AAPOR Guidance on Reporting Precision for Nonprobability Samples

In 2015 the American Association for Public Opinion Research (AAPOR) changed its position on reporting measures of precision from nonprobability samples. The updated version of AAPOR's Code of Professional Ethics and Practices allows for such reporting, provided that the measures are accompanied by "a detailed description of how the underlying model was specified, its assumptions validated and the measure(s) calculated." This document provides guidance and examples for complying with this new proviso.

While these guidelines were created for researchers using nonprobability samples, they may also be useful to those using probability-based samples subject to high nonresponse rates. In particular, the practice of stating and validating (if possible) the assumptions underpinning precision statements is equally valuable in these two contexts. Researchers using both types of samples generally rely on the assumption that their estimates are approximately unbiased – an assumption that is difficult for survey consumers to evaluate. Given that precision statements do not address other sources of error that can threaten approximate unbiasedness (e.g., noncoverage or nonresponse), we encourage consumers of all surveys to view precision statements within a broader context, such as the total survey error paradigm.[1] It is also important to note that this nonbinding document only addresses one subsection of the AAPOR Code (III.A.10); it is not comprehensive of all the disclosure elements required.

Reporting Examples for Several Common Approaches

There are a number of approaches that survey researchers use to estimate precision with nonprobability samples. AAPOR encourages researchers to implement the approach that is most appropriate for the study design. While not an exhaustive list, four commonly used approaches for quantifying the precision of statistical estimates include: resampling approaches, Bayesian credible intervals, Taylor series linearization, and application of the simple random sample (SRS) formula for margin of error. For each approach, there are certain pieces of information that a statistically-trained, independent observer would need to know in order to evaluate the study's design and the resulting estimates.

❖ Resampling Approaches

Resampling approaches, such as the bootstrap or jackknife, are preferred by many survey statisticians because they capture the variability associated with, for example, weighting adjustments. These methods approximate the variance of a survey estimator by the variability of that estimator computed from a series of subsamples taken from the survey data set.[2]

Important Details to Report

- Type of resampling approach (e.g., bootstrap, jackknife) and number of replicates used for variance estimation
- Assumed sample design (e.g., simple random sample, stratified random sampling with differential sampling rates)
- If there was clustering (e.g., multiple respondents per household), what if any adjustment was made (e.g., how were the strata and clusters defined)
- If the underlying nonprobability sample was selected using quotas, were replicates formed using the same quotas?
- Survey weights:
  - Whether or not weights were used in the calculation of the estimates
  - If weights were used, how were they computed and possibly adjusted (e.g., calibration)
  - Whether or not weights were calculated separately for each replicate
- Assumptions and any attempt to validate them

Reporting Example

To estimate the precision of estimates from this survey, we created a bootstrap confidence interval around a full sample weighted estimate of 50%. We generated 500 independent "resamples" by randomly selecting n-1

*respondents with replacement from the original survey data set. Sampling weights were adjusted for the resampling and were used to compute 500 estimated proportions. The variability in these 500 weighted estimates derived from each of the 500 "resamples" forms the basis of our standard error for the estimates reported. The bootstrap approach we applied assumed that the data came from a simple random sample. No further adjustments (e.g., calibration) were applied to the sampling weights. The confidence interval assumes that our weighted estimates are approximately unbiased. This assumption of approximate unbiasedness is based on our assertion that any differences between the survey sample and the target population on key survey outcomes are corrected by the sampling weight. [No analysis was conducted to validate that assertion. *OR* That assertion is based on a benchmarking analysis summarized in this report(link).][3]*

❖ Bayesian Credible Interval

A Bayesian credible interval is created as part of a statistical process that models the distribution of a population parameter using prior information (e.g., previous poll results) to produce an updated or posterior distribution of the estimate. While the Bayesian credible interval is in some ways analogous to the design-based confidence interval, it is computed and interpreted differently.[4]

Important Details to Report

- Prior model/distribution:
  - Source of prior(s)
  - Was the prior used non-informative (a.k.a. "flat") or informative?
- Posterior model:
  - Were sampling weights and/or design variables incorporated in the statistical models used to explain the relationship between the parameter of interest and the data collected?
  - What other predictors, if any, were included in this model?
- What type of model selection procedure was used in determining the final model?
- How was the fit of the statistical model assessed (i.e. R^2, AIC, BIC) and was the fit of the statistical model determined using cross-validation methods?
- Specific method that serves as the basis of the credible interval[5]
- Assumed sample design (e.g., simple random sample, stratified random sampling with differential sampling rates)
- If there was clustering (e.g., multiple respondents per household), what if any adjustment was made (e.g., how the strata and clusters were defined)
- Survey weights:
  - Whether or not weights were used in the calculation of the estimates
  - If weights were used, how were they computed and possibly adjusted (e.g., calibration)
- Other assumptions and any attempt(s) to validate them

Reporting Example

*The measure of precision reported for each survey estimate is its credible interval. The 95 percent credible interval for each weighted survey estimate was computed by calculating the 2.5 percent (lower bound) and 97.5 percent (upper bound) quantiles of the estimated posterior distributions that were computed using an informative prior. Parameters for the informative prior distribution were based on estimates derived from five national polls conducted between April 1-30, 2016. Each poll contained the presidential horserace question and estimates were reported on PollingAggregator.com. Since the posterior distributions are skewed, the credible intervals are asymmetric around the estimates. Details of the models, including covariates and procedures to calculate the survey weights are available in the appendix. The credible interval assumes that our weighted estimates are approximately unbiased. This assumption of approximate unbiasedness is based on our assertion that any differences between the survey sample and the target population on key survey outcomes are corrected by the weights. [No analysis was conducted to validate that assertion. *OR* That assertion is based on a benchmarking analysis summarized in this report(link).]*

❖ Taylor Series Linearization

Taylor series is the default variance estimation method in several statistical software packages. The software uses the linear Taylor series approximation of a nonlinear function and estimates the variance for that linear approximation.

Important Details to Report[6]

- Assumed sample design (e.g., simple random sample, stratified random sampling with differential sampling rates)
- If there was clustering (e.g., multiple respondents per household), what if any adjustment was made (e.g., how were the strata and clusters defined)
- Survey weights:
  - Whether or not weights were used in the calculation of the estimates
  - If weights were used, how were they computed and possibly adjusted (e.g., calibration)
- Assumptions and any attempt to validate them

Reporting Example

*All survey estimates were computed using weights. The weights were generated in a two-step process. First, a base weight was created with all cases assigned value 1. Second, this base weight was entered into a raking procedure that aligned the responding sample to American Community Survey benchmarks for gender, age, education, race, ethnicity, and region. Estimates were computed using these final weights, and standard errors for the estimates were computed using the Taylor series linearization method with these final weights. The standard errors assume that the weighted estimates used in the Taylor series linearization are approximately unbiased (or at the very least, consistent).[7] This assumption of approximate unbiasedness is based on our assertion that any differences between the survey sample and the target population on key survey outcomes are corrected by raking on the aforementioned demographics. [No analysis was conducted to validate that assertion. *OR* That assertion is based on a benchmarking analysis summarized in this report(link).]*

❖ SRS Formula with or without Adjustment for the Approximate Design Effect

Some poll reports show a margin of error (e.g., for p=50% at the 95% confidence level) computed using the classical formula that assumes a simple random sample (SRS). For a survey with 1,000 interviews, such a margin of error would look like 1.96 x SQRT[(.5 x .5)/1,000] x 100% = ±3.1. In reality, SRS conditions rarely hold due to nonresponse, weighting adjustments, and often additional factors. In many (if not most) surveys of people or households, the classical margin of error is incorrect (too small) due to departures from SRS.

A popular approach for addressing departures from SRS is to multiply the margin of error by an inflation factor. The inflation factor is computed as the square root of the approximate design effect, that is: SQRT[1 + (standard deviation of the weights / mean of the weights)^2]. This factor does not account for clustering, but there are additional adjustments that could be included for clustering. The AAPOR Code requires researchers to report whether or not their variance estimates have been adjusted for the design effect for both probability samples and nonprobability samples.

Important Details to Report

- Was the margin of error adjusted for the design effect? If so, how was the design effect computed?
- If there was clustering (e.g., multiple respondents per household), what if any adjustment was made?

Reporting Example

*The margin of error reported for this survey was computed using the classical SRS formula with an adjustment for the estimated design effect. The overall design effect for a survey is commonly approximated as the 1 plus the squared coefficient of variation of the weights. For this survey, the margin of error (half-width of the 95% confidence interval) incorporating the design effect for full-sample estimates at 50% is ± 4.5 percentage points. Estimates based on subgroups will have larger margins of error. The margin of error assumes that the weighted estimates are approximately unbiased. This assumption of approximate unbiasedness is based on our assertion that any differences between the survey sample and the target population on key survey outcomes are corrected by raking on the demographics listed in the weighting description. [No analysis was conducted to validate that assertion. *OR* That assertion is based on a benchmarking analysis summarized in this report(link).]*

❖ Declining to Estimate Precision

For some surveys (e.g., exploratory, internal research) estimating precision may not be important to the research goals. For other surveys precision measures may be relevant, but the researcher may not have the statistical resources to compute them. Under the AAPOR Code, it is acceptable for researchers working with nonprobability

samples to decline to report an estimate of variance. In such cases, it may useful to note that the survey estimators have variance, but there has been no attempt to quantify the size.

Where to Report This Information

Details about how variance was estimated can be provided in any number of locations and formats (in a study report, a Web page, a technical report, journal article, etc.). For publically released polls, if the information is not included in the study report itself, then a link or other instructions for accessing the information should be provided.

Why AAPOR Requires Disclosure of This Information

AAPOR's disclosure requirements are rooted in the idea that publically-released surveys are a scientific enterprise, and reproducibility is a core principle of science. A hypothetical independent researcher should be able to replicate the process by which the survey estimates – and precision statements about those estimates – were produced, even though the results themselves may not replicate due to factors such as nonresponse.

Acknowledgements

These guidelines were created by the AAPOR Standards Committee led by Courtney Kennedy and Trent Buskirk. A number of survey researchers and statisticians made very helpful contributions. We wish to thank Mike Brick, Jill Dever, Ariel Edwards-Levy, Natalie Jackson, Kyley McGeeney, Andrew Mercer, Anthony Salvanto, George Terhanian, Rick Valliant and Kirk Wolter for their comments.

Technical Notes

[1]For a recent review of the total survey error framework, see: Groves, Robert and Lyberg, Lars. 2010. "Total Survey Error: Past, Present, Future." *Public Opinion Quarterly*, Vol. 74 (5): 849-879.

[2]Resampling approaches may underestimate the true standard errors of estimates derived from nonprobability samples for several reasons. To explain, we contrast with the use of resampling approaches for probability samples. One potential problem is that the number of adjustments used to compute the replicate weights are likely to be fewer with nonprobability samples compared to probability samples that use resampling approaches. For example, a number of probability samples employ a base weight for differential probabilities of selection, a weighting-cell or propensity score nonresponse adjustment using information known for both respondents and nonrespondents, and a general calibration/raking step to reduce residual nonresponse and/or noncoverage bias. The impact of these adjustments is directly incorporated in the replicate subsamples to produce more accurate standard errors. However, in nonprobability samples, often little is known of the "nonrespondents" or those population members who never joined the sample in the first place, so it is unlikely that such nonresponse adjustments are incorporated in forming the replicate weights beyond general calibration/raking. Second, resampling approaches presume that the sample represents the target population in some systematic way as governed by a sampling design and that the replicate subsamples are being selected from the sample in the same way as the sample was selected from the population. In the case of nonprobability samples, there is not necessarily a stable design used to generate the sample or to govern how replicate subsamples are selected from it. Nonprobability samples may produce samples that have less variability on key outcomes than would be observed in the larger target population, especially if the self-selection mechanism governing participation in nonprobability samples is associated with the survey outcomes. In such cases, resampling approaches would produce statistical estimates that are erroneously similar from one replicate sample to the other and, thus, result in biased (too low) variance estimates.

[3]Put simply, either there was or was not an attempt to validate the assumptions underpinning the precision statement for a given survey. At present such validation efforts are quite rare for both probability and nonprobability sample surveys. In the spirit of the AAPOR Code (2015), this document is an effort to simply establish transparency on the issue. One potential approach for validating the assumption(s) would be with a benchmarking study to evaluate bias. A benchmarking study is not, however, the only potential approach, nor would it theoretically need to be conducted for each survey in a series of similarly designed surveys. The examples of "no validation analysis" and "benchmarking study" are presented simply because we anticipate these would be fairly common scenarios.

[4]Historically, a 95% confidence interval for the population mean hourly wage for workers between the ages of 16 and 20 might be estimated as ($7.50, $22.50). The interpretation of this interval rests on the repeated sampling framework – if the sample was redrawn 100 times from the population of interest and a similar confidence interval was computed estimating the mean hourly wage for workers between the ages of 16 and 20, we would expect that roughly 95 of these intervals would contain or capture the true value of the mean hourly wage for these workers in our population (i.e. the parameter). The half-width of this confidence interval is also called the "margin of error" for the point estimate for the population mean hourly wage. Notice that this interpretation does not mention the specific endpoints of the computed interval, but rather emphasizes the process by which the interval was computed. In this way, confidence intervals treat the parameter of interest as fixed and the randomness comes from the sampling design. In contrast, if the interval ($7.50, $22.50) were computed as a 95% Bayesian credible interval then we could say that there is a 95% chance that the true average wage for 16 to 20 year olds in the population is between $7.50 and $22.50 under the assumed model. Note that in this interpretation, the end points of the Bayesian credible interval are considered fixed and the population parameter (average hourly wage) is considered random. As such, the Bayesian credible intervals are byproducts of a process that attempts to model the distribution of such possible parameters by using prior information on such parameters that is then updated based on information contained in the sample via a statistical model. More information about Bayesian credible intervals in survey research can be found in AAPOR's statement on credibility intervals which is available here:
http://www.aapor.org/AAPOR_Main/media/MainSiteFiles/DetailedAAPORstatementoncredibilityintervals.pdf

[5]Credible interval estimates for a population parameter derived using a posterior distribution may not be unique and can be determined using various approaches including: (a) choosing the narrowest interval (e.g. highest posterior density interval), (b) choosing the interval where the probability of being below the interval is as likely as being above it (e.g. equal-tailed interval – which will contain the median), or (c) choosing an interval for which the mean is the central point, provided such a value exists. While generally, the most common method uses the narrowest interval, we note that this is not the only way to derive a credible interval from a posterior distribution. This aspect of credible intervals should be incorporated in the reporting so that proper interpretation can be made and comparisons to other similarly computed credible intervals also made.

[6]For many multistage sampling designs, standard errors for various statistics of interest are often derived using a "with replacement" approximation to the sampling design that typically provides conservative (i.e., larger) estimates of the standard errors. This approximation essentially relies on first stage strata (typically) and primary sampling units (PSUs) along with the sampling weights to derive standard errors of weighted survey estimates. The Taylor series approximations (as well as repeated sampling approaches) operationalized in many popular software packages will reflect variability in the estimator aggregated up to the PSU level. The Taylor series approximation for computing standard errors of weighted statistics can also be applied for without replacement designs provided that joint inclusion probabilities of selecting pairs of sampling units are provided. This is an important note when comparing standard error estimates across software packages and even within (i.e. with or without using a with replacement approximation to a multi-stage design).

[7]The unbiasedness or, at a minimum, the consistency property of the sample estimators used in computing the standard errors under the Taylor series linearization method is determined by the inferential framework implied by the underlying sampling design. Generally, nonprobability based samples have no such sampling design from which to derive a "design-based" inferential framework. In these instances, a model-based inferential framework might be used to establish the consistency (or unbiasedness) of estimators used in the computation of the standard errors under the linearization approach.