YouGov
What the world thinks

YouGov Panel Recruitment

The YouGov panel, a proprietary opt-in survey panel, is comprised of 1.8 million U.S. residents who have agreed to participate in YouGov's web surveys as well as the YouGov Online community. Participants are not paid to join the YouGov community, but do receive incentives through a points-based loyalty program to take individual surveys. Additionally, YouGov community members can share opinions on essentially any topic in the member forum, read about YouGov's proprietary research and are notified if a study the member participated in has received coverage in the media.

Panel members are recruited by a number of methods to help ensure diversity in the panel population. Recruiting methods include:

- Web advertisement join YouGov campaigns - Web search based advertising in which a respondent is directed to a landing page introducing YouGov and invited to join the community.

- Web advertising public survey campaigns - Web search based advertising in which a respondent is invited to participate in a short survey. At the conclusion of the survey the respondent would be invited to join the YouGov community to directly receive and participate in additional surveys.

- Partner sponsored solicitations – YouGov works with recruitment partners and ad networks to ensure we're getting a steady stream of new traffic. Partners will use a mix of banner advertising, co-registration and permission based email campaigns to recruit for the YouGov community.

- Member referrals – The YouGov loyalty program includes a member referral system. Members are encouraged to refer their friends to join YouGov and are rewarded with points in their loyalty program account for each referred member.

- Organic recruitment – The YouGov marketing department actively works to increase the media presence of YouGov via promoting published proprietary research. In practice, an individual may find one of our published articles online and then decide to join the YouGov community out of interest.

- Telephone-to-Web recruitment (RDD based sampling) - Telephone to Web recruitment is used on rare occasion to recruit niche audiences to the YouGov community. In the spring and summer of 2015, YouGov completed telephone interviews using RDD sampling and invited respondents to join the online panel. The goal was to boost membership among some geographical areas in anticipation of the upcoming presidential primaries. Respondents provided a working email where they could receive an electronic invitation and confirm their consent and interest in receiving and participating in YouGov Web surveys

All recruited members must pass through a double opt-in procedure, where respondents must confirm their consent again by responding to an email, the database checks to ensure the newly recruited panelist is in fact new and that the address information provided is valid.

1. Description of YouGov's sample matching and propensity score weighting methodology

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 Census.

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 stratified sample of size n is drawn from the frame. The frame is a synthetic sampling frame from the general population, such as census data. The frame is divided in stratums that are defined by crosstab by important demographic variables. And within each stratum, cases are randomly sampled to construct a probabilistic sample which serves as a target population in the sampling matching stage. 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-to-one match from our pool of opt-in respondents by nearest neighbor matching using Mahalanobis distance metric. 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 matched sample is a good approximation to the probabilistic sample of the frame. An estimated propensity score and post-stratification can be used to produce weights for the matched sample to further adjust the sample.

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).

The advantage of sample matching is that it is feasible to balance a non-random sample of a large number of covariates to reduce potential bias. The resulting samples are much less expensive to collect. Under an assumption of ignorability and some technical conditions, the matching estimator is consistent, asymptotically normally distributed, with a covariance matrix that can be estimated robustly. In this case, sample matching not only makes the marginal distribution of selected covariates less biased, it also addresses the joint distribution of these covariates.

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.

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.

The matched sample could be further weighted to reduce bias using propensity score weighting and post-stratification. Logistic regression is the technique that is generally associated with propensity scores, and it is used to determine the probability of membership in the treatment or control group, in our case the frame or matched sample, given the specific set of selection variables included. Variables commonly included are demographic variables and/or political variables. The matched cases and frame are combined and a logistic regression is estimated for inclusion in the frame. Then propensity scores are grouped into deciles of the estimated propensity score in the frame and post-stratified according to these deciles. After the propensity scores are used, the weighted matched sample should have similar characteristics to the synthetic frame of the target population.

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. YouGov has previously validated this approach in peer-reviewed research (Rivers, Sampling for web surveys, Joint Statistical Meetings, 2007).

2. Calculation of margin of error with sample matched data

Following from the above, margin of errors can in fact be estimated with non-probabilistic samples provided they have been subjected to propensity score matching.

Margin of error is the product of the critical value and the standard error. The critical value is the 97.5% quantile of the t-distribution, with degrees of freedom equal to the sample size minus one. The standard error is the square root of the variance. For propensity score matched data, the variance equals 1 plus the variance of the weights times θ *(1- θ), and then divided by sample size.

The full formula is:
qt(0.975, N - 1) * sqrt((1 + sd(weight)^2)) * sqrt(θ *(1- θ) / N)
where θ = 0.5, accounting for the estimation of maximum variance.

3. Third party validation of YouGov's methodology

YouGov's sample matching approach has been used for many years in conjunction with our proprietary online panels, and we have consistently been shown to be highly accurate when we have been compared with some of our competitors.

One example from 2016 was a head-to-head comparison of various polling companies (using non-probability online surveys) alongside a true probabilistic polling methodology (telephone random digit dialing – RDD), performed by Pew Research. The results indicated that the YouGov panel was in many cases more accurate than the probabilistic sampling. In addition, the YouGov sampling methodology was more accurate than any of its competitors (see https://today.yougov.com/topics/finance/articles-reports/2016/05/13/pew-research-yougov and associated links for a full description the study).

More recently, the New York Times compared the performance of various survey vendors following the 2018 US Congressional election. While generally skeptical of many online survey vendors (which do indeed vary greatly in the quality of the data they produce), the author of this report did highlight YouGov as being exceptional in the rigour of its methodology and the accuracy of its polling results (see https://www.nytimes.com/2019/07/02/upshot/online-polls-analyzing-reliability.html).