Sampling and Estimation in Finite Networks

Melike Oguz–Alper[1] and Li–Chun Zhang[2,3,4]

[1]Statistics Norway, e–mail: Melike.Oguz.Alper@ssb.no
[2]University of Southampton, e–mail: L.Zhang@soton.ac.uk
[3]Statistics Norway
[4]University of Oslo

Abstract

The conventional design–based inference or the randomisation approach conceives population data as a list of units where a set of measurements, which are assumed to be constant, are attributed to the individual units. The randomness is solely specified by a well–defined probability sampling design that assign probabilities to the samples selected repeatedly from a finite population. Population data may have a hierarchical structure where the lower–level units are nested within the higher–level units. Sampling and estimation techniques are well established for such finite list populations, regardless of whether the units are in a hierarchical structure or not.

The variety of data available today, however, raises other possibilities of representation. Data may contain non–nested relationships with many–to–many linkage between the units, unlike the conventional envision of data structure as a tree of units with only one–to–one or one–to–many relationships. There may be multiple types of relationship between the units. Such complex structures can often be represented by a graph consisting of a set of nodes and a set of edges. A valued graph where the measurements are attributed to the graph objects is called *network*. We can think of social networks, transportation networks, labour–flow network, communication networks, computer networks, etc. The parameter of interest is not necessarily defined only on the nodes, but the relational structure itself may be of interest. There is a large literature on the model-based inference in networks. However, the modelling approach may not always be viable especially when the underlying dynamics are too complicated or transient or subject to shocks. The randomisation approach to finite networks may be more useful in that case.

The theory of sampling and inference in finite networks is relatively underdeveloped, and the techniques are rarely applied in Official Statistics, despite some notable exceptions in the past such as multiplicity sampling including indirect sampling and adaptive cluster sampling. There is a recent article by Zhang and Patone (2017) which is a synthesis and extension of the graph sampling theory by covering all the existing network sampling techniques as special cases. They develop a general Horvitz & Thompson's (1952) estimator under arbitrary T–stage snowball sampling.

In this talk, at first, we will move quickly from the conventional design-based approach to finite list populations, to providing a formal definition of sampling in finite population networks. The existing multiplicity sampling techniques will be discussed in order to provide a better picture of how they could be, in fact, envisaged as network sampling techniques. When it comes to the inference in finite networks, target parameters are not limited to the totals of measures attributed to the nodes. Thus, we consider a design–based inference for higher–order target parameters which are defined based on the measures associated with the relationships or edges. Examples of such network parameters are network density, reciprocity, number of dyads or triads, transitivity, etc. We establish generally the relative efficiency of two types of Horvitz–Thompson estimators for network sampling. Results from a limited simulation study with an application to a labour–flow network will be presented, where the industrial sectors are the nodes and the labour flows between the sectors form the edges. The data used is retrieved from the Norwegian Income and Employment data in the administrative sources.

Keywords: Graph sampling, finite networks, network parameter, multiplicity–sampling