

## The Importance of Probability Sampling

Sampling should be designed to guard against unplanned selectiveness. A replicable or repeatable sampling plan should be developed to randomly choose a sample capable of meeting the survey's goals. A survey's intent is not to describe the particular individuals who, by chance, are part of the sample, but rather to obtain a composite profile of the population of interest. In a bona fide survey, the sample is not selected haphazardly or only from persons who volunteer to participate. It is scientifically chosen so that each person in the population will have a measurable chance of selection – a known probability of selection. This way, the results can be reliably projected from the sample to the larger population with known levels of certainty/precision, i.e. standard errors and confidence intervals for survey estimates can be constructed.

Critical elements in an exemplary survey are: (a) to ensure that the right population is indeed being sampled (to address the questions of interest); and (b) to locate (or "cover") all members of the population being studied so they have a chance to be sampled. The quality of the list of such members (the "sampling frame") whether it is up-to-date and complete is probably the dominant feature for ensuring adequate coverage of the desired population to be surveyed. Where a particular sample frame is suspected to provide incomplete or inadequate coverage of the population of interest, multiple frames should be used.

Virtually all surveys taken seriously by social scientists, policy makers, and the informed media use some form of random or probability sampling, the methods of which are well grounded in statistical theory and the theory of probability. Reliable and efficient estimates of needed statistics can be made by surveying a carefully constructed sample of a population, provided that a large proportion of the sample members give the requested information. The latter requires that careful and explicit estimates of potential non response bias and sample representativeness be developed.

Non-probability samples refer to samples where the sampling frame is not well-defined and there is no known probability of selection. In other words, there is not a full accounting of the population of interest such that a representative sample can be drawn. Mall intercept samples, most e-mail lists, commercial mail panels, and home scanner panels are non-probability samples and for which conventional sampling theory is inapplicable. Because little is known about the sample selection mechanism, the sampling distribution of sample statistics is unknown. Nonetheless, computation of standard errors under the (incorrect) assumption of random sampling is practiced.

The Knowledge Networks panel is a probability based panel sample that is representative of the U.S. population and all segments of the U.S. population. The panel sample is selected using high quality RDD sampling methods – comparable to those used by the U.S. government's RDD surveys (CDC's National Immunization Survey, for example). Unlike other Internet research that

covers only individuals with Internet access who volunteer for research, Knowledge Networks surveys are based on a sampling frame that includes both listed and unlisted phone numbers, and is not limited to current Web users or computer owners. Panelists are selected by chance to join the panel; unselected volunteers are not able to join the KN panel.

The panel sample is not an equal probability sample as there is intentional over-sampling of certain important subgroups to enhance the reliability of those subgroups and/or control recruitment costs. The Knowledge Networks panel design weights include adjustments that account for the planned design features that ensure that estimates from the panel remain statistically valid for projecting to the U.S. population and all segments of the U.S. population. A post-stratification adjustment is applied to the Knowledge Networks panel design weights to reduce variance and minimize bias due to non-sampling error. Distributions for age, race, gender, Hispanic ethnicity, education, and Census Region are used in the post-stratification adjustment. The weights resulting from applying the post-stratification adjustment to the panel design weights are the starting weights for all client surveys.

Depending on the design of the client survey, whether there is an over-sampling of certain subgroups or selection within certain subgroups (e.g. people with high cholesterol), the Knowledge Networks post-stratification weights are adjusted to reflect the design of the client survey. We provide sample weights for all completes from the client survey that sum to either the sample size or the population covered by the survey. We also provide scaled weights for one or more groups of qualified completes from the client survey if desired by the client. We examine the distributions of the weights, and trim the weights at the tails (usually 1% and 99%) to reduce the effect of outliers in data analyses. Other weights can be provided upon request such as the starting weight for the client sample or specialized non-response weights created by post-stratifying the completed sample to important benchmarks from the Knowledge Networks panel.

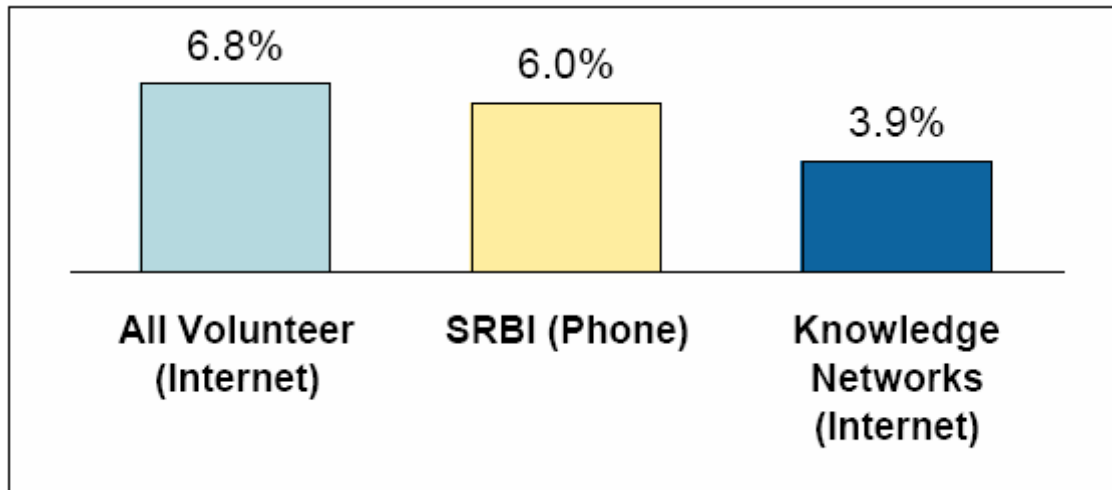
Sample weights for the Knowledge Networks Panel and client surveys selected from the panel should be used for all analyses because of the following reasons:

- Sample weights reflect the sample design of the Knowledge Networks panel
  - Over-sampling of address listed households
  - Over-sampling of Black and Hispanic households
  - Over-sampling of certain states/regions
- Sample weights reflect the sample design of the Client survey
- Sample weights reduce sampling variance – *Using independent, reliable Benchmarks*
- Sample weights reduce bias due to non-coverage, non-response, response error

Sample weights for client surveys are delivered in SPSS, SAS or ASCII file format, depending on client requirements. The weight file includes the unique identifier for each case and the different set of weights computed for the study (e.g., All completes, all qualified completes, all assignees, etc.). The file can be merged with the field data file prior to delivery to the client or sent as a separate file for the client to merge with field data.

In terms of data quality, surveys conducted using the Knowledge Networks panel consistently prove to be very high quality, i.e. low bias due to nonsampling error. The chart below presents average absolute error results from a recent independent comparison study conducted by Stanford University summarizing closeness of survey results to objective Government and U.S. Market benchmarks for about 35 key outcomes. The same survey was administered by 8 different companies during the same time period for an apples-to-apples comparison. The bar labeled “All Volunteer” collapses survey results from the 6 Internet fielded samples that are all non-probability samples. The “SRBI” bar presents results from an RDD probability based sample conducted on the phone and the “Knowledge Networks” bar presents results from the survey selected from the KN panel for the study – a probability based sample -- that was administered over the Internet. Estimates from the KN panel were closer to objective benchmarks than the all volunteer sample estimates and the SRBI RDD phone estimates as well.

**Chart 1. Average Absolute Error (%) from Independent Benchmarks**



To ensure data quality in survey results, have the ability to statistically project survey results to the population of interest and calculate valid estimates of sampling error, probability based samples should be the research tool of choice in the design and fielding of surveys.