Journal of Business Research 66 (2013) 1267–1271

Contents lists available at SciVerse ScienceDirect

Journal of Business Research

A combinatorial optimization based sample identification method for group comparisons ☆

Robyn L. Raschke [a,*], Anjala S. Krishen [b], Pushkin Kachroo [c], Pankaj Maheshwari [d]

[a] University of Nevada, Las Vegas, Department of Accounting, 4505 S. Maryland Parkway, Las Vegas, NV 89154, United States
[b] University of Nevada, Las Vegas, Department of Marketing, United States
[c] University of Nevada, Las Vegas, Department of Electrical and Computer Engineering, United States
[d] University of Nevada, Las Vegas, Department of Civil and Environmental Engineering, United States

ARTICLE INFO

Article history:
Received 1 April 2011
Received in revised form 1 September 2011
Accepted 1 November 2011
Available online 25 February 2012

Keywords:
Sample identification
Sample selection
Sample location identification
Nonprobability samples

ABSTRACT

Researchers often face having to reconcile their sample selection method of survey with the costs of collecting the actual sample. An appropriate justification of a sampling strategy is central to ensuring valid, reliable, and generalizable research results. This paper presents a combinatorial optimization method for identification of sample locations. Such an approach is viable when researchers need to identify sites from which to draw a nonprobability sample when the research objective is for comparative purposes. Findings indicate that using a combinatorial optimization method minimizes the population variation assumptions based upon predetermined demographic variables within the context of the research interest. When identifying the location from which to draw a nonprobability sample, an important requirement is to draw from the most homogeneous populations as possible to control for extraneous factors. In comparison to a standard convenience sample with no identified location strategy, results indicate that the proposed combinatorial optimization method minimizes population variability and thus decreases the cost of sample collection.

1. Introduction

Academics conducting both survey and experimental research must often weigh the costs and benefits of their sampling strategy. Because sampling can have an impact on the validity of research results, a defensible strategy is necessary (Ferber, 1977). The collection method for the data determines the classification of the sample as either a probability or a nonprobability sample. Probability sampling (e.g., simple random, stratified, or systematic) indicates that every element in the population has a known probability of being chosen in the sample for that survey. Thus, a key benefit of probability sampling is the ability to generalize the results, which allows for an estimate of the sampling error. However, probability sampling can require significant resources in both time and money. Unlike probability sampling, nonprobability sampling (e.g., convenience, quota, or judgmental) indicates that every element in the population does not have a known probability of being chosen in the sample for that survey. Therefore, the results are not as generalizable and the sampling error cannot be estimated. But, nonprobability sampling generally is less costly.

The differences between probability and nonprobability sampling are very clear and allow researchers an evaluation criterion to determine an appropriate sampling method. When faced with limited time and money, researchers usually choose the nonprobability sampling method. However, even when a nonprobability sample is the choice, the relation between variability and precision remains. Therefore, if a nonprobability sample comes from a highly variable population, the precision of the results can be in question. If the purpose of the research is for comparison (i.e., to examine the differences between two or more diverse groups of people), homogeneity of the different groups is of utmost importance. Thus, researchers need to minimize demographic differences as much as possible.

The purpose of this paper is to demonstrate a combinatorial optimization method for identifying potential data collection locations for a nonprobability sample. The substantive context of this method comes from a research project aimed at understanding the differences between urban and rural residents and their perceptions of a potential transportation tax policy. The next section of the paper describes the importance of an appropriate sampling strategy when handling targeted group comparisons. Following this, the paper presents the sample identification location problem in a substantive context that details the results of the combinatorial optimization method and demonstrates that this method provides a reasonable strategy as opposed to simply selecting a convenient location for a nonprobability sample. Next, the paper concludes with a discussion of the sampling strategy considerations necessary and the practical implications of this method.

☆ The authors thank Myla Bui-Nguyen, Loyola Marymount University; Angeline Close, University of Texas Austin; Nadia Pomirleanu, University of Nevada Las Vegas, and JBR reviewers for reading and comments of an early version of this article.

* Corresponding author. Tel.: +1 702 895 5756; fax: +1 702 895 4306.
E-mail addresses: robyn.raschke@unlv.edu (R.L. Raschke), anjala.krishen@unlv.edu (A.S. Krishen), pushkin@unlv.edu (P. Kachroo), pankaj47@gmail.com (P. Maheshwari).

0148-2963/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbusres.2012.02.024

2. The importance of sampling strategy

2.1. Sampling strategies for targeted group comparisons in survey research

For decades, social science researchers have debated the tradeoffs associated with obtaining accurate data, setting up valid experiments, and achieving reliable measures. For a research study to be accurate, the findings must be both reliable and valid. Reliability means that the findings are consistently the same even if researchers repeat the study; and validity refers to the truthfulness of the findings, which means that the study actually measures the intended elements (Calder, Phillips, & Tybout, 1982). Although many different threats to validity as well as reliability exist, internal validity is an important early consideration. Internal validity refers to the choice of the most appropriate research design for the topic of study (i.e., experimental, quasi-experimental, survey). This study deems survey research as an appropriate method, and finds that the external validity threat of selection bias impacts how well inferences from the results of the research generalize to the target population and how confident this generalization is. An issue specifically arises when nonprobability sampling is the most cost effective and realistic choice of the researcher, which makes the sampling strategy a difficult decision. Thus, researchers often face tradeoffs to achieve validity and reliability in their studies while trying to justify their choices.

To further explain the differences between selecting probability and nonprobability sampling strategies, a comparison to statistical theory is most appropriate. For probability sampling, the expectation is that, if researchers repeat the sample, then they achieve similar results and make the same sampling inferences. The only difference between the selection and the non-selection of units in the sample is the start of the random number generator. Because the sample is a finite set, probability determines the selection of a unit in the sample. This model is a design based sampling model that uses a randomization theory approach that does not need distributional assumptions. In contrast, in nonprobability sampling, the probability does not determine the selection of the units. This sampling method is a model based approach where, if the model is not true, then sampling estimates might be severely biased (Lohr, 1999).

Justification of a valid sampling method becomes even more critical when researchers seek to compare the subjective or objective characteristics of two or more homogeneous groups (Mullen, Budeva, & Doney, 2009). For example, in the cross-cultural domain, sampling strategies include convenience student sampling for experimental designs (Mikhailitchenko, Javalgi, Mikhailitchenko, & Laroche, 2009; Ueltschy, Laroche, Zhang, Cho, & Yingwei, 2009), multi-stage random sampling for descriptive survey research (Rojas-Méndez, Davies, & Madran, 2009), convenience sampling with locals for descriptive survey research (Chang & Hsieh, 2006), and restricted student sampling for descriptive survey research (Lopez, Babin, & Chung, 2009). Under ideal conditions, to achieve a sample that is representative of the comparative groups of interest, researchers must divide the population into meaningful subpopulations, or strata, that coincide with the domain context of the study. For example, if the purpose of a study is to compare educational workforce experiences of female and male engineering graduates, the basis for the strata is gender (McIlwee & Robinson, 1992). Identification of the strata specifically makes the sampling strategy more efficient if the populations of male and female engineering graduates are not equal, because random sampling from each subgroup allows the researchers to obtain more precision for their comparative groups. Thus, precision means that the variance within each subgroup is more likely to be lower than when compared to the variance in the whole population.

However, a probability sampling strategy can be too costly or impractical, leaving researchers with no choice but to select a nonprobability sampling strategy for comparing groups. The strategy is similar to probability sampling in that, initially, the strategy identifies meaningful subgroups where the variance is minimal. The specific context in this study compares the perceptions among rural and urban residents of a potential transportation tax policy, but a random selection of multitudes of locations throughout the state to draw the sample from is too costly and is not feasible. Therefore, the first challenge is to consider the optimal number of sample locations while minimizing cost and, secondly, to identify those locations that are most representative of the criteria for data collection.

2.2. Sample location identification problem: rural and urban residents in a state

The overall objective of this study is to determine what perceptual differences exist, if any, between constituents residing in urban versus rural counties within the state of Nevada with regard to a potential transportation tax policy. Prior social sciences research that uses inter-group analysis within the socio-cultural context indicates that communication and technological innovations significantly polarize rural and urban residents (Penz, 2006). This research supports the concept that these groups should also be targeted in different ways with regard to these innovations. A basis for a reasonable method is to create strata from urban and rural regions in Nevada to sample. However, the issue remains to identify which locations in both urban and rural areas are potentially representative enough for data collection. The fact that Nevada is a geographically dispersed state complicates this problem. The southern and northern regions of the state both include large populations of urban and rural residents. The evidence of these populations is the location of the two interstates that go through the northern and southernmost regions of the state (I-80 and I-15 respectively). However, no interstate connects the northern regions of the state to the southern regions. Because of the dispersion within the state, the problem centers on the extent to which a convenience sample can truly represent the population groups of interest for comparison.

To achieve reliable results for comparative purposes and to control for extraneous factors, identification of urban and rural locations within the state that accurately represent these two homogeneous groups is important. The use of a combinatorial optimization method helps to determine the total number of locations and the specific locations that are most representative of the population for the comparison groups. Operations research and engineering use combinatorial optimization with the primary goal of selecting the optimum from a set of finite objects (Schrijver, 2005).

3. A combinatorial optimization method for sample location identification

The sampling method that this paper provides is a nonprobability bi-level stratified cluster sampling technique. The first level of stratification is the division into rural and urban areas. The second level of stratification comes from dividing the urban and rural areas into their various counties. This project for data collection has limited funding, so managing costs is a major criterion. Therefore, a simple random sampling that involves collecting data at distributed geographical locations is not feasible. In light of this, the project uses census data from 2008 for the different counties to help identify sampling locations. Because the research context relates to individual perceptions of a potential transportation tax policy (Krishen, Raschke, & Mejza, 2010), the project collects the following variables in relation to working populations for each county: average travel time to work, mean household income, percentage of high school graduates, and percentage of population between the ages of 18 and 65.

To identify the appropriate county and locations, the project formulates a combinatorial optimization method to show the representation factor and cost factor in the analysis. The state represents a fixed number of locations, and the objective is to minimize the