Spatial and Spatio-temporal Epidemiology 26 (2018) 95–105

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

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

Cost-efficient case-control cluster sampling designs for population-based epidemiological studies

Thomas Ly[*], Myles Cockburn, Bryan Langholz

University of Southern California, Keck School of Medicine, Department of Preventive Medicine, 2001 North Soto Street, Los Angeles, CA 90032, USA

ARTICLE INFO

Article history:
Received 22 June 2017
Revised 10 May 2018
Accepted 18 May 2018
Available online 15 June 2018

Keywords:
Individually-matched case-control Design
Survey sampling
Cluster sampling
Two-stage sampling
Sampling weights

ABSTRACT

Cost-efficient sampling schemes for population-based case-control studies are necessary for sampling subjects in geographically dispersed populations where in-house surveys are expensive to conduct due to high interviewer travel costs that may be associated with simple random sampling. Motivated by the original study conducted by Cockburn et al. (2011) that investigated the relationship between exposure to pesticides and prostate carcinogenesis, a set of cluster-based individually matched case-control designs is presented for cost-efficient sampling of additional controls. Based on cluster sampling from the field of survey sampling, the case-control study designs presented, where one case is individually matched to three controls, use case-control status in the sampling of the primary sampling clusters. In the secondary stage, interviewer travel costs are reduced by subsampling additional controls within primary sampling clusters as opposed to selecting additional controls purely at random, which would be highly inefficient from a cost perspective. Compared to the simple random sampling (SRS) 1:1 and SRS 1:3 (one case matched to: n SRS control(s)) designs, computer simulations demonstrate that these cluster-based designs provide unbiased rate ratio estimation and statistical efficiencies that are no worse than the SRS 1:1 design and moderately less than the SRS 1:3 design. Even under situations where the intracluster correlation for the exposure variable is extremely high for the exposure of interest, the cluster-based designs have statistical efficiencies that are comparable to that of the SRS 1:1 design. Furthermore, a cost-efficiency analysis is presented that demonstrates that the cluster-based designs are more cost-efficient compared to the SRS 1:3 design.

1. Introduction

In 2011, Cockburn et al. found relationships between environmental exposure to pesticides/fungicides with plausible biologic roles in prostate carcinogenesis and prostate cancer in the California Central Valley, one of the most intensively farmed areas in the U.S. (Cockburn et al., 2011). Cases ($n = 173$) were obtained from a population-based cancer registry, and controls ($n = 162$) were identified from Medicare listings and tax assessor mailings. For control subjects, attempts were made to recruit by mail and for those that did not respond, field visits were scheduled to visit the residence in the tax assessor database. Past ambient exposures to pesticides/fungicides were derived from residential history and independently recorded pesticide and land-use data, using a novel graphical information systems (GIS) approach. In order to obtain lifetime residential histories (and details of potential confounding prostate cancer risk factors), all cases and controls were interviewed, either by telephone or in person. The study found an increased risk of prostate can-

[*] Corresponding author.
E-mail address: thomas.thantily@gmail.com (T. Ly).

https://doi.org/10.1016/j.sste.2018.05.002
1877-5845/

cer among subjects exposed to compounds that may have prostate-specific biological effects (methyl bromide (odds ratio = 1.62, 95% confidence interval: 1.02, 2.59) and a group of organocholorines (odds ratio = 1.64, 95% confidence interval 1.02, 2.63)) but not among compounds included as controls (simazine, maneb, and paraquate dichloride).

Two novel cluster-based individually matched case-control designs are presented that allow for a cost-efficient method of sampling additional control subjects as opposed to sampling additional controls at random, particularly in areas such as the California Central Valley where the population is widely scattered about a large geographical area. Furthermore, the designs apply to studies where the travel budget is limited for field interviews and where additional demographic data and possibly biospecimens are required from study subjects. The two designs to be discussed assume that the study population can be fully enumerated and random sampling of cases and controls can be carried out to conduct nested case-control studies.

The novel individually matched case-control designs are based on two-stage cluster sampling from the field of survey sampling. The methods presented use case-control status in the sampling of primary clustering units. Interviewer travel costs are reduced by subsampling additional controls within primary sampling units as opposed to selecting additional controls purely at random. These individually matched two-stage cluster sampling designs are shown to produce unbiased estimates of the relative risk, have acceptable levels of statistical efficiency, and are more cost-efficient than typical simple random sampling designs.

A brief introduction of survey sampling and cluster sampling is presented in Section 2. Section 2 also discusses the idea of cost-efficiency in sampling. Section 3 proposes two cluster-based case-control designs and their respective control sampling weights that can be incorporated as log-offset terms in statistical software packages. Section 4 presents a description of the statistical simulations used to evaluate the operating characteristics of the two proposed cluster-based designs compared to simple random sampling designs and full risk set sampling. Section 5 discusses the results of the simulations of the cluster-based case-control designs. Section 6 presents a cost-function analysis of one of the cluster-based designs proposed. In conclusion, a discussion on how to apply the cluster based designs to the Cockburn et al. (2011) study and several real world examples are presented.

2. Cluster sampling

Suppose we wanted to estimate the average household income in a city. To make the most out of the study budget we would want to obtain a statistically informative sample, with high precision, at minimal cost in terms of recruiting and interviewing subjects. This is highly dependent on the chosen sampling design. When designing a sampling strategy for a survey to collect information, both data collection costs and statistical information should be considered. The term “cost-efficiency” emphasizes both the financial cost involved in the data collection process and the statistical information obtained, with respect to a sampling design, and is generally defined as the statistical information per unit of cost. “Cost-efficiency” is used as a summary measure to compare various sampling designs. Given a survey sampling design, there always exist a tradeoff between the cost of collecting data on a sample and the statistical information obtained from the chosen sample.

The “economic cost” of a sampling strategy is the amount of time and effort that goes into sampling that is generally expressed in monetary terms. It is usually a function of the time and financial resources required for recruiting and interviewing study subjects that also includes the costs associated with traveling to the next sampled study subject’s location. “Statistical efficiency,” with respect to a given sample based on a specific sampling design, refers to the statistical information associated with estimating a statistical parameter, in this case average household income. When comparing several sampling designs, the design with that yields the lowest variance for estimating average household income is regarded as the most statistically informative or statistically efficient. A highly statistically efficient design is simple random sampling (SRS), which can generate a representative sample of all incomes in the city. However, SRS may not be a very cost-efficient design, especially if the city was fairly large and the population was geographically dispersed, which would result in higher travel costs during field recruitment and interviews. Cluster sampling, on the other hand, will never be as statistically efficient as SRS, but is more cost-efficient in that it balances the economic cost and statistical information in estimating the average household income across the large geographically dispersed city.

In survey sampling, cluster sampling is considered the most economical form of sampling where groups of population elements constitute a sampling unit rather than a single element in the population (Sheaffer et al., 1995). By sampling groups or clusters of elements that are proximal to one another, sampling becomes more economical compared to standard SRS in terms of the time and financial resources devoted to sampling. However, cluster sampling may not represent the true diversity of a population and provides less statistical information per sampled observation compared to standard SRS, this may be a small compromise when the savings in time and financial resources outweigh the slight loss in statistical information.

In cluster sampling, a population of elements is first divided into mutually exclusive clusters of elements or primary sampling units (PSU) and then a sample of the PSUs is taken. There may be no significant reduction in time and money when sampling between PSUs. However, the reduction in sampling costs comes from sampling additional subjects from among the secondary sampling units (SSU), that make up the PSUs. Examples of commonly used PSU and SSU pairs include city blocks and households within the city blocks, census tracts and people residing within census tracts, or even housing units and the people living within the units. It is important to keep these two levels in mind when applying cluster sampling. Fig. 1 is a simple illustration of a population where two-stage cluster sampling is applied. Fig. 1(A) shows a population that