

MAKING CAPTURE-RECAPTURE VIABLE

Collin Carbo¹

ABSTRACT

In the absence of an exhaustive enumeration of individuals, the problem of determining the size a finite population is a difficult one. In this paper, capture-recapture approaches are investigated from the perspectives of the "nature of the capture process" and "process of decoding unequal capture probabilities." Illustrative mathematical techniques are developed for analyzing capture-recapture data using various "capturability" distributions.

KEY WORDS: Capture-Recapture; Non-Random Sampling; Capturability distributions.

RÉSUMÉ

En l'absence d'une énumération exhaustive des individus, le problème de la détermination de la taille inconnue d'une population finie est très difficile à résoudre. Dans cet article, les méthodes de saisie-resaisie sont évaluées dans le contexte de la "nature du procédé de saisie" et "du procédé de décodage des probabilités inégales de saisie." Des techniques mathématiques et explicatives sont développées afin d'analyser les données de saisie-resaisie en utilisant diverses distributions de "capturabilité".

MOTS CLÉS: Capture-Recapture; Échantillonnage non-aléatoire; Distributions de "capturabilité".

1. INTRODUCTION

1.1 Description of the Problem

The problem of the estimation of the number of individuals in a finite population is difficult. In many cases it is impractical to do an exhaustive search(count) over whole or even some partition of the population. Furthermore, with an unknown size of the population it is usually difficult to obtain a random sample. Thus, captures are often the only practical tool available for analyzing some population. Capture and Recapture methods (CRM) specifically are characterized by the use or comparison of two or more non-random samples or captures. In the most common scenario of capture-recapture, a group of animals (individuals) is captured, marked and returned to the population. Then, at a later time, a new capture of animals is taken. The number of marked to unmarked animals observed is then used to estimate the total number of animals in the population. The basic naive CRM assumes equal probability of capture by each individual in the population in each sample

(i.e. random samples), and a fixed total population between samples. In practice these assumptions are usually false in a number of ways: (1) marked animals may become wary of traps, (2) some animals may be more easily captured, (3) animals may immigrate/migrate or be born/die between samples. The many papers on capture-recapture methods deal with various extensions or relaxation's of the assumptions of the "naive" capture-recapture.

In particular, relaxing the assumption of homogeneous capture probabilities is particularly vexing because it means that capture-recapture methodologies must deal with "captures", i.e. non-random samples. Conceptually, the difficulty can partly understood from thinking of the repeat captures of individuals as being driven by two factors, namely, the finite population size and higher capturability of particular individuals. Uncoupling these different effects turns out to be difficult.

Indeed, in this paper, it is argued that capture-recapture methods give rise to ill-posed inverse

¹Collin Carbo 40 Dunning Crescent, Regina, S4S 3W1, collin.carbo@sk.sympatico.ca

mathematical problems. An inverse problem is a problem in which one is looking for the causes from some observed effects. It often happens that the effects are insufficient without additional information to determine the causes hence the problem is ill-posed. On first thought the capture-recapture methods might seem to provide a well-defined inverse mapping in that the observed capturability distribution could stand as an estimate for unknown actual capturability distribution allowing one in some way to deduce both total population and capturabilities at the same time. However, the situation is actually more tricky than it might at first appear. For example, even in the case where all the individuals have equal capturability some individuals will by chance be captured more than others so that translation from the observed capturability distribution to the actual capturability distribution is indeed a non-trivial task.

So, the critics of capture-recapture have considerable foundation for their position that capture-recapture methods are not viable in that the CRM produce invalid and unreliable estimates. However, many ill-posed inverse problems can with the addition some information become solvable and provide valuable information. How capture-recapture methods can be made viable in a similar fashion is the focus of this paper.

1.2 Classifying Non-homogenous Capture-Recapture: Plant, Mark, List

Historically, the CRM have also been classified into open methods that deal with population variation between samples and closed methods in which population is assumed to remain constant between captures. The methods have also been classified according to whether they assume homogeneous capture probabilities across the population, or non-homogeneous capture probabilities. The open extensions have received extensive development in the literature. While the problem of the non-homogeneous capture rates has also been addressed several times in the literature it is argued that the given techniques actually contain hidden assumptions that while not invalidating the techniques given do create confusion for the practitioner. In this paper capture-recapture are examined in case of a closed population with non-homogeneous capture probabilities although many of the findings can with suitable modifications be applied to the general capture-recaptures situations.

Making capture-recapture viable in the non-homogenous case requires clearly identifying the characteristics and nature of the capture process in question. Part of the confusion and danger in the non-homogeneous case is the that there exists several different situations that may require different mathematical formatulaton. In particular, the relationship of the two capture processes is critical. To better understand and develop capture-recapture methods, three scenarios: plant, mark and list are differentiated.

In the plant process, some individuals are planted into the population and then a capture is preformed. In this case, the first capture is actually "the planted individuals" and the second capture will consist hopefully of a mixture of planted and unplanted individuals from which one hopes to make meaningful estimates of the total population. Concerns naturally arise as to whether the planted individuals behavior is similar to the non-planted individuals in regard to the capture process. Perhaps the planted individuals are more easily captured. This concern arises if the planted individuals do not have the same attribute distributions as the native population in that populations with different attributes are likely captured with different probabilities.

The mark scenario was described above. The thing to remember in this case is that the probabilities of capture of an individual remain the same for each capture. This situation will be the default situation in the paper for the most part.

The list approach is substantially different in that it involves two different capture processes. Hence, this process could also be referred to as dual list case. In this case the probabilities of capture of an individual are different between the two capture processes.

These three situations are highly different in mathematical nature, and in the author's opinion part of the confusion in the practice of Capture-Recapture methods has been the attempt to apply identical mathematical formulations to these different situations.

Other mathematical scenarios of capture-recapture exist such as the exploratory case, where the capture-recapture probabilities of all the remaining unseen individuals capture probabilities change upon detection of a given individual. For example, the

detection of an AIDS patient may change the detection probability amongst sexual partners. Also, information from capture-recapture experiments can be presented at several different levels of information such as in terms of detailed capture histories, capture-frequencies, capture counts, or just as ranked lists of individuals (where the rank indicates whether captured more often or not).

2. CAPTURABILITY DISTRIBUTIONS

2.1 Capture Frequencies

Assume that the individuals of some population of size R all have a detectability (or capturability) which is constant over many captures for each individual but can vary from individual to individual. Further assume that in each capture event just one individual is captured. Now, imagine that one starts observing some unknown population by performing a series of captures. Early on, the observed captures produce mostly new individuals. As captures continue (with replacement) the number recaptures increases and eventually the captures mostly produce individuals seen before. Over many captures, some individuals will be captured many times, others a few times, and some not at all. If one assumes that each capture must catch some individual then the sum of (capture probabilities) detectabilities is one. Since the most observable individuals tend to be seen first, let us number the individuals from most observed downward to least observable, so that p_1 is the probability of seeing the most seen individual per capture and p_2 is probability of the second most frequently seen individual and so on.

Let f_k , known as the capture frequencies, be the number of individuals observed k times. Then, the total number of different individuals observed will be given by

$$\sum_{k=1}^{\infty} f_k = Q$$

The total population $R = Q + f_0$. But how many unobserved individuals f_0 will there be? Clearly, one can not answer this question directly since one has no idea how the "hidden" part of the population's detectability is structured. All we can really be sure is that on average, with L capture events, the expectation of the sum of detectabilities of the unseen individuals will satisfy

$$L \sum_{i=Q+1}^R p_i < 1$$

If the "hidden" population has arbitrarily small observabilities then the unseen population can be arbitrarily large. Thus, one can see that all estimates of the "hidden" population depend upon the assuming some form of the unseen populations detectabilities.

2.1 Insufficiency of the Capture Frequencies

There are in fact many possible capturability distributions that can be matched (with different degrees of fit) to the observed population catchabilities and with different extensions to the unobserved population. Does a set of capture frequencies contain enough information to determine the total population size R and of necessity also capturability distribution of the population? Clearly, this cannot be the case, as one can assume many possible extensions of the observed capturability distribution to the unseen population.

This raises a serious fundamental problem with the capture-recapture method. If different configurations of total population and capturability probabilities can yield the same capture frequencies how can one hope to make meaningful estimates? At first thought, one might understandably think that this makes for a hopeless situation and capture-recapture methods are intrinsically unviable. However, this not necessarily the case. Many interesting inverse problems in the mathematical sense are ill-posed yet still provide meaningful information. The secret of making the ill-posed problems well-posed is the addition of further information and properties of the problem at hand. In the capture-recapture case, the further assumptions must be that the capturability distribution cannot be arbitrary. In short, the capturability distribution or equally the capture process must possess certain properties before the capture-recapture methods can yield viable results.

But what properties must the capture have? First, it must be that every individual of the population has some finite probability of being seen under a capture. Second, the capturabilities over the population must change reasonably smoothly. In practice this is taken to mean that the capture distribution must be definable by a few parameters.

Given this additional constraint a reasonable approach to CRM estimation is to guess a capturability distribution that one believes mirrors the actual nature of the capture process and to check that it gives a reasonable match to the observed distribution of recaptures. Thus, under the assumption that the

capture process has reasonable “smooth” behavior, meaningful population estimates can be produced.

2.2 Plausible Capturability Distributions

Among the infinite number of possible forms for underlying population catchabilities distribution the following are finite distributions have been found to be useful.

- 1) **Constant p_i 's** -- all the members of the population are equally observable with probability p . The total population R is given by $R= 1/p$. This assumption leads to the usual “naive” capture-recapture formula.
- 2) **Uniformly distributed p_i** -- the observabilities of the population are uniformly distributed over some interval $[p_R, p_1]$ p_1 is the highest observability in the population, and p_R the lowest.
- 3) **Zipf Distribution** -- the observabilities of the population follow Zipf type distribution.

$$P_k = \frac{1}{k^\alpha H(R, \alpha)}$$

where

$$H(R, \alpha) = \sum_{m=1}^R \frac{1}{m^\alpha}.$$

- 4) **Normal distribution** -- could assume that the observabilities of the population are chosen from a normal distribution. Naturally, the normal population is a continuous probability distribution but it can be approximated by finite distribution selected from it.
- 5) **Log-Normal Distribution** - that the observabilities of the population are chosen from a log-normal distribution. Studies by the author of the capturabilities distributions of actual systems seem to indicate that the capture processes, often produce capturabilities that can be thought of as being selected from a log-normal distribution.

The critical features of any proposed capturability distribution are that it depends on only a few parameters and that it reasonably give results that match with the observed capture-frequencies.

2.3 Computing the Capture Frequencies

Given a capturability distribution how can one compute the capture frequencies expected? One clearly needs to be able to do this in order to make meaningful comparisons between possible capturability distributions. While, this would appear an extremely difficult problem, in practice it turns out to be fairly simple. Assuming, that the detectabilities are independent (which they are not), one has that the

expectation for the number of the individuals being captured k times as being given by

$$f_k = \sum_{\text{individuals}} \binom{L}{k} p_i (1 - p_i)^{L-i} \quad (2.2.1)$$

where the sum gives the contributions from each individual in the population and L is the total number of captures. This formula is simply the sum of the expectations for k -captures for each individual over the entire population. Simulations with different capturabilities distributions (set of values p_i) indicate that this approximation works extraordinarily well.

2.3 Parametrization of capture probabilities

The capturable probabilities are restricted by observed capture frequencies through equation 2.2.1. While there are infinitely many equations in (2.2.1) for which $f_k = 0$ there is a largest value of k say m for which (2.2.1) is non-zero. Thus, if the population is greater than m the capture probabilities cannot be unique determined from the capture frequencies. It is not therefore too surprising that different populations with different capturability distributions can perfectly match the observed capture frequencies. However, by demanding that the capturability distributions depend on a small finite number of parameters, one is limiting extensively the freedom so that any match between the observed capture frequencies and computed capture frequencies is meaningful. This situation is somewhat similar to the situation of fitting $n+1$ data points with an n -th order polynomial. While one can always perfectly fit n -th order polynomial through $n+1$ points doing so has really no significance. It can always be done. On the other hand, the ability to fit several hundred points reasonably close to a straight line is indicative that the fit has meaning.

The ability to parametrize the capturabilities is justified in the sense that it assumes that the population has reasonably smooth behavior with a meaningful distributions of capturability over some relevant attributes. This assumption of the ability to parameterize and capturability probability distributions is the key assumption in making capture-recapture methods viable.

3. CAPTURE-RECAPTURE LIST CASE

3.1 Dual List Relationship

The list case of capture-recapture creates a more difficult problem for estimation. Given two lists in

general it is impossible to determine whether the lists are dependent or independent. However, if one has knowledge of the actual different capture processes it should be clear whether the capture processes were independent. However, independence says nothing about whether the capture probabilities for capture process of list 1 and capture process of list 2 are related. Indeed, the capture probabilities between 1st and 2nd process may be correlated or anti-correlated or uncorrelated.

In the case of being presented with two lists of unknown origin, the difficulty that arises is that one has no way of knowing if the lists are independent (arose from different capture processes) or were copied in some way from one another. The observed correlation between the lists may be due to small population size, or might be due to the fact that one list was used as an aid in compiling the second list.

3.2 Common Individuals, Missing Individuals, and Rank Correlation Information

Given that we understand in detail how the lists were prepared, it follows that we should be able to tell to some extent know whether the capture processes in each list will independent or dependent. Given two different independent captures, one tool is to use the rank correlations methods. These methods have extensive use in statistical literature, and allow one to put statistical inference limits on the probability that the lists are correlated by chance or by some other causal effect. Once one determines that the correlation is small, then the number of individuals in common or missing between the lists can be used to create population estimates.

As before, it often useful to assume a particular capturability distribution for each capture process. Assume that each capturability distributions can be approximated by a Zipf distribution. Suppose that the Zipf orderings of the two lists are uncorrelated as given by the rank correlation methods. In this case the first list is essentially a random sample relative to the second list in the sense that probability that a member of first list is on the second list depends only on the size of the first list relative to the second list and total population.

However, what if the lists are correlated? Consider the extreme case, were the capture probabilities are identical on list 1 and list 2. If α , the Zipf power happens to be near zero, then the ordering on both lists will be close to random anyway, and the naive capture

recapture formula applies. If however α is not near zero and probabilities are correlated, first few items will have high probability of matching, but later items will not. So, provided that some fraction of both lists is not in common, it follows that uncommon portion of both lists can be considered essentially randomly ordered and the naive estimate can be used again provided that one works with the missing items rather than the items in common.

Suppose that the two lists are anti-correlated. In this case, if there is any overlap it must follow from the fact that the population is really very small. Thus, in the anti-correlated case the important feature between the two lists is the number of items in common rather than the number of items not in common.

3.2 Conclusions

Understanding the nature of the capture process and the nature of the capturability distribution is essential to capture-recapture methods. On what does the capture process select? Normally, in capture studies, considerable information is obtained about each captured individual. This list of attributes however, can not be used to obtain information about the overall population because the samples are non-random. Capturing heavier individuals for example tells us nothing about the average weight of individuals in the population. However, if we can construct capturability distribution curves based on attributes, and determine what attributes contribute to the capturability and in what way, it may be possible to determine the non-random nature of the samples, and to therefore correct the samples to obtain population characteristics.

One interesting possibility for obtaining improving capture-recapture methods would be to use a combination of plant, mark and list methods on the same population. The planted individuals can be thought of as acting as a control group and provide similar benefits. The marking and recapture process provides information on the capturability distribution of the population. And lastly, the use of different capture processes, would provide information from perspective of different selection profiles.

Many scientific studies find themselves essentially working with non-random samples. Consider the problem in astronomy where a sample survey of an area of the sky is undertaken with a low powered telescope. The observed stars are definitely a non-random sample from the stars of the milky way galaxy either being closer and brighter stars. Many drug

studies work with a control group and test group chosen randomly from a non-random sample from the population. Databases of customers, subscribers, and many other sets of data are often non-random samples from the larger population. Capture-recapture methodologies thus have applications beyond the simple population size estimation problem.

REFERENCES

Burnham K.P., Overton W.S. (1979) <<Robust Estimation of Population Size when Capture Probabilities vary among animals>>, *Ecology* 60(5),1979, p927-936.

Cormack RM (1968) <<The statistics of capture-recapture methods>> *Annual Review of Oceanography and Marine biology* 6, 455-506.

Kendall M, (1945) *The Advanced Theory of Statistics Volume 1*. Charles Griffin & Company, London.

Schroeder M. (1994) *Fractals, Chaos, Power Laws*, W.H. Freeman and Company, New York.

Weil S. V., Votta L.G. (1993) <<Assessing Software Designs Using Capture-Recapture Methods>>, *IEEE Transaction on Software Engineering* 19(11), 1045-1054.