YouGov Sampling Methodology

Sampling and Sample Matching

Sample matching is a methodology for selection of representative samples from non-randomly selected pools of respondents. It is ideally suited for Web access panels, but could also be used for other types of surveys, such as phone surveys. Sample matching starts with an enumeration of the *target population.* For general population studies, the target population is all adults, and can be enumerated through the use of the decennial Census or a high quality survey, such as the American Community Survey. In other contexts, this is known as the *sampling frame*, though, unlike conventional sampling, the sample is *not* drawn from the frame. Traditional sampling, then, selects individuals from the sampling frame at random for participation in the study. This may not be feasible or economical as the contact information, especially email addresses, is not available for all individuals in the frame and refusals to participate increase the costs of sampling in this way.

Sample selection using the matching methodology is a two-stage process. First, a random sample is drawn from the target population. We call this sample the *target sample.* Details on how the target sample is drawn are provided below, but the essential idea is that this sample is a true probability sample and thus representative of the frame from which it was drawn.

Second, for each member of the target sample, we select one or more *matching* members from our pool of opt-in respondents. This is called the *matched sample.* Matching is accomplished using a large set of variables that are available in consumer and voter databases for both the target population and the opt-in panel.

The purpose of matching is to find an available respondent who is as similar as possible to the selected member of the target sample. The result is a sample of respondents who have the same measured characteristics as the target sample. Under certain conditions, described below, the matched sample will have similar properties to a true random sample. That is, the matched sample mimics the characteristics of the target sample. It is, as far as we can tell, "representative" of the target population (because it is similar to the target sample).

When choosing the matched sample, it is necessary to find the closest matching respondent in the panel of opt-ins to each member of the target sample. Various types of matching could be employed: exact matching, propensity score matching, and proximity matching. Exact matching is impossible if the set of characteristics used for matching is large and, even for a small set of characteristics, requires a very large panel (to find an exact match). Propensity score matching has the disadvantage of requiring estimation of the propensity score. Either a propensity score needs to be estimated for each individual study, so the procedure is automatic, or a single propensity score must be estimated for all studies. If large numbers of variables are used the estimated propensity scores can become unstable and lead to poor samples.

YouGov employs the proximity matching method. For each variable used for matching, we define a *distance function*, d(x,y), which describes how “close” the values x and y are on a particular attribute. The overall distance between a member of the target sample and a member of the panel is a weighted sum of the individual distance functions on each attribute. The weights can be adjusted for each study based upon which variables are thought to be important for that study, though, for the most part, we have not found the matching procedure to be sensitive to small adjustments of the weights. A large weight, on the other hand, forces the algorithm toward an exact match on that dimension.

Theoretical Background for Sample Matching

To understand better the sample matching methodology, it may be helpful to think of the target sample as a simple random sample (SRS) from the target population. The SRS yields unbiased estimates because the selection mechanism is unrelated to particular characteristics of the population. The efficiency of the SRS can be improved by using stratified sampling in place of simple random sampling. SRS is generally less efficient than stratified sampling because the size of population subgroups varies in the target sample.

Stratified random sampling partitions the population into a set of categories that are believed to be more homogeneous than the overall population, called *strata.* For example, we might divide the population into race, age, and gender categories. The cross-classification of these three attributes divides the overall population into a set of mutually exclusive and exhaustive groups or strata. Then an SRS is drawn from each category and the combined set of respondents constitutes a stratified sample. If the number of respondents selected in each strata is proportional to their frequency in the target population, then the sample is self-representing and requires no additional weighting.

The intuition behind sample matching is analogous to stratified sampling: if respondents who are similar on a large number of characteristics tend to be similar on other items for which we lack data, then substituting one for the other should have little impact upon the sample. This intuition can be made rigorous under certain assumptions.

Assumption 1: Ignorability. Panel participation is assumed to be *ignorable* with respect to the variables measured by survey conditional upon the variables used for matching. What this means is that if we examined panel participants and non-participants who have exactly the same values of the matching variables, then on average there would be no difference between how these sets of respondents answered the survey. This does *not* imply that panel participants and non-participants are identical, but only that the differences are captured by the variables used for matching. Since the set of data used for matching is quite extensive, this is, in most cases, a plausible assumption.

Assumption 2: Smoothness. The expected value of the survey items given the variables used for matching is a "smooth" function. Smoothness is a technical term meaning that the function is continuously differentiable with bounded first derivative. In practice, this means that that the expected value function doesn't have any kinks or jumps.

Assumption 3: Common Support. The variables used for matching need to have a distribution that covers the same range of values for panelists and non-panelists. More precisely, the probability distribution of the matching variables must be bounded away from zero for panelists on the range of values (known as the "support") taken by the non-panelists. In practice, this excludes attempts to match on variables for which there are no possible matches within the panel. For instance, it would be impossible to match on computer usage because there are no panelists without some experience using computers.

Under Assumptions 1-3, it can be shown that if the panel is sufficiently large, then the matched sample provides consistent estimates for survey measurements. The sampling variances will depend upon how close the matches are if the number of variables used for matching is large. In this study, over 150,000 respondents to YouGov's Internet surveys were used for the pool from which to construct the matches for the final sample.

Current Sampling Frame and Target Sample

YouGov has constructed a sampling frame of U.S. Citizens from the 2016 American Community Survey, including data on age, race, gender, education, marital status, number of children under 18, family income, employment status, citizenship, state, and metropolitan area. The frame was constructed by stratified sampling from the full 2016 ACS sample with selection within strata by weighted sampling with replacement (using the person weights on the public use file). Data on reported 2016 voter registration and turnout from the November 2012 Current Population Survey was matched to this frame using a weighted Euclidean distance metric. Data on religion, church attendance, born again or evangelical status, interest in politics, party identification and ideology were matched from the 2014 Pew U.S. Religious Landscape Survey. Characteristics of target samples vary based on the requirements of the projects. Typical general population target samples are selected by stratification by age, race, gender, education, and voter registration, and by simple random sampling within strata. At the matching stage, the final set of completed interviews are matched to the target frame, using a weighted Euclidean distances metric.

Weighting

The matched cases are weighted to the sampling frame using propensity scores. The matched cases and the frame are combined and a logistic regression is estimated for inclusion in the frame. The propensity score function may include a number of variables, including age, years of education, gender, race/ethnicity, predicted voter registration, interest in politics, born again status, ideological self-placement and inability to place oneself on an ideological scale, and baseline party identification (i.e., the profiled party identification that was collected before the survey was conducted). The propensity scores are then grouped into deciles of the estimated propensity score in the frame and post-stratified according to these deciles. The final weights may then be post-stratified by gender, race, education, and age. Large weights are trimmed and the final weights are normalized to equal sample size.