Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 2: Complex survey designs

University of Michigan

Bayesian inference for sample surveys

Survey sampling

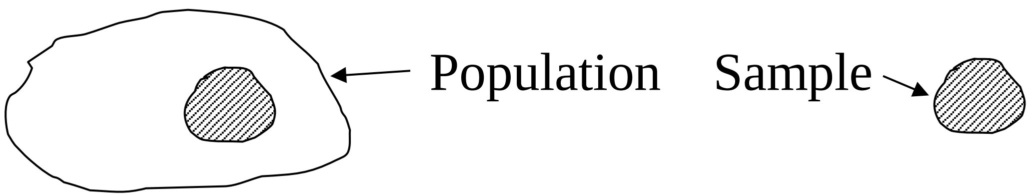
- So far we have discussed Bayes and frequentist inference for statistics in general
- We consider in this course the specific application of Bayes to survey sampling
- In this lecture we describe
  - Probability sample designs, in particular simple random sampling and more complex designs
  - Distinguishing features of survey sample inference
- In the next lecture we discuss in broad terms alternative modes of survey inference
  - Design-based, superpopulation models, Bayes

Distinguishing Statistical Features of Survey Sampling

- Major interest in *descriptive* inference about *finite population quantities*, as opposed to parameters of models (though *analytical inference* for parameters can also be of interest).
- Probability sampling – method of sampling from the population that avoids selection biases.
- Prevailing orthodoxy is *design-based* (randomization) inference: survey outcomes are treated as fixed quantities, and statistical uncertainty derives from the probability distribution that determines sample selection

Inference for a population based on a sample

- Parameters: population is thought of as drawn from an infinite "superpopulation"; Parameters are summary characteristics of this super-population, in superpopulation models (greek symbols)
- Population quantities: descriptive quantities of the population, such as means and totals (cap roman symbols)
- Statistical inference: the process of making inferences about model parameters and population quantities based on sample data.



	Parameter	Population	Sample
Mean	μ	$\bar{X}$	$\bar{x}$
SD	σ	S	s

- Inference crucially requires that sample is randomly selected from population (or an assumption that it is)

Properties of a good sampling scheme

- "representative" of the population (... whatever that means)
- demonstrably free of selection bias
- repeatable (at least in principle)
- efficient: low cost for given level of precision
- measurable precision: e.g., can quantify how close the sample estimate is to the population quantity it is estimating.
- Only probability (or random) sampling designs have these properties. Probability samples are characterized by the following two properties:
- every sample has a known (maybe zero) probability of selection
- every unit in the population has a (known) positive probability of selection.

Simple Random Sampling

- The most familiar form of probability sampling
- Simple random sampling without replacement corresponds to selecting n balls out of a well-mixed urn containing N balls (like some lotteries). For this method:
  - All possible samples of size n have an equal probability of being selected.
  - All samples of size not equal to n have zero probability of selection
  - every unit has probability n/N of selection
- SRS with replacement – units are replaced after selection, can be selected more than once
  - Impractical, but simplifies design-based theory

SRS example

- Example. Suppose the urn contains $N = 5$ balls, labeled {A B C D E}; this is our population. We select a simple random sample of $n = 2$ balls. There are 10 possible samples of size 2, namely:
- AB, AC, AD, AE, BC, BD, BE, CD, CE, DE
- Since all these samples have the same chance of being selected,
  - Pr(any size 2 sample selected) = 0.1
  - Pr(any other sample selected) = 0
  - Pr(any particular ball is included) = 0.4

Formalizing Sampling Distributions

Population units $i = 1, \ldots, N$

Sample indicator $S_i = \begin{cases} 1, & \text{unit } i \text{ selected} \\ 0, & \text{unit } i \text{ not selected} \end{cases}$

Probability sampling puts known distribution on $S = (S_1, \ldots, S_N)$

This distribution can depend on design variables Z

But not on survey outcomes Y

Simple random sampling of size n without replacement:

$$\Pr(S = s \mid Y) = 1 / \binom{N}{n}, \ \sum_{i=1}^{N} S_i = n; \ \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

$$\Pr(S = s \mid Y) = 0, \ \sum_{i=1}^{N} S_i \neq n$$

Non-random sampling methods

- Some examples of sampling methods that do not yield random samples are:
  - Sample readily accessible individuals
  - Purposive or judgmental sampling
  - Self-selected samples - volunteers, phone polls
  - Quota sampling
- These methods are less scientific and less trustworthy than probability sampling, since they are subject to hidden biases.

Neyman's (1934) paper: compared Probability Sampling versus "Purposive Sampling"

- Definition of probability sampling:
  - every sample has a *known* probability of being selected
  - every individual in the population has a positive probability of being selected
- Initially, probability sampling was equated with its basic form, simple random sampling (SRS)
  - Every sample of size *n* has *equal* chance of being selected, hence an <u>e</u>qual <u>p</u>robability of <u>se</u>lection <u>m</u>ethod (*epsem*)
  - Samples of size other than *n* have no chance of being selected
  - With and without replacement

"Purposive Sampling"

- "Non-probability sampling" – but hard to define a negative.
- Units are picked so that sample matches distribution of a characteristic known for the population.
- E.g. if we know distribution of age and gender in population, choose sample cases to match this distribution.
- A common form is *quota sampling*: interviewers are given a quota for each age group and gender and interview individuals until this quota is met

The Controversy

- Let Z = characteristic known for all units in the population (age, gender, …)
- Under simple random sampling, distribution of Z in the sample can deviate considerably from its (known) distribution in the population, purely by chance
- This “lack of representativeness” with respect to Z led some to prefer purposively picking the sample to match the population distribution of Z

Neyman's "Resolution"

- Neyman (1934) showed that we can get the best of both worlds by stratified sampling:
  - Create strata by the classifying population according to the known characteristics
  - Select a simple random sample of known size n_j from population of size N_j in stratum j
- If $f_j = n_j / N_j$ = const., results in epsem sample, retains probabilistic selection, and sample matches distribution of strata in population
- Also one can vary f_j and weight sample cases by $1/f_j$: Neyman's optimal allocation

More Complex Designs

- Neyman's paper helped to set the stage for extensions to cluster sampling, multistage sampling, greatly extending the practical feasibility and utility of probability sampling in practice
- E.g. simple random sampling of people in the US is not feasible – we do not have a complete list of everyone in the population from which to sample
- Work of Mahalanobis, Hansen, Cochran, Kish, ….

Beyond simple random sampling

- Stratified Random Sampling
  - divide population into strata (e.g. based on race)
  - select units by srs within each stratum. Different sampling fractions are allowed within strata; for example, we may over-sample minorities
  - Generally, stratifying on a variable that is related to a survey outcome increases the precision for estimating distribution of that outcome
  - More strata the better, but a sample size of at least two in each stratum is needed to provide an estimate the sampling variance

Systematic sampling from an ordered list

- For a continuous stratifying variable Z, order the population by values of Z.
- Choose a sampling interval I – the inverse of the sampling rate n/N
- Choose a random start between 0 and I, say x
- Sample units x, x+I, x+2I,…, x+(n-1)I
- Creates n implicit strata of size I, sample one unit from each stratum
- Simple and convenient, but sampling variance requires modeling assumptions

PPS sampling

- In certain applications, it is efficient to sample "large" units (firms, tax returns, transactions in an audit)…) with higher probability than "small" units – in particular when variability of outcome increases with size (as with variables like total sales, number of employees, …)
- For a continuous stratifying size variable Z, this is conveniently achieved by <u>probability proportional to size</u> (pps) sampling
- Units in the population are first ordered, either randomly or by values of Z. Then:

PPS sampling

- Associate unit i with interval (c_{i-1}, c_i), where $c_0 = 0$, $c_i = z_1 + \ldots z_i$ are <u>cumulated sizes up to i</u>, i = 1,…,n.
- Choose a sampling interval $I = z_n/n$.
- Choose a random start between 0 and I, say x
- Units corresponding to the intervals that contain the values x, x+I, x+2I,…, x+(n-1)I are sampled
- Notes:
  - Units with size greater than I are selected with probability 1. They are pre-selected and removed from the list prior to sampling from the list
  - With units randomly ordered, creates a pps sample with no implicit stratification
  - With units sorted by size, creates a pps sample with implicit stratification on size, and n implicit strata of size 1. More efficient, but sampling variance requires models

Cluster Sampling

- Group units into clusters (e.g. localities)
- Select a srs c of C clusters
- Sample all units within sampled clusters
- Useful for demographic surveys, since listing operations and interviews can focus on sampled clusters, saving on listing expense and travel time between households
- Less useful for telephone sampling, since there is no travel involved.

Two-stage sampling

- Group units into clusters (e.g. localities)
- Select a sample of size c of C clusters
- Take a simple random sample of units within sampled clusters
- A common design is to sample clusters with probability proportional to estimated size, and units within clusters with probability inversely proportional to estimated size
  - Yields an epsem sample
  - If estimated size is the true size, this yields a constant number of units in each cluster, convenient for fieldwork

Multistage sampling

- More than two stages are also possible
- E.g. sample households in two stages, and then take a subsample of individuals within households
- Or in a student sample, sample students within classes within schools
- This yields a more complex correlation structure
- The largest clusters in the hierarchy are called ultimate clusters, play an important role in design-based inference

Multistage sampling with stratification

- Efficiency is increased by stratified sampling one or more stages of selection
- E.g. stratify clusters by cluster characteristics, and take random samples of clusters within strata
- A popular design is to sample two clusters per stratum, since it allows for design-based variances to be computed.

Checking "Representativeness"

- One way of assessing representativeness is to compare distributions of known variables for the sample and the population
  - e.g. target population = U.S. Civilians
  - compare sample distribution of age, race, and sex with the population distribution from the nearest census.
  - should be done if possible, but often of limited value: really need to compare variables closely associated with the variables of interest

Distinguishing Statistical Features of Survey Sampling

- A simple and brilliant idea: simple random sampling
- Study of *complex sample designs*: designs that go beyond simple random sampling, including features like stratification, weighting and clustering
  - Simple random sampling, though simple, is not optimal or even practical in many settings
- Many practical real-world sampling issues: sampling frames, making use of administrative information, alternative modes of survey administration

Is Probability Sampling Optimal?

- Simple random sampling (or equal probability sampling in general) is an all-purpose strategy for selecting units to achieve representativeness "on average"
  - compare with randomized treatment allocation in clinical trials
- However, statisticians like optimal properties, and SRS is very suboptimal for some specific purposes…
- E.g. if distribution of X is known in population, and objective is slope of linear regression of Y on X, it's obviously much more efficient to sample equally at the two extreme values of X – this minimizes the variance of the LS slope (Royall 1970)
- But this is not a probability sample– intermediate values of X have zero chance of selection!
- For linear regression through origin, optimal design is <u>cut-off sampling</u>, which is still applied in some business surveys

Balanced Sampling

- BUT -- sampling the extremes of X does not allow checks of linearity, and lacks robustness.
- Royall and Herson (1974) argue that if linearity is a concern, choose fixed number of cases at intermediate values of X, rather leaving the sample to chance!
  - Their *balanced sampling* idea achieves robustness by matching moments of X in sample and population
- Even if sampling is random within categories of X, this is not probability sampling unless all values of X are included.

Distinctive features of survey inference

1. Primary focus on descriptive finite population quantities, like overall or subgroup means or totals
   – Bayes – which naturally concerns <u>predictive distributions</u> -- is particularly suited to inference about such quantities, since they require predicting the values of variables for non-sampled items

- This finite population perspective is useful even for analytic model parameters:

$$\theta = \text{model parameter (meaningful only in context of the model)}$$

$$\tilde{\theta}(Y) = \text{"estimate" of } \theta \text{ from fitting model to whole population } Y$$

(a finite population quantity, exists regardless of validity of model)

A good estimate of θ should be a good estimate of $\tilde{\theta}$

(if not, then what's being estimated?)

Distinctive features of survey inference

2. Analysis needs to account for "complex" sampling design features such as stratification, differential probabilities of selection, multistage sampling.

    - Samplers reject theoretical arguments suggesting such design features can be ignored if the model is correctly specified.
    - Models are always misspecified, and model answers are suspect even when model misspecification is not easily detected by model checks (Kish & Frankel 1974, Holt, Smith & Winter 1980, Hansen, Madow & Tepping 1983, Pfeffermann & Holmes (1985).
    - Design features like clustering and stratification can and should be explicitly incorporated in the model to avoid sensitivity of inference to model misspecification.

Distinctive features of survey inference

3. A production environment that precludes detailed modeling.
   - Careful modeling is often perceived as "too much work" in a production environment (e.g. Efron 1986).
   - Some attention to model fit is needed to do any good statistics
   - “Off-the-shelf" Bayesian models can be developed that incorporate survey sample design features, and for a given problem the computation of the posterior distribution is prescriptive, via Bayes Theorem.
   - This aspect would be aided by a Bayesian software package focused on survey applications.

Distinctive features of survey inference

4. Antipathy towards methods/models that involve strong subjective elements or assumptions.
    - Government agencies need to be viewed as objective and shielded from policy biases.
    - Addressed by using models that make relatively weak assumptions, and noninformative priors that are dominated by the likelihood.
    - The latter yields Bayesian inferences that are often similar to superpopulation modeling, with the usual differences of interpretation of probability statements.
    - Bayes provides superior inference in small samples (e.g. small area estimation)

Distinctive features of survey inference

5. Concern about repeated sampling (frequentist) properties of the inference.
    - Design-based inference bases the inference directly on these repeated sampling properties
    - Calibrated Bayes: model-based, but models should be chosen to have good frequentist properties
    - This requires incorporating design features in the model (Little 2004, 2006).