

Minimal Sufficiency in Rare Populations

Mohammad Moradi¹, Jennifer A. Brown² and Miriam Hodge²

1. Department of Science, Razi University, Kermanshah 67146, Iran

2. Department of Mathematics and Statistics, University of Canterbury, Christchurch 8011, New Zealand

Received: December 8, 2011 / Accepted: April 13, 2012 / Published: May 25, 2012.

Abstract: It is well understood that for conventional survey designs the set of unordered distinct units in a sample is a minimally sufficient statistic. This means that for inferential statistic of the sample, the value of the sampled units rather than the sample design is important. Sampling rare populations presents distinct challenges. Examples of rare populations are in biology with rare and endangered animals where there are only a few remaining individuals, or in social science, with the low incidence of people from an unusually high (or low) income group. Sampling rare populations tends to result in the case that many of the sample units do not contain information on the characteristic of interest (e.g., the rare animal, or people from the unusual income group). For finite rare populations the set of unordered distinct rare-units in a sample is a minimally sufficient statistic. In an example case study of a rare buttercup, the properties of the minimal sufficient estimator are explored. We compare the efficiency of the estimator for the population total based on the minimally sufficient statistic, with the standard estimator for a range of sample sizes. The variance of the minimally sufficient estimator was always smaller than the variance of the sufficient estimator. For rare populations where non-rare units can be distinguished from rare units because they have the same fixed value, the minimal sufficient statistic is the rare units, if any, in the sample.

Key words: Rare population, finite population, sufficiency, minimal sufficiency.

1. Introduction

Sampling rare populations is a very topical issue in environmental science, and in social and health sciences. In environmental science an example of a rare population is either new incursions of unwanted species [1, 2], or a declining population near local-extinction. In either category, efficient sampling methods are needed to provide managers with timely information on population change [3]. In social science, sampling special populations such as minority ethnic groups, unusual income groups, people with a low-incidence disease, and unusual household structure can be viewed as sampling for rare events [4].

The definition of a rare population varies, but generally it is defined in terms of abundance and range of a population [5-7]. We use the concept of rarity

defined in terms of sampling effort. Sampling for rare populations can be time consuming because unless some individuals are found, given the rare population exists, there is no information on the population of interest.

Sampling efficiency in species abundance surveys of rare populations can be improved by estimating subareas where the likelihood of individuals occurring is particularly high. Auxiliary information on, for example, habitat suitability can be used to guide search effort to areas most likely to contain individuals. Stratified sampling and two-stage sampling [8] are examples of auxiliary information used to partition study areas into subareas. Survey intensity among subareas can be varied to ensure search effort is focused within those subareas with the highest abundance.

Adaptive sampling and unequal probability designs [9] are examples of designs that can be used to focus

Corresponding author: Mohammad Moradi, Ph.D., academic, research field: sampling. E-mail: moradim@razi.ac.ir.

sample effort to where individuals from the rare population are most likely to be found. Adaptive cluster sampling [10] was designed for surveys of rare and clustered populations, and focuses survey effort to the immediate neighbourhood of any observed individual. Designs such as two-phase stratified sampling [11] and two-phase sequential sampling [12] are some other designs that have been developed for allocating higher survey intensity to subareas with higher abundances.

In all these situations sampling effort is targeted to ensure some of the observed sample units have non-zero values. Here we discuss minimal sufficiency statistics for rare populations where the non-rare units have zero value. We show that the minimal sufficient statistic does not require information from the sample units that have zero values, and only requires information from the rare population sample units.

2. Minimal Sufficiency in Rare Populations

2.1 Notation

We follow notation from Ref. [9]. θ is defined as $\theta = (y_1, y_2, \dots, y_N)'$ and $\theta \in \Theta$, a set of \mathbb{R}^N . Examples of θ are the mean, μ , and the population total, τ . It can be considered that the parameter consists of two components, $\theta_1 = (y_1, y_2, \dots, y_{N_1})'$ and $\theta_2 = (y_1, y_2, \dots, y_{N_2})'$ where $N_1 + N_2 = N$ and represent the rare and non-rare units in the population respectively.

The ordered sample, s_o , is the sample locations listed in the order in which they are selected, $s_o = (i_1, i_2, \dots, i_n)$. A second representation of the sample is the reduced set. As with any set there is neither order, nor duplicate elements. The reduced set will have v elements $s_R = \{i_1, i_2, \dots, i_v\}$, $v = n$ only if there are no repeated sample units as in sampling without replacement.

We define two descriptions of the data that correspond to the two descriptions of the sample: the ordered data and the reduced data set. The ordered data, d_o is a series of ordered pairs of locations and values

listed by their selection order $d_o = ((i_1, d_1), (i_2, d_2), \dots, (i_n, d_n))$. The reduced data set, d_R , $d_R = \{(i, y_i) : i \in s_R\}$ is the data with all ordering and duplication information removed. We will define two additional descriptions of the data, the distinct rare units, d_{R1} , and distinct non-rare units, d_{R2} . Both these use the definition of a rare unit. For our purposes, a rare unit is any unit whose value is above a threshold zero. The distinct rare units, d_{R1} , and the distinct non-rare units, d_{R2} divide the data into two mutual exclusive sets based on whether the y value meets the criteria for rarity, $d_{R1} = \{(i, y_i) : i \in s_R, y_i \neq 0\}$ and $d_{R2} = \{(i, y_i) : i \in s_R, y_i = 0\}$.

The distinct rare units, d_{R1} , are the focus of the theorems to follow, and they do not contain information about order, duplication, or any non-rare units.

Now that we have defined several descriptions of our data, we will define functions that move between them. The function r reduces d_o to d_R , such that $d_R = r(d_o)$. The functions $r1$ and $r2$ further reduce d_R to d_{R1} and d_{R2} respectively, such that $d_{R1} = r1(d_R)$ and $d_{R2} = r2(d_R)$.

Finally, we define some concepts that relate the data to the parameter. The sampling design, $p_\theta(s_o)$, is a function which describes the probability of selecting the ordered sample, s_o . Consistency of a parameter vector, θ , and a reduced data set, d_R occurs when each element in d_R has an equivalent element in θ . For a given consistency relationship, we define an associated subset of the parameter space, Θ_{dR} , that contains the vectors that are consistent with d_R . The selection probability, $p_\theta(d_o)$ is the probability of collecting a given data set,

$$p_\theta(d_o) = p(s_o | y_s) I_{\Theta_{dR}}(\theta) \quad (1)$$

where $I_{\Theta_{dR}}(\theta)$ is an indicator function that is one when θ is consistent with Θ_{dR} , and y_s is the set of y values in the sample.

2.2 Proof of Minimal Sufficiency

It is well established that d_R is a minimal sufficient

statistic for θ [13, 14]. We prove, for a rare population where non-rare units have zero value, that while d_R is not minimally sufficient, the reduced rare data, when expressed as an unordered set d_{R1} is a minimal sufficient statistic for θ .

Theorem 2.1 With any conventional or adaptive design satisfying Eq. (1) in a rare population where non-rare unit values are fixed, e.g., 0, and the rare unit values are larger than the fixed value, d_R is not a minimal sufficient statistic for θ .

Proof: First, we show d_{R1} is a sufficient estimator. The likelihood Eq. (1) in the rare population can be written as:

$$\begin{aligned} p_\theta(d_o) &= p(s_o|y_s)I_{\Theta_{dR}}(\theta) \\ &= p(s_o|y_s)I_{\Theta_{1dR1}}(\theta_1)I_{\Theta_{2dR2}}(\theta_2) \end{aligned}$$

We know $\theta_2=(0,0,\dots,0)$, then θ_2 is consistent with any d_{R2} and $I_{\Theta_{2dR2}}(\theta_2)=1$. We have now $I_{\Theta_{dR}}(\theta) = I_{\Theta_{1dR1}}(\theta_1) = I_{\Theta_{dR1}}(\theta)$, then

$$p_\theta(d_o) = p(s_o|y_s)I_{\Theta_{dR1}}(\theta) \quad (2)$$

which represents a factorization into two functions, one involving only the data and the other involving statistics and the parameter. Sufficiency then follows directly by the factorization theorem [15].

We prove d_{R1} is sufficient and we know r_1 is not one-to-one function of d_R therefore d_R cannot be minimal sufficient estimator in the rare population.

Theorem 2.2 With any conventional or adaptive design satisfying Eq. (1) in a rare population, d_{R1} is a minimal sufficient statistic for θ .

Proof: Let $d_o = (s_o, y_s)$ and $d_o^* = (s_o^*, y_{s^*})$ be any two data values with $p_o(s_o, y_s) > 0$ and $p_o(s_o^*|y_{s^*}) > 0$, and let r_1 be a reduction function of distinct rare units.

Suppose first that $r_1(d_o) = r_1(d_o^*)$, that is, $d_{R1} = d_{R1}^*$, $y_s = y_{s^*}$ and $\Theta_{dR1} = \Theta_{d^*R1} = \Theta_{dR}$. Then, by Eq. (2),

$$\begin{aligned} p_\theta(d_o) &= p(s_o|y_s)I_{\Theta_{dR1}}(\theta), p_\theta(d_o^*) \\ &= p(s_o^*|y_{s^*})I_{\Theta_{d^*R1}}(\theta) \end{aligned}$$

for all $\theta \in \Theta$ so that

$$f(d_o; \theta) = k(d_o, d_o^*)f(d_o^*; \theta)$$

with $k(d_o, d_o^*) = p(s_o|y_s)/p(s_o^*|y_{s^*})$.

Conversely, suppose that $f(d_o; \theta) = k(d_o, d_o^*)f(d_o^*; \theta)$ for some function $k(d_o, d_o^*)$ not depending on θ . Then substituting for f ,

$$p(d_o|y_s)I_{\Theta_{dR1}}(\theta) = k(d_o, d_o^*)p(d_o^*|y_{s^*})I_{\Theta_{d^*R1}}(\theta)$$

for all θ . Since $p(d_o|y_s)$ and $p(d_o^*|y_{s^*})$ are both positive, and the indicator function I takes on only the values 0 and 1, equality requires that the two indicator functions are 0 or 1 at the same time. Thus, $I_{\Theta_{dR1}} = I_{\Theta_{d^*R1}}(\theta)$ for all θ so that $\Theta_{dR1} = \Theta_{d^*R1}$ and hence $r_1(d_o) = r_1(d_o^*)$.

We have proved that $r_1(d_o) = r_1(d_o^*)$ if and only if $f(d_o; \theta) = k(d_o, d_o^*)f(d_o^*; \theta)$ for all θ and so that the reduction d_{R1} is a minimal sufficient statistic for θ .

3. Case Study: Buttercups

We use the example from Ref. [16] of a population of rare buttercups found in the South Island of New Zealand. The buttercups in the study are within the Lance McCaskill Nature Reserve. The Castle Hill buttercup (*Ranunculus crithmifolius* sub. *paucifolius*) is one of New Zealand's rarest plants. Locations of buttercup plants observed in a study conducted in November 1998 were mapped within the study area using 10×10 m quadrants (Fig. 1). Each of the 300 quadrants is considered a sample unit, with index labels for the units ranging from 1 to 300. The top row is labelled 1 to 15, the next row 16 to 30 and so on. The bottom row is labelled 286 to 300.

We illustrate the various descriptions of the data with a trivial sample of $n = 5$ selected with replacement. The statistic of interest for this population in the total population size, τ . The shaded quadrants in Fig. 1 are the units selected in the sample. There are three expressions of the sample that are of use to us. The sample in collection order, s_o ; the sample in ascending order by index, s ; and the unordered, reduced set of sample sites, s_R . There values are below.

The ordered sample is

$$\begin{aligned} s_o &= (i_1, i_2, i_3, i_4, i_5) \\ s &= (43, 253, 156, 290, 43) \end{aligned}$$

4. Conclusions

The minimal sufficiency of the reduced data for conventional and adaptive designs was noted in Ref. [13]. Several proofs have been offered for this [14, 17]. Here we have extended the understanding of minimal sufficiency to rare populations. Minimal sufficient statistics are usually derived from all the units in the sample. We have proven that for rare populations where non-rare units can be distinguished from rare units because they have the same fixed value, the minimal sufficient statistic is the rare units, if any, in the sample.

In our case study we use a simple random sampling and a non-rare unit value of 0, but our findings extend to unequal probability sampling. Unequal probability sampling is becoming increasingly popular for surveys of rare populations because survey effort can be targeted to the locations of the rare units. In these situations the minimal sufficient estimator will use only information from the rare units.

References

- [1] F.W. Allendorf, L.L. Lundquist, Introduction: Population biology, evolution, and control of invasive species, *Conservation Biology* 17 (2003) 24-30.
- [2] S.V. Mehta, R.G. Haight, F.R. Homan, S. Polasky, R.C. Venette, Optimal detection and control strategies for invasive species management, *Ecological Economics* 61 (2007) 237-245.
- [3] W.L. Thompson, Introduction, in: W.L. Thompson (Ed.), *Sampling Rare or Elusive Species*, Island Press, Washington, 2004.
- [4] S. Sudman, M.G. Sirken, D.D. Cowan, Sampling rare and elusive populations, *Science* 240 (1998) 991-996.
- [5] R.C. Venette, R.D. Moon, W.D. Hutchison, strategies and statistics of sampling for rare individuals, *Annual Review of Entomology* 47 (2002) 143-174.
- [6] K.J. Gaston, *Rarity*, Chapman & Hall, New York, 1994.
- [7] K.J. Gaston, What is rarity? In: W.E. Kunin, K.J. Gaston (Eds.), *the Biology of Rarity: Causes and Consequences of Rare-Common Differences*, Chapman & Hall, New York, 1997.
- [8] W.G. Cochran, *Sampling Techniques*, 3rd ed., Wiley, New York, 1997.
- [9] S.K. Thompson, G.A.F. Seber, *Adaptive Sampling*, Wiley, New York, 1996.
- [10] S.K. Thompson, Adaptive cluster sampling, *Journal of the American Statistical Association* 85 (1990) 1050-1059.
- [11] R.I.C.C. Francis, An adaptive strategy for stratified random trawl surveys, *New Zealand Journal of Marine and Freshwater Research* 18 (1984) 59-71.
- [12] J.A. Brown, M.M Salehi, M Moradi, G. Bell, D.R. Smith, An adaptive two-stage sequential design for sampling rare and clustered populations, *Population Ecology* 50 (2008) 239-245.
- [13] D. Basu, Role of the sufficiency and likelihood principles in sample survey theory, *Sankhya* 31 (1969) 441-454.
- [14] A. Chaudhuri, H. Stenger, *Survey Sampling: Theory and Methods*, Marcel Dekker, New York, 1992.
- [15] E.L. Lehman, *Theory of Point Estimation*, Wiley, New York, 1983.
- [16] J.A. Brown, Adaptive sampling of ecological populations, in: Y. Rong (Ed.), *Environmental Statistics and Data Analysis*, ILM Publications, Hertfordshire, 2011.
- [17] C.M. Cassel, C.E. Särndal, J.H. Wretman, *Foundations of Inference in Survey Sampling*, Wiley, New York, 1977.