CONTROLLED SAMPLING FOR COMPARING TWO POPULATIONS - AN EXAMPLE

Karol P. Krótki, Lorraine Porcellini, I.S.R, Temple University
Karol P. Krótki, 1601, N.Broad Street, Philadelphia, PA 19122.

Key Words: Sampling, Controlled, Two-way Comparison.

Introduction

When the objective of a study is to compare with respect to one variable two populations that are inherently different along related dimensions, a standard probability sample is not always appropriate since there is no control over extraneous variables that may be related to the key indicators of interest. This paper discusses a sample designed to study the effect of ethnicity on drinking behavior by comparing Irish and Puerto Ricans in New York City. A random sample of both populations would have almost certainly produced an Irish sample with a higher average socioeconomic level than the Puerto Rican sample. Furthermore, since the total sample size was only 1,000, limiting the analysis to the comparable sub-populations would have resulted in unacceptably small effective sample sizes. Controlled sampling was used to favor selection of areas which according to census data contained members of each ethnic group of similar socioeconomic background[1]. The paper concentrates on the preparation of the sample frame which is the innovative aspect of this design.

Background

The 1990-91 Survey of Drinking Behavior of Irish and Puerto Ricans in New York is being carried out by the Hispanic Research Center of Fordham University with the survey methodology sub-contracted out to the Institute for Survey Research.

The survey objectives call for a sample of 500 Puerto Rican and 500 Irish adults at ages 18 and over residing in households in the Greater New York City Metropolitan Area (GNYCMA).

The primary substantive objective of this survey is to measure the effect of ethnicity on drinking behavior.

The Sampling Frame

The need for an innovative approach to sampling became apparent when it was realized that straight probability samples of Puerto Ricans and Irish would result in two samples differing widely with respect to socioeconomic status.

Ten counties within the GNYCMA were included in the survey: 3 in New Jersey (Essex, Hudson, and Passaic) and 7 in New York (Bronx, Kings, Nassau, New York, Queens, Richmond, and Suffolk). 1980 Census data are available for the block groups (BGs) in these counties to assist in the creation of the sample frame. The 3 New Jersey counties consist of 1,481 BGs while the 7 counties in New York include 8,546 BGs.

Since both populations are relatively rare (13% Irish and 9% Puerto Rican), it was decided to limit the universe to those BGs with at least 40% of the population belonging to one of the two ethnic groups. This measure was taken in order to keep field costs at a reasonable level. It would have been prohibitively expensive to do a straight area probability sample of the entire GNYCMA and then to screen each household selected for eligible Irish and Puerto Rican adults.

Table 1 presents the total number of BGs together with the Irish and Puerto Rican BGs by state.

TABLE 1

Total, Irish, and Puerto Rican BGs in the GNYCMA by State

	New Jersey	New York
Total No. BG's	1,481	8,546
Irish	10	277
Puerto Rican	67	680

Since the New Jersey BGs represent a small proportion of the total number of BGs it was decided to exclude New Jersey from the universe.

Table 2 presents demographic and socioeconomic characteristics of the two universes of BGs in New York.

Table 2
Socioeconomic Characteristics of Irish and Puerto Rican BGs in the GNYCMA (NY State)
(Figures in brackets are standard deviations.)

	Irish	Puerto Rican
Number of BGs	277	680
Total Population	253,020	702,948
Ethnic Pop	115,647	384,111
Total Hhlds	83,434	237,626
Ethnic Hholds	38,254	114,278
Income median	$22,648 ($6,852)	$7,802 ($3,689)
Income,mean	$24,510 ($7,238)	$9,859 ($4,313)
Income Index[1],mean	11.56 (3.123)	4.54 (1.520)
Education[2],mean	12.98 (1.52)	11.60 (1.17)
Combined Index[3],mean	20.43 (1.98)	20.44 (3.74)

Notes:
1.Categorical version. Range 1-20. Each category representing $2,000.
2.Categorical version of % persons in BG who have 12 or more years of education. Each category represents 5%. Range 1-20.
3.Sum of the the categorical versions of income and education. Range 2-40.
Source:Summary Tape STF3A, 1980 Census of Population and Housing, U.S. Census Bureau, Washington, DC.

The table presents overall counts and means for income and education. A combined index is also calculated as the sum of the categorical versions of income and education. The table clearly shows to what extent population data for the two ethnic groups differ with respect to two basic socioeconomic variables. The median and mean incomes for the Irish BGs are more than double the Puerto Rican levels, both for the continuous and categorical versions of this variable. Even adding and subtracting twice the standard deviation gives intervals that do not overlap. Similarly, extreme differences exist for the education variable and for the combined index. The extent of the overlap can be seen in Fig. 1.

Analysis Considerations

It is obvious that straight probability sampling from these two universes would have produced two completely different samples, with almost no overlap in terms of socioeconomic status. Whereas this outcome, in and of itself, should not be a cause of concern, it does negatively impinge on the analysis that can be done of the effect of ethnicity on drinking behavior.The total sample size for each ethnic group is only 500. Given that the primary topic of interest is the effect of ethnicity on drinking it is necessary to control for extraneous variables, one of the most important being the battery of socioeconomic variables. Since in this case the sample size is small and the disparity between the two groups large, it would be impossible to carry out such a controlled analysis based on the data collected under this sample design.

The issue then becomes one of the relative merits of "comparability" and "representativity". Any probability or representative sample in this case would undoubtedly lead to a data base unsuitable for the type of analysis for which it was intended. On the other hand, a comparable pair of samples will of necessity not be representative of the universes, the Irish or Puerto Ricans in the GNYCMA[2].

The survey sponsors considered comparability to be a more important criterion than representativity. The study is somewhat experimental and preliminary in nature and as such should be designed to permit maximum investigation of the relationship between ethnicity and drinking even

at the cost of diminishing, if not eliminating, the chance of generalizing to a larger population.

Creating Matched Universes

Once the decision had been made to proceed with a non-representative, controlled sample, it was necessary to define the two populations in such a way that the corresponding samples would overlap maximally.

The matching was done at the BG level and the above-described combined index was used as the matching criteria. The process began with the two distributions of the Irish and Puerto Rican block groups. Low-score Puerto Rican and high-score Irish block groups were gradually eliminated from the universes in order to create overall means that were as equal as possible. Several trial-and-error approaches culminated in the populations described in Table 3.

Table 3
Socioeconomic Characteristics of the Irish and Puerto Rican BGsin the Matched Universes in the GNYCMA (NY State)
(Figures in brackets are standard deviations.)

	Irish	Puerto Ricans
Number of BGs	61	48
Population	50,876	52,970
Ethnic popn.	24,313	26,750
Households	20,699	17,253
"Ethnic" hhlds	9,878	7,625
Income, median	$13,989 ($4,450)	$16,417 ($5,185)
Income[1], mean	$15,970 ($4,619)	$18,711 ($8,493)
Income[1], mean	7.49 (2.20)	8.69 (2.48)
Education[2], mea	12.93 (2.29)	11.75 (1.79)
Index[3], mean	20.43 (2.59)	20.44 (3.15)

Notes: see Table 2

The matched population is graphically represented in Figure 1.

Tables 4 -7 present numerically the results of the above operations . In the case of the Irish population almost 10 % of this population was included in the ethnic BGs. Furthermore, of this reduced population, a little over 20% was included in the matched universe, which serves as the sample frame. In the case of the Puerto Rican population, these figures are about 40 % and 7% respectively.

FIGURE 1

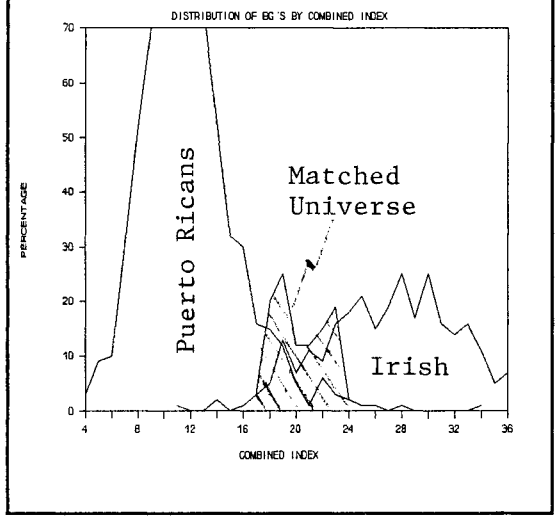


Table 4
Irish Housing Counts

	TOTAL HUS	ETHNIC HUS
Total Area (GNYCMA - New York State)	3,597,650	457,978
Proportion of Total Area HUS in Ethnic Areas -BGs>40% Irish	2.32 %	8.35 %
Proportion of Ethnic Area HUS in Sample Frame	24.81 %	25.82 %
HUS in Sample Frame	20,699	9,878

Table 5
Irish Population Counts

	TOTAL POPN	ETHNIC POPN
Total Area (GNYCMA - New York State)	9,657,946	1,258,083
Proportion of Total Area Popn in Ethnic Areas (BGs>40% IRISH)	2.62 %	9.19 %
Proportion of Ethnic Area Popn in Sample Frame	20.11 %	21.02 %
Popn in Sample Frame	50,876	24,313

Table 6
Puerto Rican Housing Counts

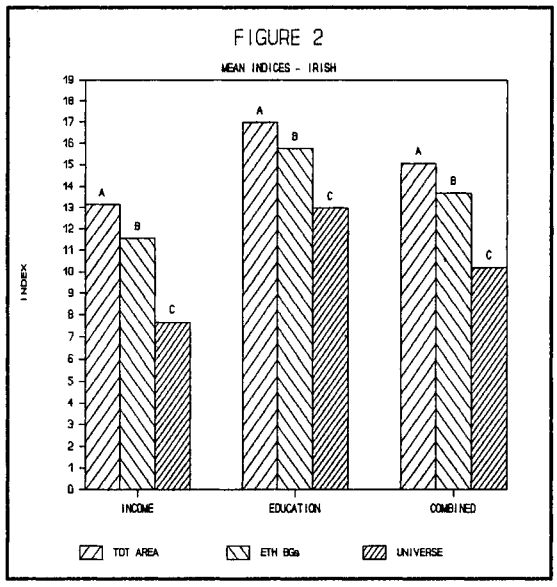
	TOTAL HUS	ETHNIC HUS
Total Area (GNYCMA - New York State)	3,597,650	319.911
Proportion of Total Area HUS in Ethnic Areas (BGs>40% PR)	6.61 %	35.72 %
Proportion of Ethnic Area HUs in Sample Frame	7.26 %	6.67 %
HUs in Sample Frame	17,253	7,625

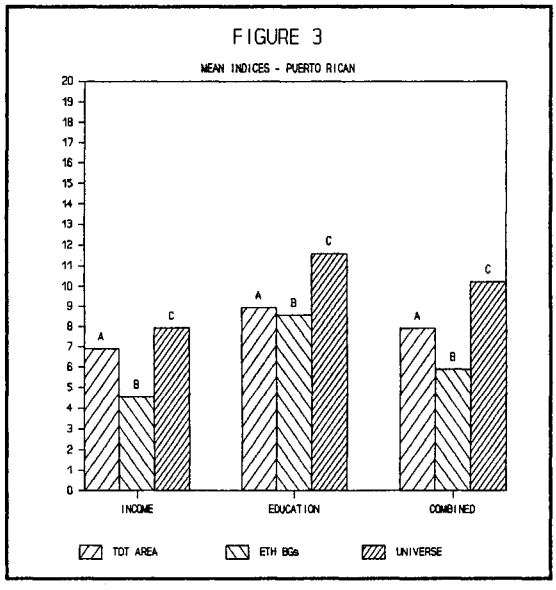
Table 7
Puerto Rican Population Counts

	TOTAL POPN	ETHNIC POPN
Total Area (GNYCMA - New York State)	9,657,946	901,553
Proportion of Total Area popn in Ethnic Areas (BGs>40% PR)	7.28 %	42.61 %
Proportion of Ethnic Area popn in Sample Frame	7.54 %	6.96 %
HUs in Sample Frame	52,970	26,750

A comparison of the income, education and combined indices for the three areas (GNYCMA, Ethnic Areas, And Matched Universe) is provided in Figures 2 and 3.

The heights of the "Matched Universe" or "Sample Frame" columns (C) are the same for both Irish and Puerto Ricans as a result of the matching process. However the first columns (A) are larger for the Irish reflecting their higher overall socio-economic status. In Figure 2 the middle columns (B) are larger than the third (C) for all indices since the lower socioeconomic Irish were selected for the matched universe. The opposite occurs in Figure 3 where the columns (B) are lower than column (C).





It would have been possible to relax the matching criteria and settle for a larger difference in the index between the two populations. This would have allowed for a larger population size (in terms of BGs) but at the expense of slightly more difficult control for socioeconomic status in analysing the effect of ethnicity on drinking behavior.

Since it had already been decided that representativity was of less importance than comparability it was felt that this principle should be

carried to its extreme. Limiting the universe to fewer BGs has no bearing on the quality of the data or the possibilities of data analysis whereas the smaller universe decreases sampling and field costs.

Sample Size and Design

The number of completed interviews is 500 for each ethnic group. This figure represents the "effective" sample size which is only a proportion of the households that must be selected and visited. Factors that must be taken into account are response rates, vacancy rates, and the eligibility rates.

Estimating these values based on 1980 Census data, it was calculated that a little over 2,000 households would have to be selected both in the Irish and Puerto Rican BGs.

A three stage cluster sample was designed using probabilities proportional to size. In the first stage BGs were first ordered by density of ethnic households, education and income to permit implicit stratification into the design. A fixed number of BGs was then selected with probability proportional to total household size. Within BGs, the 1980 Census provides information on the composition by blocks together with the population and households in each. Based on these data, blocks were combined to form listing areas with a minimum number of households. A fixed number of listing areas were selected within each block group again with probability proportional to the total number of households. Finally, in the third stage, a fixed number of households was selected in each listing area.

Conclusion

Whereas the above approach smacks of non-probability sampling in fact the design is probability based and, in theory, is an equal-probability design. Each sample element will carry an equal weight unless practical deviations from the basic design result in a need for weighting. The crucial issue is the departure from representativity, or generalizability to the original population[2]. Nevertheless, it is argued that in this situation, which is not uncommon in social research, the above design can be defended in that it provides data appropriate for analysis, whereas a straight probability sample would have made he desired analysis impossible. When the key objective of a survey is to investigate the relationship between two variables, representativity sometimes must take a back seat to practical considerations. It goes without saying that any analysis carried out on these data must clearly specify the population from which the sample was drawn.

Notes:

1. For a discussion of matching techniques which control extraneous variables through the method of selection, see Moser and Kalton, Survey Methods in Social Investigation, Basic Books Inc., 1972, pp. 220-224.

2. Ibid. p. 236.