fish in a pond, or more generally, estimating the population size of various species in the wild. Scheuren [2004] is a precursor of this work.

For example, we could take a sample of fish in a pond and tag each of the fish so selected. A day later, we could take a second sample of fish from the pond, and count the number of fish in this second sample that had been tagged the day before. Then, using the methods of this section, we could estimate the total number of fish in the pond.

Another classic paper, by Sekar and Deming [1949] was concerned with estimating birth and death rates in an area near Calcutta, India. Capture-recapture procedures also have application to (1) estimating the size of the undercount in censuses and (2) estimating the number of duplicate records on a list or a database. These are two of the specific applications we consider in depth in this work. Other possible applications include (1) estimating the number of drug addicts in the United States and (2) estimating the number of children in Massachusetts possessing a specific congenital abnormality. We discuss the general methodology first before considering specific applications. Our primary reference source for this section is the textbook by Bishop, Fienberg, and Holland [1975].

General Approach to the Two-Sample Capture-Recapture Problem

Let N be the total number of individuals in the population of interest. Let x_{11} denote the number of individuals observed to be in both samples. Let x_{12} denote the number of individuals observed to be in the first sample but not the second; and let x_{21} denote the number of individuals observed to be in the second sample but not the first. Finally, we let x_{1+} $(= x_{11} + x_{12})$ and x_{+1} $(= x_{11} + x_{21})$ denote the number of individuals in the first and second samples, respectively. We summarize the observations in a two-by-two contingency table with one missing value denoted by x_{22}:

TABLE 1

Basic 2-by-2 Contingency Table

	Second Sample		
First Sample	Present	Absent	Total
Present	x_{11}	x_{12}	x_{1+}
Absent	x_{21}	x_{22}	
Total	x_{+1}		N

The goal here is to estimate the missing value, x_{22}, which leads easily to an estimate of the total population, N. The traditional estimator of N is

$$\hat{N} = x_{11} + x_{21} + x_{12} + \frac{x_{12}x_{21}}{x_{11}}.$$

This estimator can be shown to be equal to

$$\hat{N} = \frac{x_{1+}x_{+1}}{x_{11}}.$$

and can be derived by assuming the following identity holds:

$$x_{11}x_{22} = x_{12}x_{21}.$$

The estimator, $\hat{N}$, is sometimes called the *Lincoln-Peterson estimator* in which case we could rewrite it as $\hat{N}_{LP}$.

Such dual systems estimators rely on three assumptions. The first assumption is that the samples or lists are independent. In other words,

$$P[record\ i\ on\ list\ L_1 \mid record\ i\ on\ list\ L_2] = P[record\ i\ on\ list\ L_1].$$

The second assumption is that the population of interest is homogeneous in the sense that each member of the population has an equal chance of being captured *for a given list*.

The third assumption is that there are no errors when matching records across lists. Moreover, an accurate estimate of the number of items in both lists, x_{11}, is particularly crucial to this process.

Before considering some examples of the above, we want to make two additional comments. First, most, if not all of this work, should be automated because clerical matching is too slow and too prone to produce errors. We next summarize some comments of Sekar and Deming [1949] regarding the assumption of independence.

Sekar and Deming [1949;page 106] argue that it is important to realize that "correlation signifies heterogeneity in the population [because] it implies that events that fail to be detected do not form a random sample of the whole population of events. This heterogeneity may arise only if there are differences in reporting rates among different segments of the population, resulting in the group of failures being weighted disproportionately by the different segments."

"It therefore follows that the correlation can be minimized" by (1) partitioning "the population into homogeneous groups and" then (2) "calculating the total number of events separately for each group". The grand total can then be obtained by addition. For more details, the interested reader should see the original Sekar and Deming article.

Weevils in a Box of Wheat

As Bishop, Fienberg, and Holland [1975] reports, Andrewartha [1961] describes an experiment in which about 2,000 weevils were placed in a box of wheat and allowed to disperse. An initial sample of 498 weevils was drawn without replacement from the box of wheat. Each of the weevils selected was marked with a small spot of paint and put back into the box of wheat. A week later, a second sample of 110 weevils was drawn without replacement. Twenty of those selected in the second sample were observed to have a spot of paint. Using the capture-recapture scheme described above, we obtain an estimate of

$$\hat{N}_{LP} = \frac{x_{1+}x_{+1}}{x_{11}} = \frac{498 \cdot 110}{20} = 2{,}739$$

for the total number of weevils in the box.

Estimating Birth and Death Rates in India

Sekar and Deming [1949] describe the results of a study conducted during February 1947, in an area known as the Singur Health Centre, near Calcuttta, India. The total area of the Centre is about 33 square miles. At the time of this study, the total population of the area was about 64,000 people living in about 8,300 houses.

Sekar and Deming use capture-recapture methods to estimate the number of births and deaths for residents of this area during each of the years 1945 and 1946.

The headman of each village periodically submits a list of births and deaths to a registrar. The registrar coordinates this information with a second report from each village and a list from the Maternity and Child Welfare Department. We refer to the resulting list as the "registrar's list of births and deaths" – the "R-list".

During an eleven-week period beginning on February 11, 1947 interviewers from the All-India Institute of Hygiene and Public Health visited every house within the Singur Health Centre to prepare a list of all of the births and deaths that occurred during 1945 and 1946 – the I-list.

After deleting the non-verifiable, illegible, incomplete, and incorrect items from the R-list, Sekar and Deming [1947] applied the two-sample technique described above. We summarize these corrected data in Table 2 below, a table similar to Table 6.2-1 in Bishop, Fienberg, and Holland. Table 6.2-1 is an abbreviated version of Table I on page 108 of Sekar and Deming.

TABLE 2

The Investigators' Report on the Comparison of the Lists of the Singur Health Centre

Event[1]	Year	Registrars' Lists			Interviewers' Lists
		Total	Found in Interviewer's Lists	Not Found in Interviewer's Lists	Extra in Interviewers' Lists
Births	1945	1,504	794	710	741
	1946	2,242	1,506	736	1,009
Deaths	1945	1,083	350	733	372
	1946	866	439	427	421

[1] The events referred to here are those listed as occurring in the village that did not involve institutionalized non-residents.

To illustrate the capture-recapture methodology, we consider the deaths occurring during 1946. We summarize the data of interest in a 2-by-2 table:

TABLE 3
Number of Deaths Occurring During 1946 in the Singur Health Centre

R-List	I-List	
	Present	Absent
Present	439	427
Absent	421	?

The estimate of the total number of deaths occurring during 1946 is

$$\hat{N}_{LP} = \frac{(439+421)\cdot(439+427)}{439} = 1{,}696$$

where we have rounded the result to the nearest integer.

Finally, for computing birth and death rates in this area, the population base assumed was that furnished by the household interviews.

Estimating the Number of Duplicates within a Database

A database consisting of information on individuals applying for a credit card contained the following six data elements on these individuals: Last Name, First Name, Middle Name, Social Security Number, Date of Birth, and Place of Birth. Of these, 498 pairs of records having identical Social Security Numbers were determined to be duplicates (after extensive review) while 110 pairs that agreed exactly on Last Name, First Name, Middle Name, Date of Birth, and Place of Birth were similarly determined to be duplicates. Finally, 20 pairs of records that agreed on all six data elements were determined to be duplicate records. (This is the number of duplicates in common between the two matching schemes.) How many duplicate records were to be found in the entire database?

The answer is 2,739. This is the answer to the problem about the number of weevils in a box of cereal, as the equations in both examples are identical.

Killings and Refugee Flow in Kosovo

Ball et al. [1999] estimated the number of people killed in Kosovo during the spring of 1999. They had four separate source lists of individual victims of killing:

- American Bar Association Central and Eastern Europe Law Initiative (ABA/CEELI).
- Exhumations (EXH)
- Human Rights Watch (HRW)
- Organization for Security and Cooperation in Europe (OSCE)

The data of interest are summarized in the following $2 \times 2 \times 2 \times 2$ contingency table.

TABLE 4
Number of Individual Victims of Killing
By Documentation Status (including victims with imputed dates of death)

	ABA	Yes	Yes	No	No	Total
	EXH	Yes	No	Yes	No	
HRW	OSCE					
Yes	Yes	27	32	42	123	
Yes	No	18	31	106	306	
No	Yes	181	217	228	936	
No	No	177	845	1,131	??	
Total						4,400

While there is essentially only one method of estimation -- the basic (Lincoln-Peterson) estimator -- in the case of two lists, there is much more flexibility in the case of three or more lists. Both Bishop, Fienberg, and Holland [1975] and Marks, Seltzer, and Krotki [1974] describe a variety of estimators that can be used in such a multiple systems situation. The estimates, we computed are given in the two tables below. The estimates of the number of victims based on the use of three of the four lists are computed using equations 6.4-14 and 6.4-15 on page 239 of Bishop, Fienberg, and Holland [1975]. The estimate based on the use of four lists is computed using Equation 14 on page 41 of Ball, Betts, Scheuren, Dudokovic, and Asher [2002]. We note that our estimates are in most cases slightly different from those reported in Ball et al. The exception is that our estimate based on all four lists is considerably higher than theirs.

TABLE 5
Dual System Estimates

	EXH	HRW	OSCE
ABA	7,242	9,691	5,972
EXH		6,779	7,137
HRW			5,462

TABLE 6
3-way and 4-way system estimates (under saturated model)

Lists Used	Estimated Population Total
ABA, EXH, HRW	11,811
ABA, EXH, OSCE	22,301
ABA, HRW, OSCE	12,255
EXH, HRW, OSCE	8,016
ABA, EXH, HRW, OSCE	16,942

Further Thoughts

- An assumption underlying the Lincoln-Peterson estimator is that the probabilities of being on the list frames are homogeneous (i.e., identical). Haines, Pollock, and Pantula [2000] extend the results to the case where the inclusion probabilities are heterogeneous (i.e., unequal). For example, larger farms may have a greater probability of being on a list frame than smaller farms. They also show how the inclusion probabilities can be modeled as functions of auxiliary variables using a logistic regression model. For example, in capture-recapture work, the probabilities of inclusion (i.e., capture) may vary with the age, gender, or size of the species being studied.

- When neither the homogeneity assumption nor the independence assumption is satisfied, capture-recapture procedures can produce very inaccurate results. See Winkler [2004] for some examples.

- Finally, in the Lincoln-Peterson approach, it is possible that the denominator, x_{11}, is zero. Chapman [1951] has proposed an estimator

$$\hat{N}_{CH} = \frac{(x_{1+}+1)(x_{+1}+1)}{(x_{11}+1)} - 1$$

  that circumvents this difficulty. The Chapman estimator also leads to a reduction of the bias in the estimation process. Of course, if the denominator of the estimator is close to zero, the estimates will be unstable in the sense that small changes in the value of the denominator will lead to large changes in the value of the desired estimate. For more details on the bias of such estimators, the reader should see Sekar and Deming [1949] and Chapman [1951].

References

Andrewatha, H.G., *Introduction to the Study of Animal Populations*, Chicago, University of Chicago Press, 1961.

Ball, P., W. Betts, F.J. Scheuren, J. Dudukovich, and J. Asher, *Killings and Refuge Flow in Kosovo, March – June 1999,* American Academy for the Advancement of Science, January 3, 2002.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland, *Discrete Multivariate Analysis*, MIT Press, Cambridge, MA, 1975.

Chapman, D.G., *Some Properties of the Hypergeometric Distribution with Applications to Zoological Censuses*, University of California, University of California Publication in Statistics, 1951.

Lincoln, F.C., "Calculating waterfowl abundance on the basis of banding returns", *Cir. U.S. Department of Agriculture*, Vol. 118, pages 1-4, 1930.

Marks, E.S., W. Seltzer, and K. R. Krotki, *Population Growth Estimation,* The Population Council, New York, 1974.

Petersen, C.G.J., "The yearly immigration of young plaice into the Limfiord from the German Sea", *Rep. Dan. Biol. Stn*., Vol. 6, pages 5-84, 1896.

Scheuren, F.J., "The History Corner", *The American Statistician*, Vol. 58, No. 1, 2004.

Sekar and W.E. Deming, “On a Method of Estimating Birth and Death Rates and the Extent of Registration”, *J. Amer. Stat. Assoc.*, Vol. 44, pages 101-115, 1949.

Winkler, W.E., “Methods for Evaluating and Creating Data Quality”, *Information Systems*, 29 (7), 531-550, 2004.