Wiley StatsRef:
Statistics Reference Online

Capture-Recapture for Human Populations

By *Anne Chao*

Keywords: *ascertainment data, dual-record system, epidemiology, heterogeneity, local dependence, multiple-record systems, overlapping information, population size, sample coverage*

Abstract: Capture–recapture methods, originally developed for estimating demographic parameters of animal populations, have been applied to human populations. In epidemiology and health sciences, most surveillance studies and prevalence surveys based on multiple records of incomplete lists are likely to miss some cases, and thus the number of ascertained cases in the final merged list tends to undercount the true size of the target population. Capture–recapture methods can be applied for these types of studies/surveys to provide useful estimators for the size of the target population and adjust for under-ascertainment. This article describes the history of the method and focuses on population size estimation based on the sample-coverage approach, which also provides measures to quantify the extent of overlapping information and assess the dependences among samples. The R package CARE1 (CApture–REcapture, part 1) is applied to two real examples for illustration. Other models are briefly discussed.

In a typical capture–recapture experiment in the biological sciences, we place traps or nets in the study area and sample the population several times. In the first trapping sample, a number of animals are captured; the animals are uniquely tagged or marked and released into the population. Then, in 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 demographic parameters for animal populations. Biologists 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 population 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.

National Tsing Hua University, Hsinchu, Taiwan

Update based on original article by Anne Chao, Wiley StatsRef: Statistics Reference Online, © 2014, John Wiley & Sons, Ltd

DOI: 10.1002/9781118445112.stat04855.pub2

The basic idea of the two-sample capture–recapture technique can be traced back to a 1786 paper by Pierre Simon *Laplace*, who used it to estimate the population size of France in 1802[1,2], and even earlier to John Graunt who used the idea to estimate the effect of plague on the population size of England around 1600[3]. 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 population in 1930. Two-sample models were extended to the multiple-sample case in 1938 by Schnabel[4]. So a multiple-sample capture–recapture experiment is also referred to as the *Schnabel census*. 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 Refs 2, 6–13 for the historical developments, methodologies, and applications.

The models in animal populations are generally classified as either closed or open. In a closed model, the size of a population, which is of 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 human populations.

1 Applications to Epidemiology and Health Sciences

Traditionally, epidemiologists and researchers in health sciences have attempted to enumerate all relevant cases to obtain the prevalence rates for various diseases. When multiple lists of cases ascertained by different methods are available, cases in various lists are usually merged and duplicate cases are eliminated. The overlapping information is thus ignored. This typical approach assumes complete ascertainment and does not correct or adjust for underascertainment. As LaPorte *et al*.[14] indicated, the traditional public health approaches for counting the number of occurrences of diseases are too inaccurate (surveillance), too costly (population-based registries), or too late (death certificates) for broad monitoring (*see* **Disease Registers: Overview**; **Disease Registers: Basic**). Most prevalence surveys merging several records of lists are likely to miss some cases and thus the number of cases in the merged list will be negatively biased[15–18].

The capture–recapture model has been applied to epidemiology and related sciences under the term *multiple-record systems* because most applications estimate the size of a certain target population by merging several existing but incomplete lists of names of the target population[19]. In this method, researchers count incidence of a disease in human populations much as ecologists and biologists count animals. A pioneering paper is that of Sekar and Deming[20], who used two samples to estimate the birth and death rates in India during 1945 and 1946. Wittes and Sidel[21] were the first to use three-sample records to estimate the number of hospital patients. Subsequent developments and applications were reviewed in Refs 17–19, 22. The applications cover a wide range of areas: birth defects (*see* **Birth Defect Registries**), cancers (*see* **Cancer Registries**), drug use, infectious diseases, injuries, diabetes, and mental illness among others[17–19,23]. The applications also extend to the size estimation of the elusive target populations in social studies; examples are populations of criminals, the homeless, and those with sensitive antisocial or stigmatized behaviors, or incidence of family violence, child prostitution, and drug abuse[23,24,25].

The framework to model multiple-record systems for human populations is similar to that of a closed capture–recapture setup for wildlife estimation: Each list (or sample) is regarded as a trapping sample and identification numbers and/or names are used as "tags". 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 a few lists are available, and (ii) in animal studies, there is a natural

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat04855.pub2

temporal or sequential time order in the trapping samples, whereas for epidemiological data such order does not exist in the lists, or the order may be different for some individuals[22]. Researchers in wildlife and human applications have, respectively, developed models and methodologies along separate lines. Various models/approaches are discussed after the data structure and assumptions are explained.

2 Data Structure and Assumptions

We first present the notation. Let the true unknown population size (i.e., the number of individuals in the target population) be N and all individuals can be conceptually indexed by 1, 2, … , N. There are t samples (lists, records, or sources) and they are indexed by 1, 2, … , t. Ascertainment data for all identified individuals are usually aggregated into a categorical data form. Denote $Z_{s_1 s_2 \ldots s_t}$ as the number of individuals in the category $s_1 s_2 \ldots s_t$, where $s_j = 0$ denotes absence (noncapture) and $s_j = 1$ denotes presence (capture) in the jth list. There are M identified individuals, that is, M equals to the sum of all observable cell counts. Denote n_j, $j = 1, 2, \ldots, t$, as the number of individuals identified in the jth list.

We give in Table 1 a three-list hepatitis A virus (HAV) example[22] 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 $M = 271$ cases were reported from the following three sources: (i) P-list ($n_1 = 135$ cases): records based on a serum test conducted by the Institute of Preventive Medicine of Taiwan. (ii) Q-list ($n_2 = 122$ cases): records reported by the National Quarantine Service based on cases reported by the doctors of local hospitals. (iii) E-list ($n_3 = 126$ cases): records based on questionnaires collected by epidemiologists.

In Table 1, there are seven observed cells or categories with counts 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. There is one missing cell, Z_{000}, *the number of uncounted*. The purpose is to predict Z_{000} or equivalently to estimate the total population size N $(= M + Z_{000})$.

Similar notation and interpretations can be extended to data with more than three lists. As is explained in Section 6, the HAV data represent a data set with relatively low overlapping fraction. We also give a four-list diabetes data with high overlapping fraction. The 15 observed categories with counts are shown in Table 2. These data were collected by Bruno *et al*.[26] in a community in Italy based on the following four records: diabetic clinic and/or family physician visits (List 1 with $n_1 = 1754$ cases), hospital discharges (List 2 with $n_2 = 452$ cases), prescriptions (List 3 with $n_3 = 1135$ cases), and purchases of reagent strips and insulin syringes (List 4 with $n_4 = 173$ cases). A total of $M = 2069$ cases were identified. Despite the

Table 1. Data on Hepatitis A virus[22].

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	$Z_{000} = ??$

DOI: 10.1002/9781118445112.stat04855.pub2

Table 2. Data on diabetes[26].

Diabetes list				
1	2	3	4	Data
1	1	1	1	$Z_{1111} = 58$
1	1	1	0	$Z_{1110} = 157$
1	1	0	1	$Z_{1101} = 18$
1	1	0	0	$Z_{1100} = 104$
1	0	1	1	$Z_{1011} = 46$
1	0	1	0	$Z_{1010} = 650$
1	0	0	1	$Z_{1001} = 12$
1	0	0	0	$Z_{1000} = 709$
0	1	1	1	$Z_{0111} = 14$
0	1	1	0	$Z_{0110} = 20$
0	1	0	1	$Z_{0101} = 7$
0	1	0	0	$Z_{0100} = 74$
0	0	1	1	$Z_{0011} = 8$
0	0	1	0	$Z_{0010} = 182$
0	0	0	1	$Z_{0001} = 10$
0	0	0	0	$Z_{0000} = ??$

active identification, Bruno *et al*. concluded that there were still some people who could not be identified. The purpose was then to estimate the number of missing diabetes patients and to adjust for undercount.

The basic assumptions include the following: (i) All individuals act independently. (ii) Interpretation or definition of the characteristic of the target population should be consistent for all data sources. (iii) For all sources, identification marks are correctly recorded and matched. (iv) The size of the population is approximately a constant during the study period. (v) Any individual must have a positive probability to be ascertained by any source; any nonascertainment is purely due to a small ascertainment probability rather than impossibility. (When a random sample is feasible, this assumption can be relaxed and some special types of structural zeros are permitted; see Section 3.)

Traditional approaches assume that the samples are independent (*see* **Statistical Independence**). As individuals can be cross classified according to their capture or noncapture in each list, the independence for two samples is usually interpreted from a **Two by Two Contingency Tables** in human applications[27]. This independence assumption in animal studies is expressed in terms of the more restrictive "equal-catchability assumption": all animals have the same probability of capture in each sample. In our context, this becomes an assumption of "equal-ascertainment probability". However, this assumption is rarely valid in most applications. Lack of independence among samples leads to a bias (correlation bias) for the usual estimators derived under the independence assumption. The correlation bias may be caused by the following two sources[22]:

1. Local dependence among lists 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.
2. Heterogeneity among 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 multiple independent 2×2 tables might result in a dependent table. Hook and Regal[28] provided an example.

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat04855.pub2

The above two types of dependences are usually confounded (*see* **Confounding**) and cannot be easily disentangled in a data analysis without further assumptions or without more information.

3 Two-Sample Analysis

The two-sample case is the origin of capture–recapture methodology. Even if there are more than two lists, the analysis of any pair of lists provides useful preliminary information about the dependence between samples; see later. The intuitive idea for two-sample analysis in animal populations is the following[2]: Assume that a first sample of n_1 animals is captured, marked, and released back to the population. Thus, the marked rate in the population is n_1/N, with N unknown. A second sample of n_2 animals is subsequently drawn and there are m_2 previously marked. Equating the proportion of the marked rate in the population to the marked rate in the second sample suggests that $m_2/n_2 \approx n_1/N$, which yields the following *Petersen estimator* (or the *Petersen–Lincoln estimator*) for the population size:

$$\widehat{N}_{\mathrm{P}} = n_1 n_2 / m_2 \tag{1}$$

On the basis of a hypergeometric model (in which n_1 and n_2 are regarded as fixed), *Chapman*[29] derived the following estimator to adjust the bias that arises mainly because of a small value of m_2:

$$\widetilde{N} = (n_1 + 1)(n_2 + 1)/(m_2 + 1) - 1 \tag{2}$$

Under the same hypergeometric model, both estimators have approximately the same variance given by

$$\mathrm{Var}\widetilde{N} \approx (n_1 + 1)(n_2 + 1)(n_1 - m_2)(n_2 - m_2)/[(m_2 + 1)^2(m_2 + 2)] \tag{3}$$

As the Petersen and Chapman estimates are typically skewed, a log transformation has been used to obtain a confidence interval for population size[30]. For example, for the Chapman estimator, we assume that $\log(\widetilde{N} - M)$ follows a normal distribution, implying the 95% confidence interval for the Chapman estimator with an estimated variance given in Equation (3) can be constructed as follows:

$$[M + (\widetilde{N} - M)/C, M + (\widetilde{N} - M) \times C] \tag{4}$$

where

$$C = \exp\left\{1.96\sqrt{\log\left[1 + \frac{\mathrm{var}\,\widetilde{N}}{\left(\widetilde{N} - M\right)^2}\right]}\right\}$$

The lower bound of this interval is always greater than the number of different animals actually captured in the experiment. The confidence interval can be applied not only to the Chapman estimator but also to any other population size estimators.

A critical assumption for the validity of the Petersen and Chapman estimators is that the two samples are independent. As described in Section 2, local dependence among lists and unequal catchabilities among individuals are two sources of dependences that lead to correlation bias. For example, if the two samples are positively correlated (because animals exhibit a trap-happy behavioral response or if individual heterogeneity exists and is consistent over sampling occasions), then the animals captured in the first sample are more easily caught in the second sample. Thus, the recapture rate (m_2/n_2) in the second sample tends to be larger than the true proportion of marked animals in the population n_1/N. Then it is expected that $m_2/n_2 > n_1/N$, which yields $N > n_1 n_2/m_2 = \widehat{N}_{\mathrm{P}}$. As a result, the Petersen estimator tends to underestimate the true

DOI: 10.1002/9781118445112.stat04855.pub2

size. Conversely, it tends to overestimate when two samples are negatively correlated. Similar arguments and conclusions are also valid for more than two capture occasions. The bias direction has an important implication: if the Petersen or Chapman estimates for any two samples are relatively high (low) compared to other pairwise estimates, then it indicates that the two samples are negatively (positively) dependent. The bias is quantified in Section 4.2 in terms of a dependence measure between the two samples.

When only two lists are available, three cells are observable: people identified in List 1 only, people identified in List 2 only, and people listed in both. However, there are four parameters: N, two mean capture probabilities, and a dependence measure. The data are insufficient for estimating dependence unless additional information or covariates are available. All existing methods for two lists unavoidably encounter this problem and adopt the independence assumption. This independence assumption has become the main weak point in the use of the capture–recapture method for two lists, that is, the Petersen estimator is subject to correlation bias (unless in some special cases as described later). Generally, at least three lists are required to model dependences among lists.

If one of the two lists can be obtained by a random manner (i.e., all individuals in the target population have approximately the same probability to be ascertained in that list), then heterogeneity among individuals in the other sample would not cause any correlation bias to the Petersen estimator; see Section 4.2 for a justification. In this case, the other list can be highly selective or heterogeneous with structural zeros.[22,27] Further, if there is a temporal ordering for the two lists (say, the second list is obtained after the first list), then correlation bias also vanishes when the second list is random. This is because in such a case the homogeneity of the ascertainment probabilities of the second sample implies no local dependence. All these can be intuitively explained by animal experiments[22] and justified by theory in Section 4.2.

When there are more than two lists, a variety of models incorporating dependence among samples have been proposed in the literature. Generally, there are three classes of models, the ecological model, the **Log-linear Models in Contingency Tables**, and the sample coverage approach, that allow for the above two types of dependences. We focus in Section 4 on the sample coverage models and briefly discuss the other two classes of models in Section 5 because the two classes are reviewed in other articles (see Related Articles).

4 Sample Coverage Approach (>2 Samples)

4.1 Sample Coverage: Quantifying Overlapping Fraction

For multiple-record systems, the overlapping information among lists, like the recapture proportion in animal populations, provides essential information in inferring population size. The sample coverage approach was motivated to provide population size estimators based on measures that quantify the extent of *overlapping information* and also the *dependence* among lists. This approach was proposed by Chao and Tsay[31] for the three-list case. The extension to a general case was presented by Tsay and Chao[32] and in a tutorial article with examples[22].

The concept of *sample coverage* (or simply "coverage") for a single sample was originally developed for cryptographic analyses during World War II by the founder of modern computer science, Alan Turing, and by his colleague I. J. Good[33,34]. The concept has been widely applied to the estimation of species richness[35] and animal abundances[7,9]. It was also adapted to the context of multiple records systems[22,31,32] to quantify objectively the extent of overlapping information among lists. For multiple samples, a more proper term for "sample coverage" is *overlapping fraction* or *joint coverage of samples* as will be clear in the formulation of the concept. However, in order to be consistent with previous papers, we still retain the use of the term sample coverage in the following review, but will use this and overlapping fraction interchangeably.

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat04855.pub2

The ascertainment data for all individuals can be expressed as a matrix (X_{ij}), where $X_{ij} = I[$the ith individual is captured in Sample $j]$, $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, t$, and $I[\bullet]$ denotes an indicator function. We assume for the ith individual that the probability of the records $(X_{i1}, X_{i2}, \ldots, X_{it})$ is determined by its ascertainment probability. Let P_{ij} be the ascertainment probability of the ith individual in the jth list given the records of other lists. Here, P_{ij} may depend on the records of other lists if local dependence exists within individual i. Consider the three-list case first. The sample coverage (or overlapping fraction) of List 1 given the records of the other lists, $C(L_1|L_2, L_3)$, is defined as the probability that a case which can be potentially recorded by the ascertainment method for List 1 is shared with at least one of the other given lists. The mathematical expression is

$$C(L_1|L_2, L_3) = \frac{\sum_{i=1}^{N} P_{i1} I(X_{i2} + X_{i3} > 0)}{\sum_{i=1}^{N} P_{i1}} \tag{5}$$

This quantifies the joint coverage or overlapping information between List 1 and the merged list of Lists 2 and 3. The sample coverage of all three available lists is defined as the average of the sample coverages of individual lists:

$$C = \frac{1}{3}[C(L_1|L_2, L_3) + C(L_2|L_1, L_3) + C(L_3|L_1, L_2)] \tag{6}$$

Contrary to most people's intuition, multi-list sample coverage can be very accurately and efficiently estimated using only information contained in the sample itself[31], as long as the sample is reasonably large. Consider the estimation of $C(L_1|L_2, L_3)$ in Equation (5) in the following intuitive way. Any case in List 1 must be in one of the four categories (100), (101), (110), and (111). All cases in the latter three categories overlap with at least one of the other lists, implying that the overlapping fraction of List 1 can be estimated as $(Z_{110} + Z_{101} + Z_{111})/n_1 = 1 - Z_{100}/n_1$. Similar estimators can be obtained for Lists 2 and 3. Thus, the estimated sample coverage is expressed as an average (over three lists) of the three fractions:

$$\widehat{C} = \frac{1}{3}\left[\left(1 - \frac{Z_{100}}{n_1}\right) + \left(1 - \frac{Z_{010}}{n_2}\right) + \left(1 - \frac{Z_{001}}{n_3}\right)\right] = 1 - \frac{1}{3}\left[\frac{Z_{100}}{n_1} + \frac{Z_{010}}{n_2} + \frac{Z_{001}}{n_3}\right] \tag{7}$$

Note that Z_{100}, $Z_{010,}$ and Z_{001} are the numbers of individuals recorded only in one list (i.e., singletons). Singletons do not contain any overlapping information. The second expression in Equation (7) shows that sample coverage estimator is the one-complement of the average of fractions of singletons. Similar estimators can be obtained for the general t-sample cases.

4.2 CCV Measures: Quantifying Dependences among Samples

In the sample coverage approach, dependence among samples is modeled by the *coefficient of covariation* (CCV) for two or more samples. This is an extension of the "coefficient of variation" of one sample to multiple samples. The dependence measure CCV between Samples j and k is defined as

$$\gamma_{jk} = \frac{1}{N}\sum_{i=1}^{N} \frac{E[(X_{ij} - \mu_j)(X_{ik} - \mu_k)]}{\mu_j \mu_k} = \frac{1}{N}\sum_{i=1}^{N} \frac{E(X_{ij}X_{ik})}{\mu_j \mu_k} - 1 \tag{8}$$

where $\mu_j = E(n_j)/N$ denotes the mean probability of being listed in the jth sample. The magnitude of γ_{jk} measures the *degree of dependence* between Samples j and k. Two samples are independent if and only if $\gamma_{jk} = 0$. Two samples are positively (negatively) dependent if $\gamma_{jk} > 0$ ($\gamma_{jk} < 0$). From the second expression in Equation (8), positive (negative) dependence means that the average probability of jointly being listed in the two samples is greater (less) than that in the independent case (i.e., $\mu_j \mu_k$).

DOI: 10.1002/9781118445112.stat04855.pub2

The relative bias of the Petersen estimator (bias divided by the estimate) for Samples j and k is approximately $-\gamma_{jk}$[20,31]. That is, we can quantify the correlation bias of the Petersen estimator:

$$\text{Correlation bias} = E(\widehat{N}_{\text{P}}) - N \approx -\gamma_{jk}\ E(\widehat{N}_{\text{P}}) \tag{9}$$

This gives a theoretical justification that the Petersen estimator overestimates when two samples are negatively correlated, whereas it underestimates when two samples are positively correlated, as stated in Section 3. Moreover, consider the special case that there is no local dependence. Let the two sets of probabilities, $\{P_{ij};\ i=1,\ 2,\ \ldots,\ N\}$ and $\{P_{ik};\ i=1,\ 2,\ \ldots,\ N\}$, denote the ascertainment probabilities of N individuals for Samples j and k, respectively. In this special case, the two heterogeneous samples are independent if and only if $\gamma_{jk}=0$, that is, $N^{-1}\sum_{i=1}^{N} P_{ij}P_{ik} = \mu_j\mu_k$, which means that the covariance between the two sets of probabilities is zero. Thus, if one sample is random so that the set of probabilities for that sample is homogeneous, then it suffices to assure independence of the two samples provided no local dependence exists; the other nonrandom sample can be heterogeneous with structural zeros.

The CCV measure can be similarly defined for more than two samples. For example, the CCV measure for three samples, j, k, and m, is defined as

$$\gamma_{jkm} = \frac{1}{N}\sum_{i=1}^{N} \frac{E[(X_{ij}-\mu_j)(X_{ik}-\mu_k)(X_{im}-\mu_m)]}{\mu_j\mu_k\mu_m} \tag{10}$$

In the independent case, dependence measures of any number of samples are zero.

4.3 Coverage-based Population Size Estimators

Notice that if neither local dependence nor heterogeneity exists, then the sample coverage defined in Equation (6) reduces to $C = D/N$, where

$$D = \frac{1}{3}[(M - Z_{100}) + (M - Z_{010}) + (M - Z_{001})] = M - \frac{1}{3}(Z_{100} + Z_{010} + Z_{001}) \tag{11}$$

Here, $(Z_{100} + Z_{100} + Z_{100})/3$ represents the average of the non-overlapping cases. Thus, D can be interpreted as the average (over three lists) of the shared or overlapping cases. We summarize the following estimation procedures for the three-list case.

When the three sources are independent so that all dependence measures are zero, we have $N = E(D)/E(C)$, implying a simple population size estimator[31]:

$$\widehat{N}_0 = D/\widehat{C} \tag{12}$$

It can also be intuitively thought of as the ratio of overlapping cases to overlapping fraction. Here, the sample coverage estimator $\hat{C}$ is given in Equation (7).

When dependence exists among samples, Chao *et al.*[31] took into account the dependences and derived an adjustment formula to adjust the above simple estimator $\widehat{N}_0$ based on a function of two-sample CCVs. Inferences of population size depend on whether data contain sufficient overlapping information:

1. If data contain sufficient overlapping information, the following explicit size estimator is suggested with a bootstrap standard error (s.e.) described farther below:

$$\widehat{N} = \left[\frac{Z_{+11} + Z_{1+1} + Z_{11+}}{3\widehat{C}}\right] \div \left\{1 - \frac{1}{3\widehat{C}}\left[\frac{(Z_{1+0} + Z_{+10})\,Z_{11+}}{n_1 n_2} + \frac{(Z_{10+} + Z_{+01})Z_{1+1}}{n_1 n_3} + \frac{(Z_{0+1} + Z_{01+})Z_{+11}}{n_2 n_3}\right]\right\} \tag{13}$$

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat04855.pub2

Here, a rough guideline about the condition "sufficient overlapping information" is that the estimated sample coverage $\hat{C}$ should be at least 55% and the estimated bootstrap s.e. of $\widehat{N}$ is less than one-third of the population size estimate. A general property about the explicit estimator $\widehat{N}$ is that the correlation bias relative to N is negligible when there is no three-sample dependence[18]. When there is no local dependence, the relative correlation bias also vanishes if individual heterogeneity follows a gamma distribution, which covers a wide range of heterogeneity types[31].

2. For relatively low sample coverage (<55%) case or the estimated bootstrap s.e. of $\widehat{N}$ exceeds one-third of the population size estimate, data do not contain sufficient information to accurately estimate the population size. In this case, the following "one-step" estimator $\widehat{N}_1$ is suggested: (The estimator is called *one-step* because it is obtained by one iterative step from the adjustment formula[22])

$$\widehat{N}_1 = \frac{D}{\widehat{C}} + \frac{1}{3\widehat{C}}[(Z_{1+0} + Z_{+10})\hat{\gamma}_{12} + (Z_{10+} + Z_{+01})\hat{\gamma}_{13} + (Z_{01+} + Z_{0+1})\hat{\gamma}_{23}] \tag{14}$$

where the CCV estimates are

$$\hat{\gamma}_{12} = \widehat{N'}\frac{Z_{11+}}{n_1 n_2} - 1, \hat{\gamma}_{13} = \widehat{N'}\frac{Z_{1+1}}{n_1 n_3} - 1, \hat{\gamma}_{23} = \widehat{N'}\frac{Z_{+11}}{n_2 n_3} - 1 \tag{15}$$

and

$$\widehat{N'} = \frac{D}{\widehat{C}} + \frac{1}{3\widehat{C}}\left[(Z_{1+0} + Z_{+10})\left(\frac{D}{\widehat{C}} \cdot \frac{Z_{11+}}{n_1 n_2} - 1\right) + (Z_{10+} + Z_{+01})\left(\frac{D}{\widehat{C}} \cdot \frac{Z_{1+1}}{n_1 n_3} - 1\right) + (Z_{01+} + Z_{0+1})\left(\frac{D}{\widehat{C}} \cdot \frac{Z_{+11}}{n_2 n_3} - 1\right)\right]$$

This one-step estimator can be regarded as a lower (upper) bound for positively (negatively) dependent samples. Most data sets used in epidemiological applications tend to have a net positive dependence. Thus, the one-step estimator is often used as a lower bound.

The above three population size estimators ($\widehat{N}_0, \widehat{N}, \widehat{N}_1$) are referred to as the *sample coverage population size estimators*. A bootstrap method (*see* **Bootstrap with Examples**) was proposed[31] to obtain estimated s.e. for each of the above three estimators, and to construct the resulting 95% confidence interval using a log transformation as in Equation (4). A relatively low overlapping fraction means that there are many singletons, and consequently a large s.e. is usually associated with the estimator $\widehat{N}$ in Equation (13).

The estimation procedure and related properties for the general t-sample case are parallel to those for the three-sample case[22,32]. All coverage-based population size estimators and statistics can be obtained from the R package CARE1; see Section 6.

5 Other Models

5.1 Ecological Models

Pollock proposed a sequence of models mainly for wildlife studies to relax the equal-catchability assumption[36]. This approach models 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 behavioral responses to capture; and (iii) model M_h, which allows heterogeneous animal

DOI: 10.1002/9781118445112.stat04855.pub2

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. Local 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} are useful for epidemiological studies. Various estimation procedures have been proposed[2,7,11−13,36].

As the CCV measures defined in Section 4.2 take into account both local dependences among lists and heterogeneity among individuals, the sample coverage approach conceptually encompasses all ecological models as special cases in model formulation. However, the inference procedures for the two approaches are different. Under ecological models, natural time ordering or sequential ordering is incorporated, whereas no ordering is considered in the sample coverage approach. Estimators for some ecological models can be computed from the R package SpadeR (species prediction and diversity estimation in R) which can be freely downloaded from the author's website at *http://chao.stat.nthu.edu.tw/softwareCE.html*.

5.2 Loglinear Models

The loglinear model approach is a commonly used technique for analyzing discrete data. Loglinear models that incorporate list dependence were first proposed by Fienberg[37] for dealing with human populations. Cormack[38] proposed the use of this technique for several ecological models.

In this approach, the data are regarded as a form of an incomplete 2^t **Contingency tables** (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. For three lists, the most general model is a model with three 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 **Model Selection: Akaike's Information Criterion**. The chosen model is then projected onto the unobserved cell to obtain the number of missing cases.

Local 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, that is, **Rasch Model** and their generalizations[39]. As 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 two review papers[17,18].

Model selection in the loglinear models may be difficult because two equally fitted models might produce quite different estimates. An adequate fit to the observed cells may not necessarily imply an adequate estimate for the count of the unobservable cell. Also, the existence of heterogeneity in data might result in the lack of a reliable estimate[17]. As the number of lists increases, the number of adequate models increases rapidly and thus model selection is a problem. In contrast, no model selection or model comparison is needed in the sample coverage approach; no further difficulty arises when the number of lists increases.

6 Examples

The R package CARE1, available in the CRAN (Comprehensive R Archive Network) (*http://cran.r-project.org/web/packages/*) and also in the author's website, can be used to analyze epidemiological data based on t incomplete lists of individuals, $t \geq 2$. For $t = 2$, CARE1 computes the Petersen and Chapman

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat04855.pub2

estimator along with s.e. and confidence intervals. For $t > 2$, CARE1 provides three coverage-based population size estimators and related statistics; the output also includes the results for any pair of samples. For $t > 5$, it may take longer running time due to the bootstrap procedures. Below we demonstrate the application of CARE1 to two data sets (given in Tables 1 and 2) and interpret the results.

6.1 HAV Data (Low Overlapping Fraction)

The following steps show how to run CARE1 in R for three-list HAV data (given in Table 1) and the output.

```
# Install CARE1 package from CRAN
install.package("CARE1")

# Import CARE1
library(CARE1)

# Import HAV categorical data
df <- data.frame(P=c(1,1,1,1,0,0,0),
                 Q=c(1,1,0,0,1,1,0),
                 E=c(1,0,1,0,1,0,1),
                 data=c(28,21,17,69,18,55,63))

# View import data
df
  P Q E data
1 1 1 1   28
2 1 1 0   21
3 1 0 1   17
4 1 0 0   69
5 0 1 1   18
6 0 1 0   55
7 0 0 1   63

# Transform summarized data to individual data
dat <- df[rep(1:nrow(df), time=df$data),-4]

# Transform observed data to CARE1 input format
HAV <- as.record(dat)

# Print the CARE1 input format
HAV

001 010 011 100 101 110 111
 63  55  18  69  17  21  28

# Main step to obtain population size estimates
CARE1.print(HAV)

# Output (three parts)
(1) NUMBER OF IDENTIFIED CASES IN EACH LIST:
 n1  n2  n3
135 122 126
```


DOI: 10.1002/9781118445112.stat04855.pub2

```
(2) ESTIMATES BASED ON ANY PAIR OF SAMPLES:
      Petersen Chapman se cil ciu
pair12      336     334 29 289 403
pair13      378     374 36 319 461
pair23      334     331 30 285 404

Note 1: Refer to Seber (1982, pages 59 and 60) for Petersen estimator
       and Chapman estimators as well as s.e. formula.
Note 2: A log-transformation is used is used to obtain the confidence
       interval so that the lower limit is always greater than the
       number of ascertained; see Chao (1987, Biometrics, 43, 783-791)
       for the construction of the confidence interval.

(3) SAMPLE COVERAGE APPROACH:
        M       D  Chat est   se cil   ciu
Nhat-0 271 208.667 0.513 407   26 365   467
Nhat   271 208.667 0.513 971  688 411  3778
Nhat-1 271 208.667 0.513 508   53 425   636

Warning: The estimated sample coverage (overlapping fraction is too)
low, so Nhat is unstable.
Warning: The estimated bootstrap s.e. of Nhat exceeds one-third of the
population size estimate, so Nhat is unstable.

Parameter estimates:
         u1   u2   u3  r12  r13  r23
Nhat-0 0.33 0.30 0.31 0.21 0.08 0.22
Nhat   0.14 0.13 0.13 1.89 1.57 1.91
Nhat-1 0.27 0.24 0.25 0.51 0.34 0.52


Definitions for the sample coverage approach:

M: number of individuals ascertained in at least one list.
D: the average (over all lists) of the overlapping cases.
Chat: sample coverage estimate, see Eq. (14) of Chao et al. (2001).
est: population size estimate.
se: estimated standard error of the population size estimate based on
    the bootstrap method.
cil: 95% confidence interval lower limit (using a log-transformation).
ciu: 95% confidence interval upper limit (using a log-transformation).
Nhat-0: population size estimate under independence assumption.
Nhat: population size estimate for sufficiently high sample coverage
     cases; see Eq. (16) of Chao et al. (2001).
Nhat-1: one-step population size estimate for low sample coverage
cases; see Eq. (17) of Chao et al. (2001).
u1,u2,u3 etc.: estimated mean ascertainment probabilities depending on
the estimate of N.
r12,r13,r23 etc.: estimated coefficient of covariation (CCV) depending
       on the estimate of N.
```

The first part of the output gives the number of identified individuals in each of the three lists. The second part gives the Petersen and Chapman estimates along with s.e. and confidence intervals for any pair of lists. These estimates can be used as preliminary analysis to detect possible dependence among lists. As described in Sections 3 and 4, if a Petersen or Chapman estimate for two samples is relatively high (low) compared to other pairwise estimates, then it reveals that the two samples are negatively (positively) dependent. However, for the HAV data set, the Petersen and Chapman estimates for the three pairs of

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat04855.pub2

lists are in the range of 331 to 378. The narrow range of these estimates would not indicate the possible direction of dependence at this stage.

The third part gives the coverage-based population size estimators along with related statistics. The estimated sample coverage or overlapping fraction is $\hat{C} = 51.3\%$ (Chat in the output), which is considered to be low. The average (over three lists) of the overlapping cases is $D = 208.667$. If independence among samples is assumed, then a population size estimate would be $\widehat{N}_0 = D/\widehat{C} = 407$ (Nhat-0 in the output), with a bootstrap s.e. 26 based on 1000 bootstrap replications. The 95% confidence interval lower limit (cil) is 365 and the upper limit (ciu) is 467. Even with the same input data, the bootstrap s.e. estimate and confidence interval are likely to be different for replicated runs of CARE1 due to resampling variation in the bootstrap procedures.

Incorporating the dependences between any two samples, we have $\widehat{N} = 971$ (Nhat in the output), but a large estimated bootstrap s.e. (688) renders the estimate useless. This extremely large s.e. also signifies that these data with a relatively low sample coverage do not contain enough information to correct for undercount or to provide an accurate population size estimate, and thus at best we can only provide a minimum or maximum number for the population size.

CARE1 provides the estimated mean ascertainment probabilities μ_1, μ_2, and μ_3 (u1, u2, and u3 in the output) as well as the estimated CCV measures γ_{12}, γ_{13}, and γ_{23} (r12, r13, and r23 in the output) for all pairs of samples. All these estimates depend on the value of N, and thus their estimates are given for each of the population size estimates. Regardless of population size estimates, all numerical values consistently show that the three lists have approximately the same mean ascertainment probabilities, and the CCV estimates reflect positive dependence for any pair of lists. Consequently, the estimator $\widehat{N}_0$ under the independence assumption would generally underestimate. We recommend the use of $\widehat{N}_1 = 508$ (Nhat-1 in the output) as a lower bound, with an estimated s.e. of 53 with a 95% confidence interval (425, 636) based on 1000 bootstrap replications.

After the three surveys, the National Quarantine Service of Taiwan conducted a screen serum test for the HAV antibody for all students of the college at which the HAV outbreak occurred. The conclusive final figure of the number infected was about 545. Thus, this example presents a very valuable data set with the advantage of a known true parameter. Our estimator $\widehat{N}_1 = 508$ provides a satisfactory lower bound. This example shows the need for undercount correction and also the usefulness of the capture–recapture method in estimating the number of missing cases.

6.2 Diabetes Example (High Overlapping Fraction)

The following steps show how to run CARE1 in R for four-list Diabetes data (given in Table 2) and the output. The explanation of the notation included in the output is omitted; see the HAV example for details.

```
library(CARE1)


# Import diabetescategoricaldata
df <- data.frame (V1=c(1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0),
                  V2=c(1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0),
                  V3=c(1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0),
                  V4=c(1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1),
                  data=c(58,157,18,104,46,650,12,709,14,20,7,74,8,182,10))
# View import data
df
```


DOI: 10.1002/9781118445112.stat04855.pub2

```
   V1 V2 V3 V4 data
1   1  1  1  1   58
2   1  1  1  0  157
3   1  1  0  1   18
4   1  1  0  0  104
5   1  0  1  1   46
6   1  0  1  0  650
7   1  0  0  1   12
8   1  0  0  0  709
9   0  1  1  1   14
10  0  1  1  0   20
11  0  1  0  1    7
12  0  1  0  0   74
13  0  0  1  1    8
14  0  0  1  0  182
15  0  0  0  1   10
```

```
# Transform summarized data to individual data
dat <- df[rep(1:nrow(df), time=df$data), -5]

# Transform observed data to CARE1 input format
diabetes <- as.record(dat)

# Print the CARE1 input format
diabetes
```

```
0001 0010 0011 0100 0101 0110 0111 1000 1001 1010 1011 1100 1101 1110 1111
  10  182    8   74    7   20   14  709   12  650   46  104   18  157   58
```

```
# Main step to obtain population size estimates
CARE1.print (diabetes)

# Output (three parts)
(1) NUMBER OF IDENTIFIED CASES IN EACH LIST:
  n1   n2   n3   n4
1754  452 1135  173

(2) ESTIMATES BASED ON ANY PAIR OF SAMPLES:
      Petersen Chapman se  cil  ciu

pair12     2353    2351 58 2250 2478
pair13     2185    2185 22 2146 2233
pair14     2264    2261 88 2117 2468
pair23     2060    2057 77 1922 2224
pair24      806     803 47  725  913
pair34     1558    1555 67 1445 1712

(3) SAMPLE COVERAGE APPROACH:
          M       D  Chat  est se  cil  ciu
Nhat-0 2069 1825.25 0.803 2272 25 2228 2328
Nhat   2069 1825.25 0.803 2609 78 2477 2784
Nhat-1 2069 1825.25 0.803 2458 50 2372 2568

Parameter estimates:
         u1   u2   u3   u4   r12  r13  r23  r14  r24  r34
Nhat-0 0.77 0.20 0.50 0.08  0.03 0.04 0.10 0.00 1.82 0.46
Nhat   0.67 0.17 0.44 0.07  0.11 0.19 0.27 0.15 2.24 0.67
Nhat-1 0.71 0.18 0.46 0.07  0.04 0.12 0.19 0.09 2.05 0.58
```


Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat04855.pub2

The interpretation of all output is similar to that of the HAV data. We only briefly summarize the conclusions. From the second part of the output, the Petersen and Chapman estimates for two pairs of lists, (2, 4) and (3, 4), are substantially lower than the other pairwise estimates, signifying strong positive dependence between Lists 2 and 4, and also Lists 3 and 4. This is also reflected by the relatively large CCV estimates for γ_{24} and γ_{34} (r24 and r34 in the output). Thus, the population size estimator under the independence assumption (Nhat-0 in the output) is not proper.

The sample coverage for these data is estimated to be 80.3%. As the coverage estimate is sufficiently high, an accurate population size estimate is expected. The recommended estimate is $\widehat{N} = 2609$ (Nhat in the output) with an estimated s.e. of 78 using 1000 bootstrap replications. The corresponding 95% confidence interval is (2477, 2784). See Chao *et al*.[22] for comparisons of these results with those obtained from the loglinear models. The two approaches have different assumptions and different advantages and limitations. A common limitation is that sufficient overlapping information is required to yield reliable population size estimates. Otherwise, only a lower bound (for generally positively correlated samples) is useful as in the HAV data.

Related Articles

Capture—Recapture Methods; **Capture – Recapture Sampling Designs**; **Capture – Recapture Models, Spatially Explicit**; **Capture—Recapture Methodology**.

References

[1] Cochran, W.G. (1978) Laplace's ratio estimators, in *Contributions to Survey Sampling and Applied Statistics* (ed A. David), Academic Press, New York, pp. 3–10.
[2] Seber, G.A.F. (1982) *The Estimation of Animal Abundance*, 2nd edn, Griffin, London.
[3] Hald, A. (1990) *A history of probability and statistics and their applications before 1750*, Wiley, New York.
[4] Schnabel, Z.E. (1938) The estimation of the total fish population of a lake. *Am. Math. Mon.*, **45**, 348–352.
[5] Darroch, J.M. (1958) The multiple recapture census I. Estimation of a closed population. *Biometrika*, **45**, 343–359.
[6] Borchers, D.L., Buckland, S.T., and Zucchini, W. (2002) *Estimating Animal Abundance: Closed Populations*, Springer, London.
[7] Chao, A. (2001) An overview of closed capture–recapture models. *J. Agric. Biol. Environ. Stat.*, **6**, 158–175.
[8] Chao, A. and Huggins, R.M. (2005) Classical closed-population capture–recapture models, in *The Handbook of Capture–Recapture Methods* (eds S. Amstrup, T. McDonald, and B. Manly), Princeton University Press, Princeton, pp. 22–35.
[9] Chao, A. and Huggins, R.M. (2005) Modern closed-population capture–recapture models, in *The Handbook of Capture–Recapture Methods* (eds S. Amstrup, T. McDonald, and B. Manly), Princeton University Press, Princeton, pp. 58–87.
[10] McCrea, R.S. and Morgan, B.J.T. (2014) *Analysis of Capture—Recapture Data*, CRC Press, Boca Raton, Florida.
[11] Schwarz, C.J. and Seber, G.A.F. (1999) A review of estimating animal abundance III. *Stat. Sci.*, **14**, 427–456.
[12] Seber, G.A.F. (1986) A review of estimating animal abundance. *Biometrics*, **42**, 267–292.
[13] Seber, G.A.F. (1992) A review of estimating animal abundance II. *Int. Stat. Rev.*, **60**, 129–166.
[14] LaPorte, R.E., McCarty, D.J., Tull, E.S., and Tajima, N. (1992) Counting birds, bees, and NCDs. *Lancet*, **339**, 494–495.
[15] Hook, E.B. and Regal, R.R. (1992) The value of capture–recapture methods even for apparent exhaustive surveys. *Am. J. Epidemiol.*, **135**, 1060–1067.
[16] Hook, E.B. and Regal, R.R. (1995) Capture–recapture methods in epidemiology: methods and limitations. *Epidemiol. Rev.*, **17**, 243–264.
[17] International Society for Disease Monitoring and Forecasting (1995) Capture–recapture and multiple-record systems estimation I: history and theoretical development. *Am. J. Epidemiol.*, **142**, 1047–1058.
[18] International Society for Disease Monitoring and Forecasting (1995) Capture–recapture and multiple-record systems estimation II: applications in human diseases. *Am. J. Epidemiol.*, **142**, 1059–1068.
[19] El-Khorazaty, M.N., Imery, P.B., Koch, G.G., and Wells, H.B. (1977) A review of methodological strategies for estimating the total number of events with data from multiple-record systems. *Int. Stat. Rev.*, **45**, 129–157.

DOI: 10.1002/9781118445112.stat04855.pub2

[20] Sekar, C. and Deming, W.E. (1949) On a method of estimating birth and death rates and the extent of registration. *J. Am. Stat. Assoc.*, **44**, 101–115.
[21] Wittes, J.T. and Sidel, V.W. (1968) A generalization of the simple capture–recapture model with applications to epidemiological research. *J. Chronic Dis.*, **21**, 287–301.
[22] Chao, A., Tsay, P.K., Lin, S.H., *et al.* (2001) The applications of capture-recapture models to epidemiological data. *Stat. Med.*, **20**, 3123–3157.
[23] Böhning, D. (2008) Recent development in capture–recapture methods and their applications. *Biom. J.*, **50**, 954–956.
[24] Böhning, D., Suppawattanabodee, B., Kusolvisitkul, W., and Viwatwongkasem, C. (2004) Estimating the number of drug users in Bangkok 2001: a capture–recapture approach using repeated entries in one list. *Eur. J. Epidemiol.*, **19**, 1075–1083.
[25] Van der Heijden, P.G.M., Cruyff, M., and Böhning, D. (2014) Capture–recapture to estimate crime populations, in *Encyclopedia of Criminology and Criminal Justice* (eds G.J.N. Bruinsmaand and D.L. Weisburd), Springer, Berlin, pp. 267–278.
[26] Bruno, G.B., Biggeri, A., LaPorte, R.E., *et al.* (1994) Application of capture–recapture to count diabetes. *Diabetes Care*, **17**, 548–556.
[27] Chao, A., Pan, H.Y., and Chiang, S.C. (2008) The Petersen–Lincoln estimator and its extension to estimate the size of a shared population. *Biom. J.*, **50**, 957–970.
[28] Hook, E.B. and Regal, R.R. (1993) Effects of variation in probability of ascertainment by sources ("variable catchability") upon capture–recapture estimates of prevalence. *Am. J. Epidemiol.*, **137**, 1148–1166.
[29] Chapman, D.G. (1951) Some properties of the hypergeometric distribution with applications to zoological censuses. *Univ. Calif. Publ. Stat.*, **1**, 131–160.
[30] Chao, A. (1987) Estimating the population size for capture–recapture data with unequal catchability. *Biometrics*, **43**, 783–791.
[31] Chao, A. and Tsay, P.K. (1998) A sample coverage approach to multiple-system estimation with application to census undercount. *J. Am. Stat. Assoc.*, **93**, 283–293.
[32] Tsay, P. and Chao, A. (2001) Population size estimation for capture–recapture models with applications to epidemiological data. *J. Appl. Stat.*, **28**, 25–36.
[33] Good, I.J. (1953) On the population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.
[34] Good, I.J. (2000) Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval enigma. *J. Stat. Comput. Simul.*, **66**, 101–111.
[35] Bunge, J. and Fitzpatrick, M. (1993) Estimating the number of species: a review. *J. Am. Stat. Assoc.*, **88**, 364–373.
[36] Pollock, K.H. (1991) Modelling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future. *J. Am. Stat. Assoc.*, **86**, 225–238.
[37] Fienberg, S.E. (1972) The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, **59**, 591–603.
[38] Cormack, R.M. (1989) Loglinear models for capture–recapture. *Biometrics*, **45**, 395–413.
[39] Darroch, J.N., Fienberg, S.E., Glonek, G.F.V., and Junker, B.W. (1993) A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *J. Am. Stat. Assoc.*, **88**, 1137–1148.

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat04855.pub2