 ISBN: 9789073592421

CPS Paper

A Computational Analysis of Snowball Sampling for the Estimation of Means

Author: Mr João Gabriel Malaguti

Coauthors: Alinne de Carvalho Veiga, Letícia de Carvalho Giannella

Submission ID: 589

Reference Number: 589

Presentation File

abstracts/ottawa-2023_a510fed25175c5497ca947cc4aff42be.pdf

Files/Uploads

CPS0019_JoãoMalaguti

Brief Description

Snowball sampling is a non-probabilistic sampling method, widely used in the social sciences for its ability in reaching hard-to-reach populations, such as drug users, victims of domestic violence and queer people.

However, due to being a non-probabilistic method, formal equations for the standard error of the mean do not exist, making analyses more complex.This study bypasses this issue by making use of computational statistics, in particular Monte Carlo simulations, which allows for approximate estimations based on different controlled scenarios in order to improve the understanding of the sampling method.The effects of connection density, i.e.

the number of connections one person has inside a population, and different indication probabilities, i.e.

the probability one person has to indicate another to join the sample, were investigated on the estimation of the mean and its standard error for snowball samples.

Abstract

Snowball sampling is a non-probabilistic sampling method, widely used in the social sciences for its ability in reaching hard-to-reach populations, such as drug users, victims of domestic violence and queer people.

To put it simply, snowball sampling starts from a small initial sample and the respondents are asked to refer other people (those who belong to the population of interest) to join the sample. The referral request is repeated to these new respondents and, like a snowball rolling down a hill, the sample slowly increases in size.

However, due to being a non-probabilistic method, formal equations for the standard error of the mean do not exist, making analyses more complex. While there is a class of estimation methods (called respondent-driven sampling estimators or RDS estimators) that claim to be able to estimate the standard errors, these methods rely on a number of very heavy assumptions that are hardly met in practice.

This study bypasses these issues by making use of computational statistics, in particular Monte Carlo simulations, which allows for approximate estimations based on different controlled scenarios in order to improve the understanding of the sampling method. For this, random graphs (also called Erdos-Rényi graphs or Gilbert graphs) were used to propose/simulate populations. Graphs are mathematical models used to understand relationships, formed by two basic pieces: the node (representing the elements or individuals) and the line (representing the relations). Along with these graphs, a variable of interest y, exponentially distributed, was generated.

The effects of connection density, i.e. the number of connections one person has inside a population, and different indication probabilities, i.e. the probability one person has to indicate another to join the sample, were investigated on the estimation of the mean and its standard error for snowball samples.

These computational experiments revealed some interesting results: the connection density affects the sample size exponentially, while the probability of indication affects the sample size linearly. Another result found was that assuming simple random sampling consistently overestimates the standard error of the mean and therefore can be taken as an upper bound for the standard error under snowball sampling, though research with more complex graphs is still needed to confirm these findings.