USES OF SAMPLING METHODOLOGY IN EPIDEMIOLOGIC RESEARCH

Esa Läärä
University of Oulu, Finland, esa.laara@oulu.fi

Health surveys, conducted by public health agencies, are used to address descriptive epidemiological questions on the distribution of health-related characteristics in a finite target population. The study population in these surveys is typically obtained by complex multi-stage sampling. Etiological research questions in turn concern the causal effects of suspected risk factors on the risk of getting a given disease or other health outcome. The relevant universe can be the whole humankind or a defined domain of it, and the study population is commonly a highly selected convenience sample. Statistical analysis on the parameters of interest (like hazard ratios from a Cox model) views this population as a simple random sample of an imagined superpopulation of similar kind of people.

An epidemiological study population is either closed (cohort) or open (dynamic population). The term study base refers to the experience of the study population in time. A longitudinal base consists of individual follow-up times of the study population, these times often being left-truncated and right-censored. All cases of the outcome occurring in the study base are expected to be ascertained. For risk factor data an ideal option would be complete enumeration of the base, a.k.a. the full cohort design. This can be feasible for register or questionnaire data, but not for risk factors to be measured from e.g. blood samples for a big population needed for an adequate number of cases of a rare outcome.

A highly cost-efficient alternative, historically known as the case-control study, employs a two-phase outcome-selective sampling strategy. Here, risk factor data are collected from all cases. In addition, for consistent estimation of the distribution of those factors in the whole study base, and eventually of the interesting effect parameters, a set of "controls" from it are randomly sampled with a small sampling fraction. The main design options are (A) case-noncase sampling, (B) case-cohort sampling, and (C) density sampling, these differing as to the sampling frame of the controls. In (A) they are sampled from those yet free from the outcome at the end of the relevant risk period. In (B) a subcohort, a simple or stratified sample of the whole study cohort, is selected. In (C) the controls are sampled from the base during the follow-up. In its main variant, nested case-control study, for each new case, one or more controls are sampled from the pertinent risk-set comprising members of the study population who are outcome-free and yet under follow-up at the time of diagnosis of the case. The three designs have their pros and cons, each having its own special niches of applicability. Statistical efficiency can be improved in designs (A) and (C) by close matching, i.e. stratified sampling on a few important determinants of the outcome, and/or on occasion of measurements from biobank material.

Binary logistic regression and the Cox proportional hazards model are the paradigmatic frameworks for statistical analysis in these designs; often a conditional logistic or a stratified Cox model is called for. Weighted versions of the pertinent estimating functions have received increased popularity. Approaches like those based on post-stratification and calibrated weights are also adopted for utilizing auxiliary data available in the whole study population for more efficient estimation of the parameters of interest that describe the effects of those risk factors only obtainable from the second phase sample.

Lively and fruitful exchange of ideas takes nowadays place between statisticians developing survey sampling methodology and those working with methods for design and analysis in epidemiology.

Selected references

Borgan, Ø., Breslow, N.E., Chatterjee, N., Gail, M.H., Scott, A., Wild, C.J. (editors) (2018). *Handbook of Statistical Methods for Case-Control Studies.* Chapman and Hall/CRC.
Keogh, R., Cox, D.R. (2014). *Case-Control Studies*. Cambridge University Press.
Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Wiley.