Preventive Veterinary Medicine 151 (2018) 40–45

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

Validity of a two-stage cluster sampling design to estimate the total number of owned dogs

Oswaldo Santos Baquero*, Marcos Amaku, Ricardo Augusto Dias, José Henrique Hildebrand Grisi Filho, José Soares Ferreira Neto, Fernando Ferreira

Department of Preventive Veterinary Medicine and Animal Health, School of Veterinary Medicine and Animal Science, University of São Paulo, Av. Prof. Orlando Marques de Paiva, 87, Cidade Universitária, São Paulo, SP, CEP: 05508-270, Brazil

ARTICLE INFO

Keywords:
Dog population management
Sample
Estimate
Validity

ABSTRACT

Estimates of owned dog population size are necessary to calculate measures of disease frequency and to plan and evaluate population management programs. We calculated the error and bias of estimates of the total number of owned dogs using a two-stage cluster sampling design. The estimates were conditioned on sample composition as well as on size and heterogeneity of the spatial distribution of owned dog populations. For this, we simulated nine cities that differed systematically in size (number of census tracts) and heterogeneity (variance of the number of dogs per census tract). Then, we defined 16 scenarios to calculate the sample composition using an algorithm that incorporated data from a pilot sample, estimates of cost, and prior specifications of the expected error and confidence level. In three additional scenarios of predefined sample composition, the numbers of primary and secondary sampling units were: 30 × 30, 50 × 20 and 65 × 15. Finally, for each city and sample composition, we selected primary sampling units (census tracts) with probability proportional to its size and with replacement, and secondary sampling units (households) by simple random sampling. For each city and composition we selected 500 samples, totaling 85500 samples. The distribution of errors conditioned on the sample composition and city showed that estimates were accurate (average mean bias = 0.006%, maximum mean bias = 0.3%). All sample compositions resulted in errors between 4% and 7% in cities with low heterogeneity. In cities with high heterogeneity, the errors for the various compositions ranged as follows: 8–11% (calculated), 11–13% (65 × 15), 12–14% (50 × 20) and 15–17% (30 × 30). The sample size of predefined compositions was between 33% and 87% lower than the sample size of calculated compositions. Therefore, the predefined compositions have an operational advantage (reduced sampling effort) and simplify the sampling design (calculation of sample composition is not needed). Furthermore, the expected error of estimates under different scenarios is known for each predefined composition. In the absence of information about the heterogeneity of the cities, the 65 × 15 is the more conservative composition.

1. Introduction

Owned dogs can comprise 95% of the total dog population (Matos and Alves, 2002), are an important source of recruitment of street dogs (Baquero et al., 2016) and have higher contact rates with humans. The management of owned dogs and the diseases they suffer depend on knowledge of the population size to follow the population dynamics or to calculate prevalences and incidences. Downes et al. (2013) conducted a systematic review of methods to estimate owned dog and cat populations and they discovered selection and information sources of bias but did not study statistical sources of bias. They also found marked differences in estimates from the same population calculated using different methods. Furthermore, only one of the studies that they considered reported confidence intervals. This lack of error measures precludes sound inferences and comparisons of population parameters over time.

Two-stage cluster sampling designs are widely used to estimate household parameters (household-level variables) because they do not rely on sampling frames with all households of the target population and have economic advantages due to the lowest dispersion of the selected sampling units (Levy and Lemeshow, 2008). The first stage comprises the selection of primary sampling units, which are sets of primary sampling units; whereas the second stage is used to select secondary sampling units, which are the elements with the parameters to be estimated (Levy and Lemeshow, 2008). It is known that for a fixed sample composition, the greater the variance of the target parameter in

* Corresponding author.
E-mail address: baquero@usp.br (O.S. Baquero).

https://doi.org/10.1016/j.prevetmed.2017.12.017
Received 28 August 2017; Received in revised form 28 November 2017; Accepted 24 December 2017

the primary sampling units, the greater the error of the estimates (Levy and Lemeshow, 2008).

The number of dogs per household is a household parameter and thus, the owned dog population size can be estimated using two-stage cluster sampling designs. However, the error of estimates depends on the sample composition and on the variance of the number of dogs per households in primary sampling units. Therefore, we made a simulation study based on real data, to calculate the error of estimates of owned dog population sizes for different sample compositions, under different scenarios of population size and variance of the number of dogs per census tract (heterogeneity of spatial distribution). We also calculated the bias of estimates to make a complete characterization of the validity of estimates.

2. Methods

We simulated cities that differed systematically in size (number of census tracts) and heterogeneity of the spatial distribution (variance of the number of dogs per census tract). In each city, we took samples that differed in composition; some compositions were calculated and others predefined. Finally, we calculated the errors and biases. The following subsections describe the methods to simulate cities, calculate sample compositions, select sampling units, and calculate errors and biases.

2.1. Simulation of cities

We simulated nine cities varying in size (number of census tracts) and heterogeneity (variance of the number of dogs per census tract) to represent scenarios ranging from favorable (small and low heterogeneity) to challenging (large and high heterogeneity) for obtaining precise estimates. This resulted in three groups of 100, 300 and 500 census tracts respectively. Within each group there was a city of low heterogeneity (homogeneous), one of high heterogeneity and one at the midpoint.

For each census tract, the number of households that composed it was drawn from an empirical cumulative distribution function fitted to the observed number of households per census tract, according to census data from the city of Votorantim, São Paulo (IBGE, 2010) (Fig. 1). The number of dogs per households was simulated as follows:

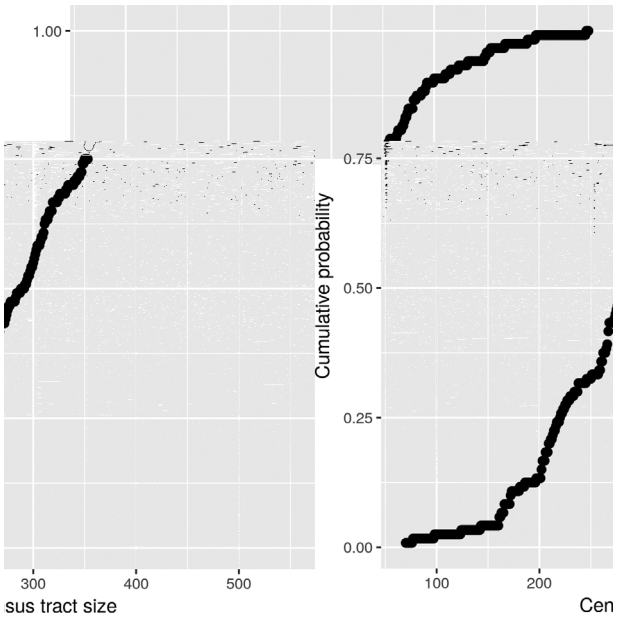

Fig. 1. Empirical cumulative distribution function fitted to the observed number of households per census tract in Votorantim, 2010.

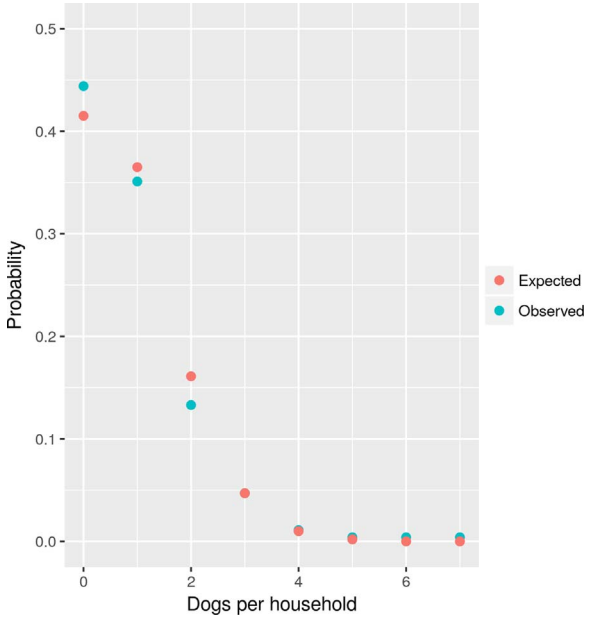

Fig. 2. Distribution of dogs per household in Votorantim, 2010.

1. Calculation of the mean number of dogs per household (maximum likelihood estimate of the Poisson parameter λ) observed in a survey conducted in Votorantim, São Paulo (Baquero et al., 2015). The calculated λ was 0.88.
2. In homogeneous cities (low heterogeneity) we allocated dogs to households using the fitted λ (Fig. 2). The mean variance of these cities was 6370.
3. In cities of mid heterogeneity, we allocated dogs to households using an equidistant sequence λ_i, $i = \{0.5, \ldots, 1.5\}$ with as many values as households in the city. The number of dogs in household i was given by λ_i. The mean variance of these cities was 15501 (approximately 2.5 times the variances of cities with low heterogeneity).
4. In cities of high heterogeneity, we allocated dogs to households using an equidistant sequence λ_i, $i = \{0.3, \ldots, 2\}$ with as many values as households in the city. The number of dogs in household i was given by λ_i. The mean variance of these cities was 32069 (approximately 2.5 times the variances of cities with mid heterogeneity).

2.2. Algorithm to estimate the sample composition

The aim of the estimation of the sample composition was to find the number of census tracts and the number of households per census tracts that should be sampled to obtain estimates with an error specified in advance in an algorithm. Our algorithm was based on the procedures described by Levy and Lemeshow (2008), which consider data collected in a pilot study, a cost function and a prior error specification (Fig. 3). We combined different pilot compositions, costs and errors (Table 1) to test how estimates of sample composition change when the values of these variables are doubled.

The pilot sample was used to calculate the total number of dogs X_{ij}, in each household j, of the census tract i. The variances of the dogs per census tract (interclass) σ_{1x}^2 and between households (intraclass) σ_{2x}^2 were given by:

$$\sigma_{1x}^2 = \frac{\sum_{i=1}^{M} (X_i - \overline{X})^2}{M}$$

$$\sigma_{2x}^2 = \left(\frac{1}{N}\right) \sum_{i=1}^{M} \left(\frac{N_i}{(N_i - 1)}\right) \sum_{j=1}^{N_i} (X_{ij} - \overline{X_i})^2$$