Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Measuring Uncertainty with Multiple Sources of Data

Sharon L. Lohr[1]

Abstract

In a probability sample with full response, the margin of error provides a reliable, theoretically justified measure of uncertainty. When combining estimates from multiple samples or administrative data sources, however, traditional margins of error underestimate the uncertainty—differences among statistics from various sources often exceed the estimated sampling variability. I examine methods that have been proposed for measuring uncertainty from combined data sources, with application to estimating smoking prevalence and sexual assault rates, and outline some possible directions for research.

Key Words: Margin of Error; Calibration; Complex Surveys; Bayesian Model Averaging; Sexual Assault; Smoking Prevalence.

1. Introduction

1.1 Combining data to meet statistical needs

The challenges facing official statistics are well known. Survey response rates are decreasing. At the same time, there is increased demand for more granular information, available faster and with less expense. The theme of this conference was how to meet these increased demands by combining data or statistics from multiple sources: surveys, administrative data, sensor data, or data gleaned from the internet.

The hope is that the statistics from the combined data more accurately (or less expensively) estimate the quantities of interest. One major challenge, however, is assessing the accuracy of statistics from combined data. Most survey estimates are accompanied by confidence intervals that indicate the magnitude of the uncertainty about the estimates resulting from sampling variability. Statistics from administrative and sensor data typically do not have sampling variability and are usually presented without measures of uncertainty. But they do have other types of errors, and we can learn about those errors by studying multiple data sources.

In this paper, I discuss some of the issues for measuring uncertainty for three methods commonly used to combine data—multiple-frame methods, hierarchical models, and calibration—from the perspective of three examples. Two of the examples are presented in this section, and the third is introduced in Section 5.

1.2 Smoking among U.S. adults

Figure 1.2-1 shows point estimates and confidence intervals for the percentage of U.S. adults who have smoked at least 100 cigarettes in their lifetime. Siegfried et al. (2017) computed the statistics that are displayed in Figure 1.2-1 from five nationally representative sample surveys: the Tobacco Use Supplement of the Current Population Survey (TUS-CPS), the Population Assessment of Tobacco and Health (PATH) Study, the National Health Interview Survey (NHIS), the National Survey of Drug Use and Health (NSDUH), and the National Adult Tobacco Survey (NATS).

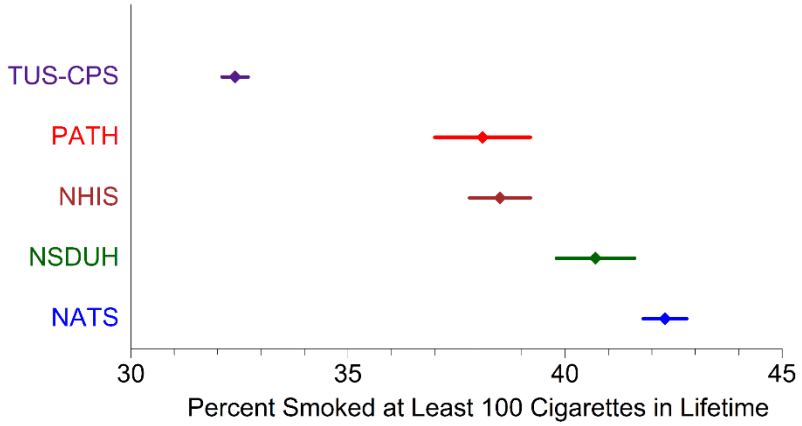
All of the surveys were conducted at approximately the same time (2013-2015) and had target populations that included civilian noninstitutionalized U.S. residents age 18 and over, so that an estimate for U.S. adults age 18 and over could be computed from each. Some of the surveys asked the question about smoking at least 100 cigarettes

[1]Sharon L. Lohr, Professor Emerita, Arizona State University, Tempe, AZ, 85287-1804, USA, www.sharonlohr.com

during one's lifetime directly; for the others, the estimate could be computed from several questions. But all of the point estimates in Figure 1.2-1 are, in theory, estimating approximately the same population quantity.

The point estimates differ more than one would expect from the narrow confidence intervals, which reflect only the sampling error. The estimate from TUS-CPS, in particular, is much lower than the others, and the estimate from NATS is substantially higher.

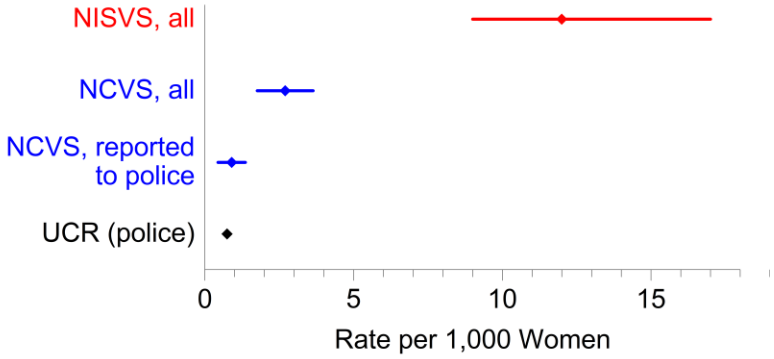
Figure 1.2-1
Percentage of U.S. adults who have smoked at least 100 cigarettes, with 95 percent confidence intervals



1.3 Sexual assault in 2015

Figure 1.3-1 displays four estimates of rape and sexual assault in the U.S. in 2015, along with 95 percent confidence intervals. Lohr (2019) discusses the data sources for these estimates.

Figure 1.3-1
U.S. 2015 female rape/sexual assault rates from three sources, with 95 percent confidence intervals



The statistics in the top line are from the 2015 National Intimate Partner and Sexual Violence Survey (NISVS), found in Table 1 of Smith et al. (2017). The statistics in the second and third lines are from the 2015 National Crime Victimization Survey (NCVS; Bureau of Justice Statistics, 2018a, 2018b). The second line includes all rapes and sexual assaults that women reported in the 2015 NCVS; the third line includes only the victimizations that NCVS respondents said were known to the police.

The two surveys have a number of differences. The NISVS statistic estimates the number of U.S. adult women, per 1,000 women, who experienced at least one completed or attempted rape in the previous 12 months. The NCVS statistic estimates the number of rape (completed and attempted) and sexual assault victimizations per 1,000 women in the previous 12 months. The ages considered also differ: the NISVS interviews women age 18 and older; the NCVS also interviews persons age 12 to 17 (who typically have one of the highest rates for sexual assault victimization). Thus, the NCVS includes more types of crime (other types of sexual assault as well as rape), includes multiple victimizations per woman, and includes an additional age group that typically experiences high rates of sexual assault. One would expect from these considerations that the NCVS statistic would be larger than the NISVS statistic; instead, the NCVS estimate is much smaller. Much of the difference can be explained by the different questions used in the two surveys: although the definitions of rape are similar for the NCVS and NISVS, the questions are actually measuring different population quantities. But there are other explanations for the differences as well (Lohr, 2019).

The NCVS estimate of rapes and sexual assaults reported to the police is closer to the fourth estimate in Figure 1.3-1, from the Federal Bureau of Investigation's Uniform Crime Reporting System (UCR). The UCR tallies crimes known to law enforcement agencies. Table 1 from FBI (2016) reported 126,134 rapes in 2015 but did not give breakdowns by sex; to obtain the estimated victimization rate for females, I apportioned these to men and women according to the victim proportions in the 2015 National Incident Based Reporting System data (Puzzanchera et al., 2017).

The UCR data are intended to be a census of crimes known to law enforcement agencies. The FBI does not report a measure of uncertainty for the estimates, or provide a measure of the variability from the imputation procedure used for missing data. The other types of errors that might go into a mean squared error, such as variability in crime classification and recording among law enforcement agencies, have not been well studied. Thus, the point estimate for the UCR statistic in Figure 1.3-1 is unaccompanied by a confidence interval, but that does not mean the estimate has no error; we just do not have a good measure of the uncertainty.

In Figures 1.2-1 and 1.3-1, the variability between estimates from the different sources exceeds the within-source variability given by the confidence intervals. For the statistics from NCVS and NISVS, this occurs in part because the questions asked for the NISVS elicit more reports of sexual assault than the questions asked for the NCVS; for the smoking statistics, the differences may arise from nonsampling errors and methods for taking the surveys (Siegfried et al., 2017).

2. Methods for combining data

Lohr and Raghunathan (2017) and National Academies of Science, Engineering, and Medicine (2017) describe statistical methods that may be used to combine data. These include (1) record linkage, (2) small area estimation, (3) imputation, (4) multiple-frame methods, (5) hierarchical models, and (6) calibration. In this paper, I briefly describe some issues involved in measuring uncertainty for the last three of these methods.

All of the methods have formal procedures for measuring uncertainty, but the measures can be unrealistically small. Uncertainty for statistics from combined data arises from:

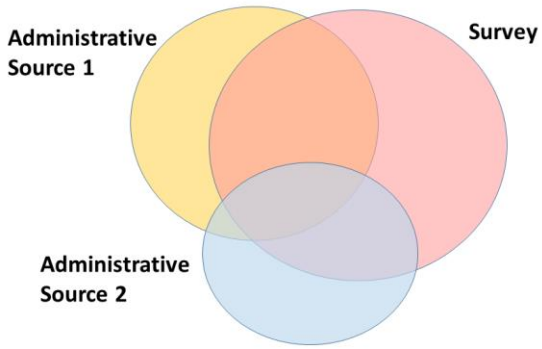
A. Sampling error from the individual sources
B. Nonsampling error from the individual sources
C. Differences across sources
D. Statistical models used to combine the data

Most published measures of uncertainty include sampling variability, but, as we saw in the examples from Section 1, the measures do not include nonsampling errors and other differences that can cause statistics from different sources to vary.

3. Multiple-frame methods

Figure 3-1 shows a multiple-frame structure with three potential sources of data. Each source covers only part of the population. Multiple-frame surveys are often used with independent sample surveys, but in Figure 3-1 two of the sources are administrative data such as the UCR or electronic health records, which may be censuses.

Figure 3-1
Three sources (frames) of data



The estimator of a population total for Figure 3-1 is the sum of the seven estimated domain totals, where three domains are found in a single data source, three are in exactly two sources, and the last is in all three sources.

What about uncertainty? The variance of the estimated population total for a multiple-frame survey is a function of the covariance matrices of each source's vector of estimated domain totals (Lohr, 2011). But this variance captures only uncertainty from sampling error, which does not affect the administrative data sources. The mean squared error of the estimated total may be much larger than the variance if some of the domain estimates are biased for the corresponding domain population totals.

The differences among estimates for the surveys in Figures 1.2-1 and 1.3-1 give reason to doubt the multiple-frame assumption that each source provides unbiased estimates of the same population domain quantities. However, the NCVS estimate of rapes reported to the police is consistent with the FBI (2016) statistic, and there is potential for using multiple-frame-type methods with those two data sources. To assess the uncertainty of combined estimates, it is necessary to learn more about the errors affecting both sources of data.

When all or some sources have incomplete coverage of the population, multiple-frame methods can give estimates that apply to the population defined by the union of all frames. But one must know how many frames each data point, from each source, belongs to. Is the incident reported by an NCVS respondent also in the FBI records (and thus in the overlap of the two frames), or is it in the part of the population covered only by the NCVS? Conversely, is a rape reported to law enforcement also in the scope of the NCVS (which it will not be if it occurred to a child, resident of a foreign country, or person in an institution)? If the domain classifications are inaccurate, estimates of population totals may be biased and estimates of uncertainty too small. Lohr (2011) and Lin (2013) presented methods for adjusting estimates and their variances for possible domain misclassification.

Aspects of the multiple-frame paradigm also apply to many of the other methods of combining data. A fundamental problem for many methods is ascertaining the coverage and overlap of different sources; for hierarchical models, small area estimation, imputation, and calibration, it is often implicitly assumed that the sources cover the same population. Multiple-frame methods can help assess that assumption and be used together with other statistical models when sources have incomplete coverage.

4. Hierarchical models

Hierarchical models, often used in biostatistics for meta-analysis, formally capture heterogeneity among sources in the estimates of uncertainty. To illustrate, consider the model of Manzi et al. (2011) for the mean $\bar{y}_{dj}$ from domain d and data source j:

$$\bar{y}_{dj} = \theta_d + \delta_{dj} + e_{dj}, \tag{4.1}$$

where θ_d is the overall mean in domain d; δ_{dj} is a random effect for source j's deviation from the overall domain mean, assumed to follow a $N(\Delta_j, \tau_j^2)$ distribution; and e_{dj} is the sampling error for $\bar{y}_{dj}$. In equation (4.1), the domain means from source j are assumed to have average bias Δ_j. This and related models allow one to explicitly model the bias from each source.

For the model parameters in equation (4.1) to be identifiable, one needs to provide additional information about the parameters θ_d, defining a source or combination of sources as unbiased for a function of the parameters. If combining the estimates on smoking in Figure 1-2.1, which source, if any, is unbiased? Hierarchical models have strong assumptions on bias, model form, and population coverage (the issues affecting multiple-frame surveys also affect hierarchical models). But these assumptions are made explicit and some of them can be tested empirically.

One tremendous advantage of hierarchical models is that they capture heterogeneity among data sources in the posterior distribution of the parameters. The posterior variance may be larger than each individual source's sampling variance.

5. Calibration

5.1 Variance of calibrated estimators

Finally, let's look at calibration, probably the most commonly used method for combining survey data with information from another source. The variable of interest in the survey is denoted by y, and the vector of auxiliary information, measured in both the survey and an external source of control totals, is denoted by $\boldsymbol{x}$. Calibration adjusts the survey weights so that $\widehat{\boldsymbol{X}}$ (the estimated population total of $\boldsymbol{x}$ from the survey) equals $\boldsymbol{X}$ (the population total of $\boldsymbol{x}$ from the external source).

For the special case of poststratification, the variance of the poststratified estimator $\hat{Y}_{ps} = \boldsymbol{X}'\widehat{\bar{\boldsymbol{Y}}}$, where $\boldsymbol{X}$ is the vector of control totals for the G poststrata and $\widehat{\bar{\boldsymbol{Y}}} = (\hat{Y}_1/\hat{X}_1, \ldots, \hat{Y}_G/\hat{X}_G)'$ is the vector of poststrata means estimated from the survey. The variance of $\hat{Y}_{ps}$ is

$$V(\hat{Y}_{ps}) \approx \boldsymbol{X}'V\left(\widehat{\bar{\boldsymbol{Y}}}\right)\boldsymbol{X}. \tag{5.1}$$

Poststratification almost always reduces the variance of the estimator of the population total. But whether it reduces the mean squared error when there is nonresponse depends on the nonresponse mechanism and on the properties of the control totals $\boldsymbol{X}$.

Dever and Valliant (2010, 2016) showed that if the control totals come from an auxiliary survey (for example, the American Community Survey is often used as a source of control totals) instead of a census, and thus have sampling variability, the variance in equation (5.1) can substantially underestimate the variance of the poststratified estimator. In that case, the calibration estimator with estimated control totals is $\hat{Y}_{EC} = \widehat{\boldsymbol{X}}_{aux}'\widehat{\bar{\boldsymbol{Y}}}$, where $\widehat{\boldsymbol{X}}_{aux}$ is an estimator of the vector of control totals from the auxiliary survey, and it has variance

$$V(\hat{Y}_{EC}) \approx \boldsymbol{X}'V\left(\widehat{\bar{\boldsymbol{Y}}}\right)\boldsymbol{X} + \bar{\boldsymbol{Y}}'V(\widehat{\boldsymbol{X}}_{aux})\bar{\boldsymbol{Y}}. \tag{5.2}$$

Here, $\bar{\boldsymbol{Y}}$ is the population vector of poststratum means. The second term in equation (5.2) can have the same order of magnitude as the first term (and even be larger if the auxiliary survey has high variability).

Note what Dever and Valliant (2010) did in equation (5.2). Control totals from an auxiliary survey do not equal the population values because of sampling variability. They may be too high for some poststrata and too low for others. This creates a bias for $\hat{Y}_{EC}$, but we do not know the direction or size of the bias, and the variance in equation (5.1), which is conditional on the control totals, does not capture it. Equation (5.2) transforms the uncertainty about the bias into a variance so that it is reflected in the confidence interval for the estimator.

The variances in equations (5.1) and (5.2) assume that the variables $\boldsymbol{x}$ in the administrative data or auxiliary survey are the same as the variables $\boldsymbol{x}$ in the main survey. This is not necessarily true. If race and ethnicity categories are used for poststratification, for example, the main survey may define or measure the categories differently than the administrative data or auxiliary survey—the definitions may differ, one source might use multiple-race categories while the other does not, or the questions or context may elicit different responses.

If the main survey has full response and the $\boldsymbol{x}$ variables are consistent across sources, then the variance in equation (5.2) usually coincides with the mean squared error of the estimator, for any choice of auxiliary variables and calibration model.

When there is nonresponse, however, different choices for the calibration model may give different answers. The variance of the estimate in equation (5.2) is conditional on the model choice, and does not include possible bias resulting from that choice.

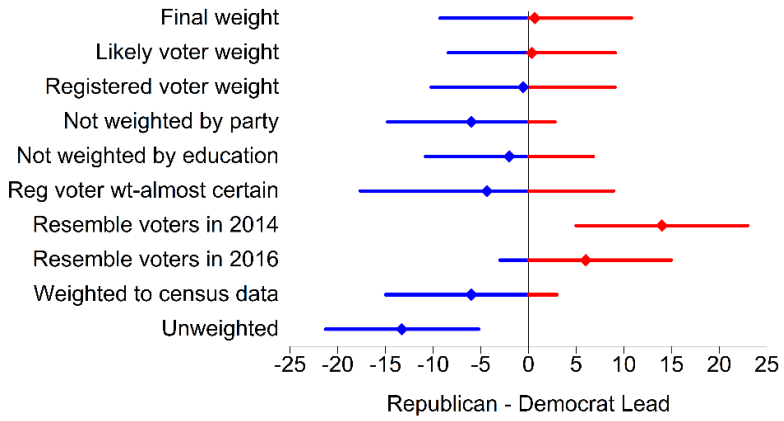
5.2 U.S. 2018 midterm election polls

To study the difference calibration models can make, let's look at polling data from *The New York Times* (Andre et al., 2018) for the 2018 midterm election in Illinois' Sixth Congressional District. The telephone poll was conducted from September 4 to 6, 2018; Cohn (2018) describes the methodology. For this district, 36,455 calls were made to likely voters, and 512 persons responded, resulting in a response rate of 1.4 percent.

Republican Peter Roskam was projected to receive 45 percent of the vote and Democrat Sean Casten was projected to receive 44 percent; each of these statistics had margin of error 4.7 percentage points (the remaining 11 percent in the poll were undecided). Thus, as of September 2018, the poll showed a one-point estimated lead for the Republican candidate, well within the margin of error of nine percentage points.

But the response rate was low and the poll was conducted two months before the election. Strong assumptions underlie what variables are used in the weighting, who is projected to vote, and what undecided persons are expected to do.

Figure 5.2-1
Estimates for Illinois Congressional District 6 under different weighting and turnout models



Andre et al. (2018) did something few other pollsters do: they showed what the results would be under different weighting models and voter turnout assumptions, and provided alternative weights on the data set. Figure 5.2-1 displays 95 percent confidence intervals for the percentage-point Republican lead under different weighting and voter turnout models. The red part of each line indicates the Republican was projected to lead; the blue part indicates the Democrat was projected to lead.

The point estimates depend heavily on which model is used for weighting, and the variability among the estimates from different weighting models rivals that from the sampling error.

6. Including uncertainty from weighting models

6.1 Bayesian model averaging

The differences among the estimates in Figure 5.2-1 suggest that the confidence intervals based on sampling error alone underestimate the uncertainty of the estimator. Bayesian model averaging can be used to include uncertainty about the model. Hoeting et al. (1999) give an overview; Lohr and Brick (2017) apply the method to the Literary Digest Poll of 1936.

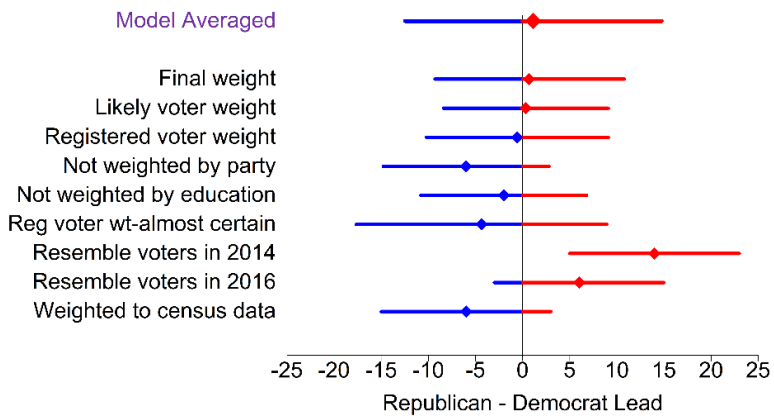
Consider an application with weighting models $M_1, M_2, \ldots, M_K$ and data D. The posterior distribution for model M_k given the data is $pr(M_k|D)$. Then the posterior distribution for a parameter θ given the data is

$$pr(\theta\,|D) = \sum_{k=1}^{K} pr(\theta\,|M_k,\ D)\ pr(M_k|\ D). \tag{6.1}$$

Thus, the posterior mean for θ is a weighted average of the estimates, weighted by the posterior probabilities of the models. The posterior variance includes the sampling variability of each estimate as well as the variability among the estimates from the different weighting schemes.

Figure 6-1.1 shows a Bayesian model-averaged posterior prediction interval for the polling data. I did not have enough information from the data to obtain data-driven posterior distributions for the models (after all, these are models for assumptions about nonresponse mechanisms and I do not have the data from the nonrespondents), so I adopted a prior distribution for the models based on the subjective descriptions in Andre et al. (2018). I gave the unweighted model a prior probability of zero, since it seems that some type of weighting would be necessary.

Figure 6.1-1
Estimates for Illinois Congressional District 6, including Bayesian model-averaged estimate



The model-averaged interval in Figure 6-1.1 is wider than any of the intervals for the individual weighting schemes, reflecting my prior uncertainty about the weighting models. The point estimate is close to that of the model with the final weights, but the interval estimate is wider, reflecting the additional uncertainty about the weighting. Like the Dever and Valliant (2010) method, Bayesian model averaging transforms the uncertainty about the weighting model from a potential bias to a variance.

I did this analysis two months before the election, in September, 2018. What happened in the November election? According to the Illinois State Board of Elections (2018), 146,445 persons voted for Roskam (Republican), and 169,001 persons voted for Casten. The (Republican – Democrat) difference was approximately (– 7) percentage points.

6.2 Challenges

Of course the prior distributions are key to the estimates, and these depend on assumptions. In other applications of Bayesian model averaging, for example when one is considering different regression models, the data have information about which models fit well. But models for nonresponse are predicting data that are not observed. I used subjective prior probabilities for the weighting models, but posterior probabilities could be calculated for the models if information is available from the sampling frame or an external source. Another option would be to include information from past data.

The theory of design-based inference was developed in order to avoid problems of subjectivity with survey samples. Design-based inference avoids model assumptions, needs no subjective priors, and has elegant and beautiful mathematical theory. Neyman (1934, p. 592) argued that the "representative method" (probability sampling) is superior because the "construction of the confidence belt is quite independent of any arbitrary assumption concerning the values of θ." The representative method "makes it superfluous to make any appeals to the Bayes' theorem" (Neyman, 1934, p. 562).

When there is nonresponse, however, you *must* make model assumptions. As soon as you adopt a nonresponse model, you have become a Bayesian. But you are a Bayesian who has complete and utter belief in the particular nonresponse model you chose for the weighting; you have placed a certainty prior on one nonresponse model.

Many steps need to be taken, and many issues need to be resolved, before these methods are ready for use with official statistics. It is easy to obtain lower variability by considering only weighting models that give similar answers, and standards are needed (just as standards have been developed for reporting response rates). One option would be to register weighting models and prior distributions before collecting the data, similarly to the registration of clinical trial protocols and proposed analyses.

A Bayesian model averaging approach has the advantage that, like hierarchical models, the assumptions about the nonresponse models are explicit. We always have to make assumptions about missing data, but for most surveys these assumptions are hidden in the technical notes, or the details of the weighting decisions may remain unpublished. With a Bayesian model, the assumptions are set out for all to see and they can be evaluated.

7. Discussion

7.1 Using multiple data sources to explore error properties

There are a number of reasons why the survey estimates in Figures 1.2-1 and 1.3-1 differ. For the NISVS and NCVS, the survey questions are likely responsible for most of the difference in the estimates. The NCVS asks explicitly whether the respondent was raped or sexually assaulted; the NISVS asks about behavior and events that occurred. Lohr (2019) discusses other potential reasons for the difference between NISVS and NCVS, and Siegfried et al. (2017) discuss reasons for the differences among the smoking estimates.

Different weighting models for nonresponse may also explain some of the differences for the survey estimates. The NCVS and NISVS both use weighting but the final weighting models are not described in detail. In addition, the

NISVS response rate is about 40 to 50 percentage points lower than the NCVS response rate; the NISVS estimates may be more sensitive to weighting adjustments than the NCVS estimates because the NISVS has more nonresponse.

One way to evaluate the surveys may be possible when the FBI finishes its transition to the National Incident-Based Reporting System, which collects details about characteristics of victims and offenders for each offense. Right now only about 40 percent of agencies participate in the system, but the detailed information could be used to study potential nonresponse bias in the NCVS (or potential underreporting errors in the law enforcement statistics).

Similarly, for the different lifetime smoking estimates, we can use the some sources to explore error properties in the other sources. For example, the CPS allows proxy interviews; does that help explain why its estimates are lower than the others? Willis et al. (2017, p. 3) state that the lower estimates from TUS-CPS are "likely for a variety of reasons involving differences between surveys in technical and survey administration features."

7.2 The zeroth problem

Colin Mallows (1998) wrote about how statisticians often become involved too late in a research problem. Statisticians need to be involved in what he called the zeroth problem, the problem of determining what data sources are relevant to the problem. He argued: "Statistical arguments often fail to convince because the basis for their assumptions is not spelled out."

When using multiple sources for inference, the zeroth problem is evaluating the quality and properties of the individual data sources, and assessing which are relevant for answering the research questions.

Only after that evaluation can we move on to combining data. All of the methods for combining rely on models, and the measures of uncertainty for the combined data sources depend on the measures for the individual data sources, which may be underestimates. Those underestimated individual variances are inherited by the combined estimate.

Having a diversity of sources, particularly if the sources have different types of nonsampling errors, can help us understand that underestimation and explore the error properties of each individual source. They can help us study quality in a systems-level approach. Measuring uncertainty in individual studies, and in combined data, is a systems-level problem, and thus it needs a systems-level solution: one that includes measurement and nonresponse errors as well as variability that results from different weighting models.

References

Andre, M., Buchanan, L., Bloch, M., Bowers, J., Cohn, N., Coote, A., Daniel, A., Harris, R., Katz, J., Rebecca Lieberman, R., Migliozzi, B., Murray, P., Pearce, A., Quealy, K., Weingart, E., and White, I. (2018), "We Polled Voters in Illinois's 6th Congressional District", *The New York Times*, https://www.nytimes.com/interactive/2018/upshot/elections-poll-il06-1.html, accessed September 10, 2018.

Bureau of Justice Statistics (2018a), "Rates of Rape/Sexual Assaults by Sex and Reporting to the Police, 1993-2016", Generated using the NCVS Victimization Analysis Tool at www.bjs.gov on September 11, 2018.

Bureau of Justice Statistics (2018b), "Standard Errors for Rates of Rape/Sexual Assaults by Sex and Reporting to the Police, 1993-2016", Generated using the NCVS Victimization Analysis Tool at www.bjs.gov on September 11, 2018.

Cohn, N. (2018), "Our Polling Methodology", *The New York Times* (September 6), https://www.nytimes.com/2018/09/06/upshot/live-poll-method.html, accessed September 10, 2018.

Dever, J., and Valliant, R. (2010), "A Comparison of Variance Estimators for Poststratification to Estimated Control Totals", *Survey Methodology*, 36, pp. 45-56.

Dever, J., and Valliant, R. (2016), "General Regression Estimation Adjusted for Undercoverage and Estimated Control Totals", *Journal of Survey Statistics and Methodology*, 4, pp. 289-318.

Federal Bureau of Investigation (2016), *Crime in the United States, 2015*. Washington, D.C.: Federal Bureau of Investigation

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial", *Statistical Science*, 14, pp. 382-401.

Illinois State Board of Elections (2018), *Official Canvass, General Election, November 6, 2018*, Springfield, IL: Illinois State Board of Elections.

Lin, D. (2013), *Measurement Error in Dual Frame Estimation*, Ph.D. Dissertation, Southern Methodist University.

Lohr, S. L. (2011), "Alternative Survey Sample Designs: Sampling with Multiple Overlapping Frames", *Survey Methodology*, 37, pp. 197-213.

Lohr, S. L. (2019), *Measuring Crime: Behind the Statistics, Boca Raton*, FL: CRC Press.

Lohr, S. L., and Brick, J. M. (2017), "Roosevelt Predicted to Win: Revisiting the 1936 Literary Digest Poll", *Statistics, Politics and Policy*, 8, pp. 65-84.

Lohr, S. L., and Raghunathan, T. E. (2017), "Combining Survey Data with Other Data Sources", Statistical Science, 32, pp. 293-312.

Mallows, C. (1998), "The Zeroth Problem", *The American Statistician*, 52, pp. 1-9.

Manzi, G., Spiegelhalter, D. J., Turner, R. M., Flowers, J., and Thompson, S. G. (2011), "Modelling Bias in Combining Small Area Prevalence Estimates from Multiple Surveys", *Journal of the Royal Statistical Society: Series A*, 174, pp. 31-50.

National Academies of Sciences, Engineering, and Medicine (2017). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*, Washington, D.C.: The National Academies Press.

Neyman, J. (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection", *Journal of the Royal Statistical Society*, 97, pp. 558-625.

Puzzanchera, C., Smith, J., and Kang, W. (2017), "Easy Access to NIBRS Victims, 2015: Victims of Violence", https://www.ojjdp.gov/ojstatbb/ezanibrsdv/, accessed September 11, 2018.

Siegfried, Y., Morganstein, D., Piesse, A., and Lohr, S. (2017), "Why Independent Surveys with the Same Objective Yield Different Estimates", *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 935-952.

Smith, S. G., Zhang, X., Basile, K. C., Merrick, M. T., Wang, J., Kresnow, M.-J., and Chen, J. (2018). *The National Intimate Partner and Sexual Violence Survey (NISVS): 2015 Data Brief.* Atlanta, GA: National Center for Injury Prevention and Control, Centers for Disease Control and Prevention.

Willis, G., Hartman, A., Reyes-Guzman, C., Seaman, E .L., Gibson, J. T., Goettsche, E., Chomenko, D., Mangold, K., and Block, M. (2017), *The 2014-2015 Tobacco Use Supplement to the Current Population Survey*, Rockville, MD: National Cancer Institute.