Survey sampling basics

Summary

In this chapter we focus on the basics of sampling in the context of survey sampling. We cover three of the most common probability sampling approaches: simple random sampling, stratified sampling, and cluster sampling. We also briefly discuss non-probability sampling methods: convenience sampling, quota sampling, and snowball sampling.

Introduction

Sampling consists of selecting a subset of units from our study population (or data generating process) for the purpose of generalizing the results obtained on this subset to the entire study population.

Even those of us not familiar with survey sampling are likely to have encountered sampling in randomized algorithms, for example, random forests, or resampling methods, such as cross-validation and bootstrap. The underlying principles are the same as in survey sampling, as is the main reason for applying sampling – sampling is typically used as a means to reduce the resources required for the enquiry. In random forests and bootstrap the resources we are trying to reduce are computational, but survey sampling is more complex, because time, money, human, and other resources have to be considered.

Sampling methods can be divided into two fundamentally different groups: probability sampling and non-probability sampling. In probability sampling every unit population has a non-zero probability of being selected and that probability is known. Non-probability sampling approaches violate at least one of those two criteria, some units can't be selected or are selected in a way such that the probability of their selection is not known. Simple random sampling is an example of probability sampling, while convenience sampling - for example, selecting units that are close at hand – is an example of non-probability sampling.

The main characteristic of these two groups of approaches to sampling is that in probability sampling we can, using the laws of probability, derive the uncertainty in our results. In other words, we can quantify how representative the sample is of our population. In non-probability sampling, however, that is not possible, at least not without making additional assumptions.

The key component of sampling is the *sampling frame*. The sampling frame is a list of all units in the study population. In other words, it defines a set of units from which a researcher can select a sample of the study population. Without a sampling frame, probability sampling is not possible. However, in cases where our study population has a hierarchical structure, we can avoid the need for a detailed sampling frame for parts of the study population that were not selected at the higher level. We will discuss this in more detail in the Cluster sampling scheme.

In practice, if resources permit, probability sampling is always preferred over non-probability sampling. However, even when in situations where we can use probability sampling, the application is rarely ideal. Two main issues we face are coverage and non-response:

- Coverage: Ideally, our sampling frame would cover the entire study population. However, in practice, the sampling frame often does not include all units of the study population (under-coverage) or it includes some units that are not in the study population (over-coverage). For example, if we survey University of Ljubljana (UL) students, we might, if we are not careful, include non-UL students that are only spending a semester here (over-coverage), but miss UL students that are studying abroad this semester (under-coverage).
- Non-response: In practice, not all units can be measured. In particular, when dealing with human participants, not all participants will respond. Non-response is essentially a missing values problem.

Both coverage and non-response introduce bias into our results. In general, dealing with these issues requires careful consideration and the gathering of additional information. We will discuss non-response bias it in more detail in a later chapter. In this chapter we will, unless noted otherwise, assume that there are no coverage or non-response issues. In other words, the only source of uncertainty will be the sampling itself.

Probability sampling

In our analyses we will focus on estimating the mean. The basic principles are the same for other quantities of interest, but each requires a different model.

The Football Manager dataset that we use throughout, we obtained from Kaggle.

Simple random sampling

Recommended readings

INTRODUCTION

Sample representativeness.

s <- all possible samples p(s) <- probability mass/density of a sample

fixed-size/random size (p(s) = 0 if |s| != n then it is fixed) equal prob/unequal probability sampling (different inclusion probabilities)

Sampling scheme: method of selecting (an algorithm)

P SAMPLING SCHEMES

SIMPLE RANDOM SAMPLING

Fixed sampling scheme. Equal probability sampling. All samples of size n are equally possible.

With replacement/without replacement. The Finite Population Correction Factor (FPC) is used when you sample without replacement from more than 5% of a finite population. It's needed because under these circumstances, the Central Limit Theorem doesn't hold and the standard error of the estimate (e.g. the mean or proportion) will be too big.

UNEQUAL PROBABILITY

There are many sampling designs with une- qual inclusion probabilities and fixed sample size (see for instance Brewer and Hanif, 1983, and Tillé 2006). All the sampling designs included in this class use auxiliary information to draw the samples

POISSON SAMPLING

Missing value is poisson sampling. Random sample size.

Stratified sampling

tratification is a very simple idea. The pop- ulation is split into H non-overlapping sub- sets called strata Uh H h „ , =... 1 . Next, a random sample is selected in each stratum independently from the other strata. The final sample is the union of the samples drawn in each stratum.

t can be shown that the precision of the stratified estimator of the population total is improved compared to the estimator computed using simple random sampling without replacement when the variable of interest y is very homogeneous in the strata.

1. Proportional allocation

2. Optimal allocation (Neyman allocation) vs simple sampling

Cluster sampling

luster sampling is used for efficiency costs (for instance, when the population units are geographically spread and it takes time and money to sample using simple random sampling). It is also applied when it is difficult to construct the sampling frame (for instance, when we do not have a list of people living in households, but it is easy to construct the sampling frame of their households).

can be single or multi level

Will have mugher error than simple sampling if the clusters are highly correlated! (inter cluster corellation)

NON-PROBABILISTIC SAMPLING SCHEMES

Nonprobability sampling is any sampling method where some elements of the population have no chance of selection (these are sometimes referred to as 'out of coverage'/'undercovered'), or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is nonrandom, nonprobability sampling does not allow the estimation of sampling errors. These conditions give rise to exclusion bias, placing limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population

Convenience

Sometimes known as grab or opportunity sampling or accidental or haphazard sampling. A type of nonprobability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, readily available and convenient. The researcher using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative enough.

Judgement sampling

The researcher chooses the sample based on who they think would be appropriate for the study. This is used primarily when there is a limited number of people that have expertise in the area being researched

Quota sampling

Snowball & respondent-driven sampling

Additional readings

- Salganik, M. J., & Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. Sociological methodology, 34(1), 193-240.

Homework

a. Write down three questions from your everyday live that you often reply to with an uncertain answer. Answer those questions uncertainly with natural language. Then answer those questions using probabilistic answers that roughly correspond to the natural language questions. At least one of the questions must have uncountably many answers (truths) and at least one of the questions must have infinitely but countably many answers.

b. Install Stan and run the bernoulli toy example from the installation instructions. You may use CmdRStan (recommended), RStan, PyStan, or any other interface.

Submit a A5 report that contains (a) and the histogram of the posterior of the parameter from (b).