2

The Importance of Probability-Based Sampling Methods for Drawing Valid Inferences

Gary Langer

Before 1936, data on populations generally were collected either via a census of the entire population or "convenience" sampling, such as straw polls. The latter, while quick and inexpensive, lacked a scientific, theoretical basis that would justify generalization to a broader population. Using such methods, the Literary Digest correctly predicted presidential elections from 1916 to 1932 – but the approach collapsed in 1936. The magazine sent postcards to 10 million individuals selected from subscriptions, phone books, and automobile registration records. Through sampling and self-selection bias, the 2.4 million responses disproportionately included Republicans, and the poll predicted an easy win for the losing candidate, Alf Landon.

George Gallup used quota sampling in the same election to draw a miniature of the target population in terms of demographics and partisanship. Using a much smaller sample, Gallup correctly predicted Franklin D. Roosevelt's win. This set the stage for systematic sampling methods to become standard in polling and survey research. (See, e.g., Gallup and Rae 1940.)

But quota sampling turned out not to be a panacea. The approach suffered a mortal blow in the 1948 presidential election, when Gallup and others erroneously predicted victory for Thomas Dewey over Harry Truman. While a variety of factors was responsible, close study clarified the shortcomings of quota sampling. Replicating the U.S. population in terms of

G. Langer (✉)
Langer Research Associates, New York, USA
e-mail: glanger@langerresearch.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_2

cross-tabulations by ethnicity, race, education, age, region, and income, using standard categories, would require 9,600 cells, indicating a need for enormous sample sizes. Further, "The microcosm idea will rarely work in a complicated social problem because we always have additional variables that may have important consequences for the outcome" (Gilbert et al. 1977). And bias can be introduced through interviewers' purposive selection of respondents within each quota group.

After spirited debate, survey researchers coalesced around probability sampling as a scientifically rigorous method for efficiently and cost-effectively drawing a representative sample of the population. In this technique, each individual has a known and ideally non-zero probability of selection, placing the method on firmly within the theoretical framework of inferential statistics. As put by the sampling statistician Leslie Kish, "(1) Its measurability leads to objective statistical inference, in contrast to the subjective inference from judgment sampling, and (2) Like any scientific method, it permits cumulative improvement through the separation and objective appraisal of its sources of errors" (Kish 1965).

In modern times, high-quality surveys continue to rely on probability sampling. But new non-probability methods have come forward, offering data collection via social media postings and most prominently though opt-in online samples. These often are accompanied by ill-disclosed sampling, data collection, and weighting techniques, yet also with routine claims that they produce highly accurate data. Such claims need close scrutiny, on theoretical and empirical bases alike.

Opt-in surveys typically are conducted among individuals who sign up to click through questionnaires on the Internet in exchange for points redeemable for cash or gifts. Opportunities for falsification are rife, as is the risk of a cottage industry of professional survey respondents. One study (Fulgoni 2006) found that among the 10 largest opt-in survey panels, 10 percent of panelists produced 81 percent of survey responses, and 1 percent of panelists accounted for 24 percent of responses.

An example of further challenges in opt-in online surveys is their common and generally undisclosed use of routers to maximize efficiency of administration, albeit at the cost of coverage. As an illustration, participants may be asked if they are smokers; if so, are routed to a smoking survey. If not smokers, they may be asked next if they chew gum. If yes, they are routed to a gum-chewers survey. If not, they may next be asked if they use spearmint toothpaste, and so on. Unbeknownst to sponsors of the toothpaste study, smokers and gum chewers are systematically excluded from their sample.

The approach, then, raises many questions. Who joins these poll-taking clubs, what are their characteristics, and what do we know about the reliability and validity of their responses? Are respondent identities verified? Are responses validated? What sorts of quality control measures are put in place? What survey weights are applied, how were they obtained, and what is their effect? What claims are made about the quality of these data, and how are these claims justified?

Purveyors of opt-in online sampling often point to the declining response rates and increasing costs of probability-based telephone surveys, topics that are addressed later in this book. But these arguments are hardly a constructive defense of alternative methodologies, nor do they recognize the wealth of research identifying response rates as a poor indicator of data quality. Rather than pointing toward potential deficiencies in existing methods, it seems incumbent on the proponents of alternative non-probability methods to construct a reasoned defense of the approach, including a theoretical basis for its validity.

Empirical research consistently has found validity in scientific probabilistic sampling methods. Results for non-probability opt-in panels have been more concerning. An extensive review of existing literature, the AAPOR Report on Online Panels, published by the American Association for Public Opinion Research in 2010, recommended that "researchers should avoid nonprobability online panels when one of the research objectives is to accurately estimate population values." This report concluded, "There currently is no generally accepted theoretical basis from which to claim that survey results using samples from nonprobability online panels are projectable to the general population. Thus, claims of 'representativeness' should be avoided when using these sample sources" (Baker et al. 2010). (Subsequent to this presentation, an AAPOR report on non-probability sampling, in 2013, again noted the absence of a theoretical framework that would support statistical inference.)

In a large empirical study in 2011, Yeager and his colleagues compared seven opt-in online sample surveys with two probability sample surveys, finding that the probability surveys were "consistently highly accurate" while the opt-in samples were "always less accurate … and less consistent in their level of accuracy" (Yeager et al. 2011). The authors also found little empirical support for the claim that some non-probability panels are consistently more accurate others. They reported that weighting did not always improve accuracy of these panels, and they found no indication that higher completion rates produce greater accuracy. A report on data produced for a study by the Advertising Research Foundation found similar problems, as did

an independent analysis of 45 individual data quality studies (Baker 2009; Callegaro et al. 2012). These confirm the fundamental issue: the absence of theory that would predict accurate, reliable results from non-probability samples.

Even if they can't be used to generalize about a broader population, it has been suggested that non-probability approaches are sufficient for evaluating associations among variables and for tracking trends over time. However, an empirical study on propensity to complete the U.S. Census, comparing otherwise identical probability-based and non-probability surveys, indicated otherwise. It found "systematic and often sizable differences between probability sample telephone data and non-probability Internet data in terms of demographic representativeness of the samples, the proportion of respondents reporting various opinions and behaviors, the predictors of intent to complete the Census form and actual completion of the form, changes over time in responses, and relations between variables" (Pasek and Krosnick 2010). More study is warranted, but the picture to date is bleak.

Another recent trend is to evaluate information made publicly available on social networks such as Facebook and Twitter. The appeal of these datasets is their size and scope. Data can be collected on a minute-by-minute basis in vast quantities on nearly any topic imaginable. While these forms of data may hold great potential for social scientists, they also present unique challenges. For example, it may be assumed that a Twitter or Facebook post represents one individual expressing his or her actual opinion on something once. In fact some users may post multiple times, using a single account or multiple accounts. Postings may not reflect the self-initiated expression of actual attitudes, but rather may be part of orchestrated campaigns. Accounts may be created by interest groups, corporations, or paid public relations agents. Posts may be produced by automated computer programs known as "bots." Fake accounts can be purchased in bulk. All of these forms of information exist within the same datasets.

Regardless of their source, selecting relevant postings and extracting meaning from them are further challenges. Many postings include slang, irony, sarcasm, abbreviations, acronyms and emoticons, or lack identifiable context. Tests of automated coding systems indicate highly inconsistent results. And again we face the lack of theoretical justification to make inferences about a broader population.

What does the future hold for non-probability samples? Can they be "fixed"? Some researchers suggest the use of Bayesian adjustment, or a return to sample matching. While further research is welcome, what has been lacking to date is the required transparency that must underlie any such

evaluation. Non-probability methods should be held to the same analytical standards and evaluated on the same basis as probability samples with regard to claims of accuracy, validity, and reliability. Full disclosure of methods and data quality metrics is crucially important. And the Holy Grail remains the development of an online sampling frame with known probabilities of selection, bringing the enterprise into harmony with sampling theory.

Probability sampling requires ongoing evaluation as well. Some organizations implement poor-quality sampling designs and suboptimal execution and analysis. Coverage is an ongoing concern, and the potential impact of declining response rates needs continuing examination. So does work on probability-based alternatives to traditional telephone methods, such as address-based sampling, mixed-mode designs, and others that may be developed in the future.

Areas for future research:

- Expanded empirical research into the validity and reliability of non-probability survey data
- Efforts to develop a theoretical framework under which such samples may support inference
- Improved assessment and methods of analysis of social media data
- Continued examination of probability-based methods
- Development and implementation of transparency standards

References and Further Reading

Baker, R. (2009). Finally, the Real Issue? Retrieved 2009, from http://regbaker.typepad.com/regs_blog/2009/07/finally-the-real-issue.html

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., et al. (2010). Research Synthesis: AAPOR Report On Online Panels. *Public Opinion Quarterly*, *74*(4), 711–781. http://doi.org/10.1093/poq/nfq048

Callegaro, M., Villar, A., Krosnick, J. A., & Yeager, D. S. (2012). A Systematic Review of Studies Investigating the Quality of Data Obtained with Online Panels. Presented at the Annual Meeting of the American Association for Public Opinion Research, Orlando, FL.

Gallup, G. H., & Rae, S. F. (1940). The Pulse of Democracy: The Public Opinion Poll and How it Works. New York: Simon and Schuster.

Gilbert, J. P., Light, L. R., & Mosteller, F. (1977). Assessing Social Innovations: An Empirical Base for Policy. In W. B. Fairley (Ed.), *Statistics and Public Policy*. Reading, MA: Addison-Wesley Pub Co.

Kish, L. (1965). Survey Sampling. New York: John Wiley & Sons, Inc.

Pasek, J., & Krosnick, J. A. (2010). Measuring intent to participate and participation in the 2010 Census and their correlates and trends: Comparisons of RDD telephone and nonprobability sample. *Survey Methodology*.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples. *Public Opinion Quarterly*, *75*(4), 709–747. http://doi.org/10.1093/poq/nfr020

Gary Langer is a survey research practitioner. He is president of Langer Research Associates and former long-time director of polling at ABC Network News. Langer is a member of the Board of Directors of the Roper Center for Public Opinion Research, a trustee of the National Council of Public Polls, and former president of the New York Chapter of the American Association for Public Opinion Research. His work has been recognized with 2 News Emmy awards, 10 Emmy nominations, and AAPOR's Policy Impact Award. Langer has written and lectured widely on the measurement and meaning of public opinion.