# A Simulation Study to Compare the Bias and Efficiency of Sample Mean Obtained from Probability and Nonprobability Samples

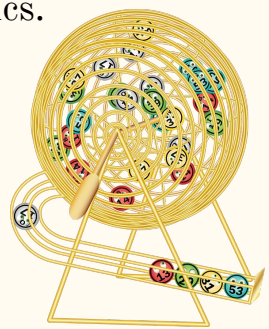
Sarah Grunblatt, Harun Mazumder, & Chaoyi Zeng

# BACKGROUND

# PROBABILITY VS. NON-PROBABILITY

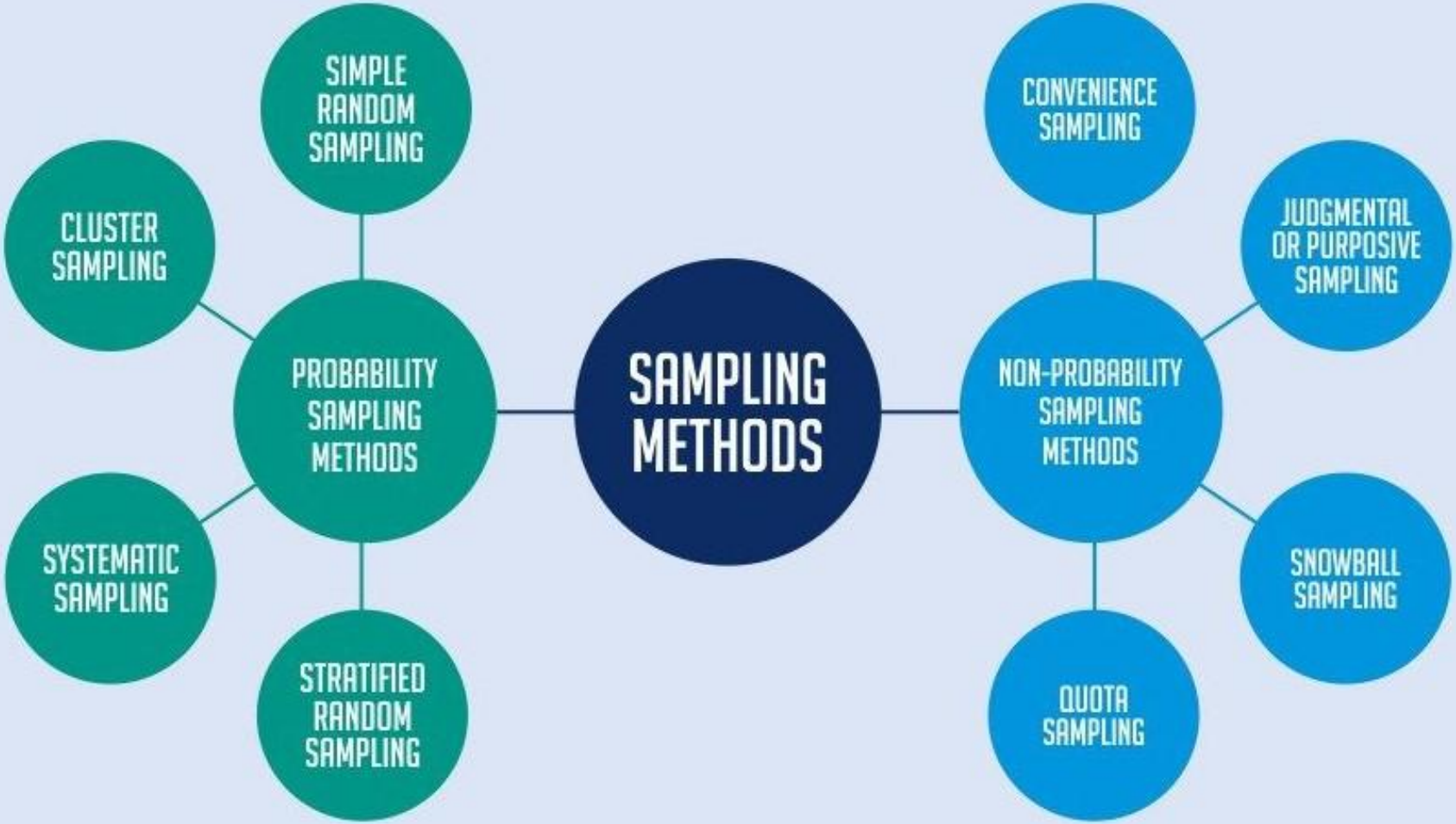
**Probability sampling involves random selection of the population sample.**

- Equal chance of random selection for all.
- It depends on the rationale of probability theory (the mathematical odds that the sample will accurately reflect the larger population).
- Confidence intervals are able to be determined for the statistics.
- More accurate and rigorous and is therefore preferred by researchers.
- Ex. A lottery drawing



**Non-probability does NOT involve random selection of the population sample.**

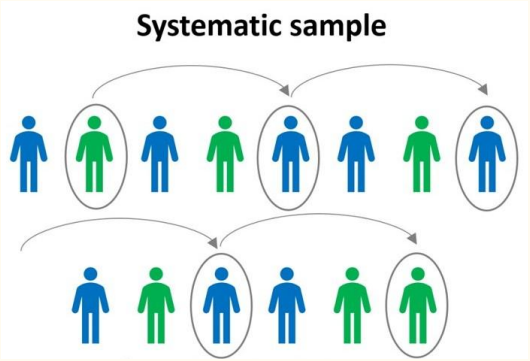
- At times, it can be very difficult to calculate the odds of an event (not equal probability).
- As such, it is difficult to know if the sample reflects the distribution of the larger population.
- In some circumstances of social research, random sampling may not be feasible, practical, or theoretically sensible.
- In these cases, nonprobability methods are implemented.
- It is cost- and time-effective, easy to use, and can work with a small population.

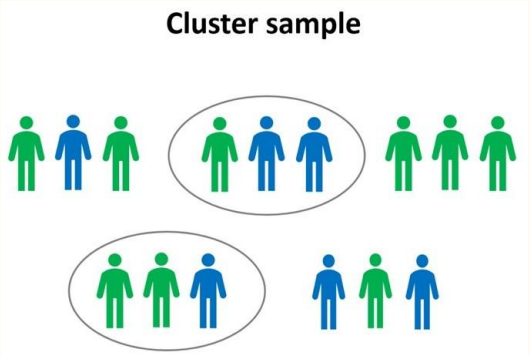



https://www.questionpro.com/blog/types-of-sampling-for-social-research/

# BASIC PROBABILITY SAMPLING TECHNIQUES

1. **Simple Random Sampling --** Randomly selecting participants until the desired sample size is met. Equal probability of selection for all.
2. **Systematic Sampling --** A starting point is randomly selected. The remainder of the sample is selected at a set interval.
3. **Stratified Sampling --** Samples are randomly selected from groups. May be unequal in size.
4. **Cluster Sampling --** The population is divided into clusters. A certain number of clusters are randomly selected. All participants in those clusters are included. May have multiple stages.

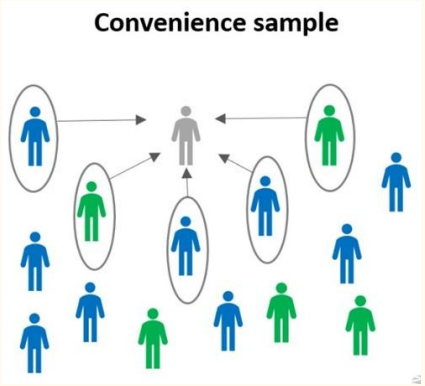



https://dataz4s.com/statistics/probability-sampling/

# BASIC NON-PROBABILITY SAMPLING TECHNIQUES

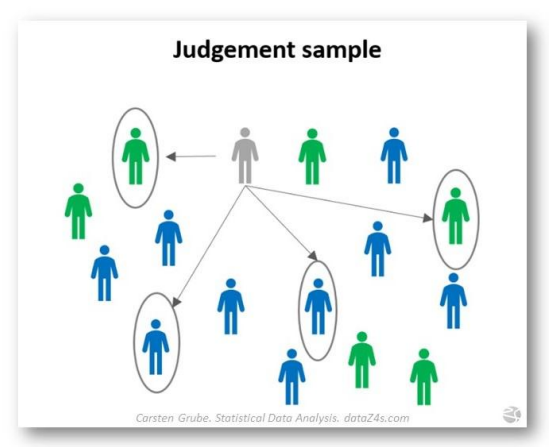
1. **Convenience, Haphazard, or Accidental Sampling**
   - Collecting a sample via the easiest possible manner.
   - Often at one location and time to whomever is available.
   - Some potential respondents may accidentally be missed or unconsciously avoided.
   - Ex. family, coworkers, grocery store, shopping mall, etc.
2. **Consecutive or Total Enumerative Sampling**
   - Similar to convenience sampling except no subjects are missed. (ALL eligible individuals are included.)
   - Good method to minimize sampling bias.
   - The best non-probability method because the sample is a better representation of the entire population.
   - Ex. consecutive houses on a street




https://dataz4s.com/statistics/non-probability-sampling/

# BASIC NON-PROBABILITY SAMPLING TECHNIQUES

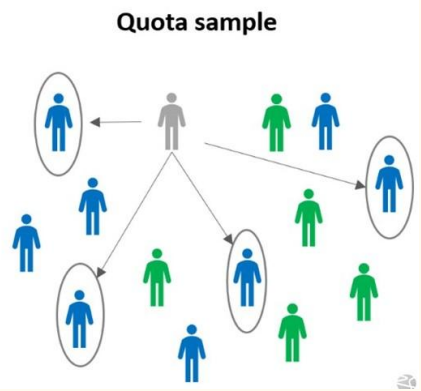
3. **Judgement or Purposive Sampling**
    - Researchers choose participants to enroll that meet certain criteria based on the study's focus.
    - Ex. Religion
    - 3 Types:
        i. **Deviant Case --** Cases that are significantly different from the main pattern
        ii. **Case Study --** Limited to one group (typically with a specific characteristic)
        iii. **Ad Hoc Quotas --** A quota is set. Participants are chosen until the quota(s) is met.

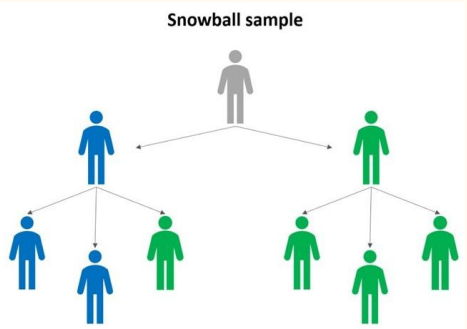



https://dataz4s.com/statistics/non-probability-sampling/

# BASIC NON-PROBABILITY SAMPLING TECHNIQUES

4. **Quota Sampling**
    - Enrolling every participant until the proportion of the groups in the sample matches the proportion of the groups in the population.
    - Ex. Tobacco use among different age groups (100 participants per age category)
5. **Snowball Sampling**
    - Participants recruit other members for the study. Often implemented with hard-to-reach populations.
    - Ex. Heroin drug users




https://dataz4s.com/statistics/non-probability-sampling/

# MOTIVATIONS & OBJECTIVE

# MOTIVATION

Many different sampling methods exist, and each has its own merits.

Typically, nonprobability sampling is cheaper and easier to implement.

In this study, we desire to compare what effect nonprobability sampling (particularly convenience sampling) will have on the sample mean, especially in regards to bias and efficacy, as opposed to results that could potentially be obtained by probability sampling methods.

**We aim to better understand if there truly a significant enough difference in the results to justify implementing one method over another.**

# OBJECTIVE

The objective of this exercise is to design a simulation study that compares the bias and efficiency of the sample means obtained from probability and nonprobability samples.

# LITERATURE REVIEW

# #1: SAYS WHO? THE SIGNIFICANCE OF SAMPLING IN MENTAL HEALTH SURVEYS DURING COVID-19

- Public Health interventions (partiularly regarding mental health) as a result of the COVID-19 pandemic will most certainly be necessary.

- Data is needed to determine how quickly, to whom, how much, and what type of resources will be needed.

- “An immediate priority is collecting high-quality data on the mental health effects of the COVID-19 pandemic across the whole population and vulnerable groups”. (The Lancet Psychiatry)

- Data is needed that represents the true need arising from the pandemic.

# #1: SAYS WHO? THE SIGNIFICANCE OF SAMPLING IN MENTAL HEALTH SURVEYS DURING COVID-19

- The desire for quick information has driven the rapid propagation of online surveys using **nonprobability** and **convenience** samples, some of which claim to be representative.

- Many are receiving widespread media attention, which frequently and greatly impacts public opinions and reactions in varying ways.

- "Acting on misleading information could be worse than having no information at all." **Convenience sampling should not exclusively be relied on** to drive policy and resources because they are prone to substantial bias.

- **CONCLUSION:** These quick methods can be valuble for initial information gathering to identify future research avenues using less biased sampling methods.

# #2: SOCIAL MEDIA, WEB, AND PANEL SURVEYS

***Using Non-Probability Samples in Social and Policy Research***

- Online surveys are growing in popularity due to their low cost and easy administration.
- Inherent selection biases: topical self-selection and economic self-selection.
- An empirical comparison of benchmark data grounded in a comprehensive population registry with:
  - **Two river samples** (Facebook and web-based sample) and
    - (**Convenience** Sampling -- Potential respondents are invited via ads and invitations online.)
  - **One panel sample** (from a major survey research company)
    - (The method of first selecting a group of participants through a **random** sampling method and then asking that group for information.)

# #2: SOCIAL MEDIA, WEB, AND PANEL SURVEYS

***Using Non-Probability Samples in Social and Policy Research***

- The river samples (**convenience** sampling) diverge from the benchmark on demographic variables and yield much higher frequencies on non-demographic variables, even after demographic adjustments
  - Attributed to topical self-selection.
- The panel (**random**) sample is closer to the benchmark.
- No differences between samples when examining the characteristics of a non-demographic subpopulation.
- **CONCLUSION:** "Non-probability online surveys do not replace probability surveys, but augment the researcher's toolkit with new digital practices, such as exploratory studies of small and emerging non-demographic subpopulations."

# #3: INTEGRATING PROBABILITY AND NONPROBABILITY SAMPLES FOR SURVEY INFERENCE

- Survey data collection costs have risen to a point where many survey researchers and polling companies are **abandoning large, expensive probability-based samples in favor of less expensive nonprobability samples.**

- **Is there a way to bridge the gap** and potentially incorporate the timeliness and cost-effectiveness of nonprobability sampling with the accuracy and unbiasedness of probability sampling?

- This article proposed a method of combining these approaches in a way that exploits their strengths to overcome their weaknesses within a Bayesian inferential framework.

# #3: INTEGRATING PROBABILITY AND NONPROBABILITY SAMPLES FOR SURVEY INFERENCE

- Used simulated data to evaluate supplementing inferences based on small probability samples with prior distributions derived from nonprobability data.

- **CONCLUSION: Informative priors based on nonprobability data can lead to reductions in variances and mean squared errors for linear model coefficients.**

- The method is also illustrated with real probability and nonprobability survey data.

# #3: INTEGRATING PROBABILITY AND NONPROBABILITY SAMPLES FOR SURVEY INFERENCE

- This method can lead to **cost savings** for a fixed variance (or MSE) if the nonprobability sample units are significantly cheaper to interview than the probability sample units.

- **Computational efficiency:** It is easily implemented using freely available software such as R and any statistical software that allows Bayesian inference and specification of the prior distributions for the linear regression model.

- *NOTE: The R code used to obtain results is available in the online supplementary material.*

# ADDITIONAL ARTICLES

### #4: *Inference From Non-probability Surveys With Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques*

"The results show that **statistical matching outperforms** PSA in terms of bias reduction and Root Mean Square Error (RMSE), and that **simpler prediction models**, such as linear and k-Nearest Neighbors, provide **better** outcomes than bagging algorithms."

### #5: *Measures of Selection Bias in Regression Coefficients Estimated From Non-probability Samples*

"Developed **model-based indices of selection bias for regression coefficients** estimated from non-probability samples and evaluated the utility of these indices in different settings."
"This work has important implications for other studies in a variety of disciplines that are employing so-called big data, large volunteer samples, or **convenience samples to make statements about relationships between variables in target populations**, especially concerning genetics and genomics."

# A Simulation Study to Compare the Bias and Efficiency of Sample Mean Obtained from Probability and Nonprobability Samples

Sarah Grunblatt, Harun Mazumder, & Chaoyi Zeng

# METHODS

# MEASURES OF SELECTION BIAS IN REGRESSION COEFFICIENTS ESTIMATED FROM NON-PROBABILITY SAMPLES

Selection bias is a serious potential problem for inference about relationships of scientific interest based on samples without well-defined probability sampling mechanisms. Motivated by the potential for selection bias in (a) estimated relationships of polygenic scores (PGSs) with phenotypes in genetic studies of volunteers, and (b) estimated differences in subgroup means in surveys of smartphone users, we derive novel measures of selection bias for estimates of the coefficients in linear regression models fitted to non-probability samples, when aggregate-level auxiliary data are available for the selected sample and the target population. The measures arise from normal pattern-mixture models that allow analysts to examine the sensitivity of their inferences to assumptions about non-ignorable selection in these samples. We examine the effectiveness of the proposed measures in a simulation study, and use them to quantify the selection bias

# INFERENCE FROM NON-PROBABILITY SURVEYS WITH STATISTICAL MATCHING AND PROPENSITY SCORE ADJUSTMENT USING MODERN PREDICTION TECHNIQUES

Online surveys are increasingly common in social and health studies, as they provide fast and inexpensive results in comparison to traditional ones. However, these surveys often work with biased samples, as the data collection is often non-probabilistic because of the lack of internet coverage in certain population groups and the self-selection procedure that many online surveys rely on. Some procedures have been proposed to mitigate the bias, such as propensity score adjustment (PSA) and statistical matching. In PSA, propensity to participate in a nonprobability survey is estimated using a probability reference survey, and then used to obtain weighted estimates. In statistical matching, the nonprobability sample is used to train models to predict the values of the target variable, and the predictions of the models for the probability sample can be used to estimate population values. In this study, both methods are compared using three datasets to

# ALL CAPS

?

# ALL CAPS

?

# ALL CAPS

?

# ALL CAPS

?

# ALL CAPS

?

# ALL CAPS

?

# PURPOSE

# ALL CAPS

?

# ALL CAPS

?

# ALL CAPS

?

# ALL CAPS

?

# ALL CAPS

?

# DISCUSSION

# ALL CAPS

?

# ALL CAPS

?

# ALL CAPS

?

# ALL CAPS

?

# ALL CAPS

?

# ALL CAPS

?

# CONCLUSION

# ALL CAPS

?

# ALL CAPS

?

# ALL CAPS

?

# ALL CAPS

?

# ALL CAPS

?

# ACKNOWLEDGEMENTS


- Professor Oral

# WORKS CITED

Trochim, W. M., & Donnelly, J. P. (2001). Research methods knowledge base.

# A Simulation Study to Compare the Bias and Efficiency of Sample Mean Obtained from Probability and Nonprobability Samples

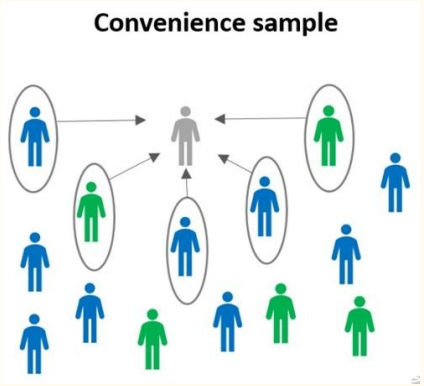
Sarah Grunblatt, Harun Mazumder, & Chaoyi Zeng

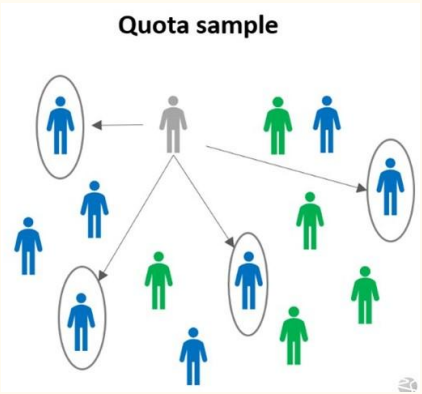
# ALL CAPS

?

# BASIC NON-PROBABILITY SAMPLING TECHNIQUES

1. **Convenience Sampling --** Collecting a sample via the easiest possible manner, often at one location and at one time to whomever happens to be available.
2. **Judgement or Purposeful Sampling --** Researchers choose participants to enroll based on meeting certain criteria based on the study's focus. Ex. Religion
3. **Quota Sampling --** Enrolling every participant until the proportion of the groups in the sample matches the proportion of the groups in the population.
4. **Snowball Sampling --** Participants recruit other members for the study. Often implemented with hard-to-reach populations.




https://dataz4s.com/statistics/non-probability-sampling/

# STATISTICAL DATA INTEGRATION IN SURVEY SAMPLING: A REVIEW

Finite population inference is a central goal in survey sampling. Probability sampling is the main statistical approach to finite population inference. Challenges arise

due to high cost and increasing non-response rates. Data integration provides a timely

solution by leveraging multiple data sources to provide more robust and efficient inference than using any single data source alone. The technique for data integration varies

depending on types of samples and available information to be combined. This article

provides a systematic review of data integration techniques for combining probability

# A REPLICATION APPROACH TO CONTROLLED SELECTION FOR CATCH SAMPLING INTERCEPT SURVEYS

Highlights

- We examined a constrained draw replication approach compared to traditional probability sampling for intercept surveys.
- A simulation study was used to compare the accuracy, precision and bias of catch estimates derived from both designs.
- At high replication levels, the new method produced equally precise, design unbiased estimates as the traditional design.
- This approach may help fisheries scientists incorporate additional customizable constraints into their base survey designs.

# HOUSEHOLD SAMPLING DESIGNS: DIFFERENCES AND SIMILARITIES BETWEEN PROBABILITY SAMPLING AND ROUTE AND QUOTA SAMPLING.

This paper compares the representation quality of three face-to-face household surveys. Two of them used probability samples and the other one selected the ultimate sampling units by using random route and quota sampling, with non-responses resulting in 'automatic' substitutions. The hypothesis to be tested is that random route sampling and quota sampling (with substitution) provide similar representative quality as home sampling (without substitution) based on the local population register. Marked differences were found in education level in the probability samples, where the deviations exceeded 25%. A different picture emerged when comparing employment variables, where quota sampling overestimated both the labour force participation rate (by 2.5% points) and unemployment rates (9.5% points).

# CALIBRATING NON-PROBABILITY SAMPLES WITH PROBABILITY SAMPLES USING LASSO

Due to declining telephone survey response rates, it has become challenging for election pollsters to capture voting intentions in a timely way. This has lead to the expanded use of samples obtained from non-probability web surveys. Because nonprobability samples can suffer from selection bias, we develop a model-assisted calibration method using adaptive LASSO regression – estimated-controlled LASSO (ECLASSO). This method yields consistent estimates of population totals as long as a subset of the true predictors is included in the prediction model, thus allowing large numbers of possible covariates to be included without risk of overfitting. We apply ECLASSO to predict the voting results for the U.S. 2014 midterm election.
Key Words: Probability survey; Propensity weighting; General regression estimator; Model-assisted calibration; Election polls. 1. Introduction

Non-probability samples are an increasing part of life for the survey analyst. This is

# GIS/GPS-ASSISTED PROBABILITY SAMPLING IN RESOURCE-LIMITED SETTINGS

It is rather challenge to draw probability samples for epidemiology and global health research that involves specific geographic area and resource-limited countries and regions. Based on authors' published work, in this chapter we introduce an innovative probability sampling method using the GIS technology for probability spatial sampling, the GIS and GPS technologies to connect the sampled geographic area with residential houses and residents, and the random digits method to select individual participants. With this method, data requirement and cost are minimized while implementation can be achieve in a short period. Most part of the method has been tested and used in a developing country to sample rural residents, rural-to-urban migrants and urban residents.

# DOUBLY ROBUST INFERENCE WHEN COMBINING PROBABILITY AND NON-PROBABILITY SAMPLES WITH HIGH-DIMENSIONAL DATA

We consider integrating a non-probability sample with a probability sample which provides high-dimensional representative covariate information of the target population. We propose a two-step approach for variable selection and finite population inference. In the first step, we use penalized estimating equations with folded-concave penalties to select important variables and show the selection consistency for general samples. In the second step, we focus on a doubly robust estimator of the finite population mean and reestimate the nuisance model parameters by minimizing the asymptotic squared bias of the doubly robust estimator. This estimating strategy mitigates the possible first-step selection error and renders the doubly robust estimator root-n consistent if either the sampling probability or the outcome model is correctly specified.

# STATISTICAL ANALYSIS WITH NON-PROBABILITY SURVEY SAMPLES

***Developing inferential procedures with non-probability survey samples***

Nevertheless, non-probability survey samples are biased samples, from which no valid inferences about the target population can be obtained immediately. A popular tool for bias correction is the propensity score associated with each unit in the population, which is defined as the probability of selection conditional on observed auxiliary variables. Propensity scores need to be estimated in practice, but existing estimation methods are mainly derived on an ad hoc basis. This thesis establishes a general framework for statistical inferences with non-probability survey samples when relevant auxiliary information is available from a reference probability survey sample. Under this framework, we develop a rigorous procedure of estimating propensity scores. The main idea of the procedure is to approximate the required but unknown population-level information by its estimate based on the reference sample.

# CRITICAL REVIEW OF SAMPLING TECHNIQUES IN THE RESEARCH PROCESS IN THE WORLD

Sampling techniques are the component of research which play great role in validity of the research result. Without good sampling good research conduction is impossible. It has two major types namely probability and non probability sampling. The probability sampling consists of simple random sampling, systematic sampling, stratified sampling, and multi stage sampling while in non probability sampling quota sampling, cluster sampling, purpose sampling, judgment sampling, snow ball sampling, expert sampling and convenience samplings are included Seeing to its importance the present study was arranged since, 8th April, 2020. The major objective was that to critically review the sampling techniques in the research process in the world. Total 14 articles were studied and analyzed the situation what methodology is better for conducting research. Hundred percent respondents told that the probability sampling is better than the non probability sampling for conducting research but this methodology is more expensive and time consuming which further delay the result of the study while the non probability sampling is not more expensive and time consuming in the world but its result is doubtful and does not mostly valid for implementation of generalization for population from which the sample has been selected. The result further explore that both methodologies have different advantages on their places but it is necessary for researchers to use proper methodology for their research and analyze the situation for the solution of problems. Research is a systematic and objective attempt for the solution of the problems which play great role for the development of a country and without good research the development of the country is impossible. Now a day the weighted economies of the world are China, America, India and Japan. They all conduct research for the development of their economy enhancement. The developed