Survey Methods

Chandler Zachary
MATH 6330
CU Denver

Submitted May 14, 2020

1 Introduction

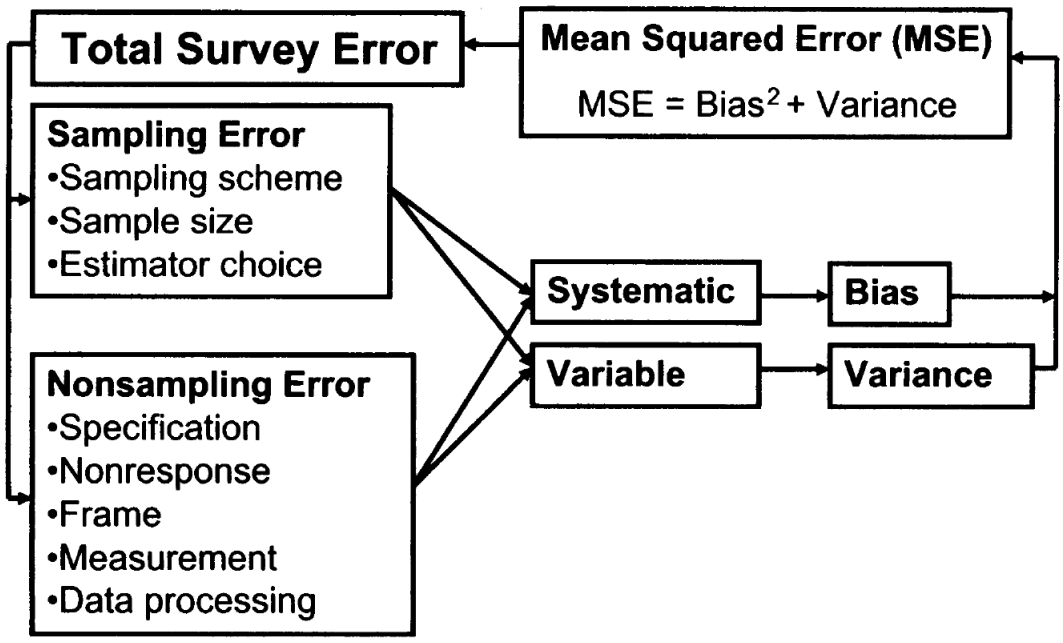
Surveys are constructed to illuminate some interesting fact about a whole population that can only be obtained by direct, verbal engagement. Policy analysts may want to know some detail about healthcare consumption, pollsters about voting preferences, or senior managers about workplace behaviors, to take a meager few examples. Thus, they may ask targeted questions of people who have health insurance, registered voters, and employees to make inferences based on the responses about the overall populations represented. In all cases, the only way to obtain the information is by talking to people or asking them to submit written information. Insofar as surveys enable statisticians to make such inferences, survey design rests on the principles of sampling design. It is not enough to talk with people. One must talk with the right people to represent the population of interest and minimize sources of selection bias, as well as with a sufficient number of people to find an effect.

Statisticians need to verify that the effects they find with a survey are not polluted by errors related to measurement. This problem, however, does not lie solely in the sampling design. That is, it does not arise solely from surveying the wrong people for the population of interest. It can also result from asking the wrong questions of the right people or by asking the right questions in confusing ways. This suggests that the *total error* to be minimized in conducting a survey is composed of two sources: *sampling error* and *non-sampling error*. Sampling error is the result of surveying only a subset of people rather than the entire population, and it can be caused by mis-specifying the sample or by failing to understand the probability mechanism underlying the estimation procedure. Non-sampling error is everything else related to measuring and recording the covariates of interest.

Minimizing total error in the survey process is critical to prevent jeopardizing the quality of inference. Figure 1 is a helpful aid to relate total survey error to statistical principles. I will elaborate further on the details in this graphic, but for now, it suffices to say that the components of survey error, sampling and non-sampling, can be matched to systemic and variable errors that translate into bias and variance, the components of mean squared

error. Thus, minimizing survey error can be thought of as minimizing mean squared error, a common objective for good experimental design. Total survey error is a framework for

Figure 1



Source: Total Survey Error: Design, Implementation, and Evaluation. Biemer, Paul P. (2010).

enforcing survey quality from the design through the implementation and into the analysis. It provides a tool for matching sources of error in the survey process to sources of error in inference. The framework is developed at length in the text *Introduction to Survey Quality* (Biemer and Lyberg, 2003), and I will follow much of its exposition in the following work. In the next section, I will expand on the sources of non-sampling error and use examples from the National Research Council's (NRC's) report "Estimating the Incidence of Rape and Sexual Assault," conducted as an analysis of the National Crime Victimization Survey (NCVS) by the Bureau of Justice Statistics (BJS) (National Research Council, 2014). The NCVS[1] surveys a representative sample of 160,000 people annually why were victims of non-fatal personal crimes or property crimes. The NRC report gives detailed explanations for their claims about the various sources of error in the NCVS in the context of the total

[1]The survey homepage is: https://www.bjs.gov/index.cfm?ty=dcdetail&iid=245.

survey error framework. The NRC report examines the legal and methodological difficulties in counting incidence of sexual crimes and was used to update the NCVS. In the following section, I will mention a few useful sampling strategies for surveys. Before continuing, I must point out that this paper makes some assumptions about the reader's background in statistical theory. An upper division course in statistics should be sufficient.

2 Survey Quality

2.1 Total Survey Error and the Survey Process

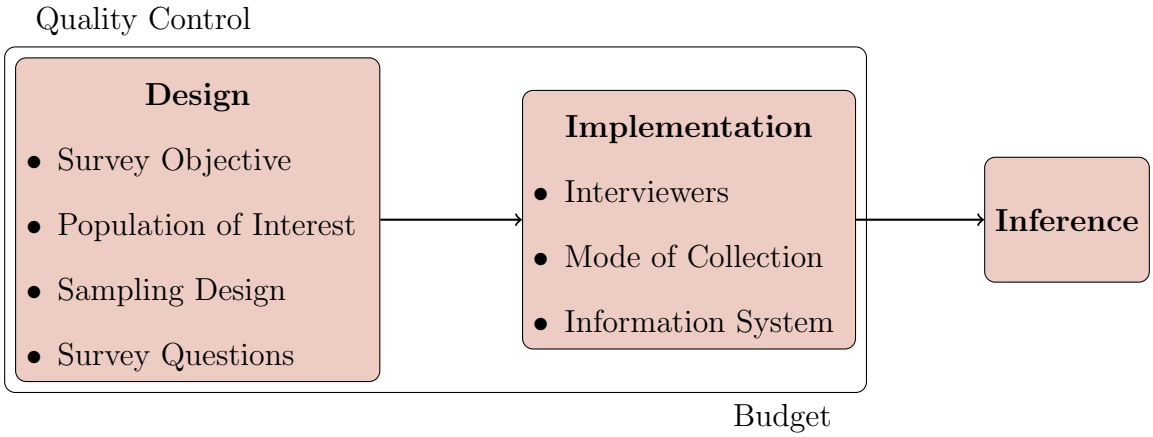
Recalling Figure 1, there are two key components of error that must be monitored throughout a survey: systemic and variable. Systemic errors tend to reinforce each other leading to an overall poor estimate of the parameter of interest. Variable errors should tend to disagree with each other and cancel each other out. When this does not occur, it affects the precision of the estimate of the parameter of interest. The reader familiar with undergraduate statistical theory will recognize these as bias and variance, respectively. Systemic and variable error together are the channel whereby minimizing total survey error achieves minimizing mean squared error. A primary objective of the survey process, then, is to define the population and frame, specify covariates, write questions, and collect data in a way that minimizes these sources of error.

It is important to recognize that non-sampling errors can not necessarily be quantified at the beginning of a survey. In other words, there is no *ex ante* way to claim, "Specification error will contribute 47.3% of systemic and 12.8% of variable error." It is true that some formulas exist for anticipating, say, the impact of frame error. The goal of a survey is to estimate the population parameters. In theory, one can not know what those parameters are, so quantifying error in the early stages is impossible. The utility of the total survey framework lies in gauging relative contributions from different sources of error once budget constraints, requisite sample size, and Types I and II error levels have been determined. The

research team can select from among different survey designs the design that simultaneously satisfies these constraints and minimizes error. After data have been collected, errors can be assessed and techniques implemented to correct for errors, some of which will be discussed later.

Although many details of the survey process can be specific to the context of a given survey, the general structure can be depicted as in Figure 2. One can segment the survey

Figure 2



A simplified view of the survey process.

process into three phases: design, implementation, and inference. Design and implementation are tied together by robust quality control procedures and budgetary considerations. In the design phase, the objective is set, the population and parameters of interest are determined, the sampling design is constructed, and the questions are written. The design phase may be iterative. One may start with a certain population in mind and refine it as information unfolds and changes the assumptions of the survey design. After beginning to write questions, one may realize that covariates or parameters are not well specified. This was my experience in preparing the survey for the CO Women's Chamber of Commerce. After realizing that there was insufficient baseline data for measuring business outcomes among women in Colorado, I was forced to change my empirical design. Questions were introduced, revised, and scrapped as we realized we were not measuring the right predictors of women's

business success. Thoughtful effort in the design phase is critical because it can anticipate problems in implementation and prevent expending effort and budget unnecessarily.

The implementation phase requires selecting the appropriate mode of data collection and information system for capturing and processing data. The mode of collection can be in-person, phone, internet, or even SMS text. Choosing the right mode presents its own unique set of challenges, not least of which is training interview staff to consistent apply protocols for data collection without compromising the sampling design. The information system must be easy enough to use that technology does not impair data collection. Finally, in the inference phase, analysis is conducted of the survey data, and results are produced. This is the final opportunity for updating the data to accommodate non-sampling error, as we will see later. It is also important that the technology make results easy to understand and timely for the end user.

Throughout each part of the survey process, countermeasures exist for mitigating and managing the effects of error. In the design phase, expert consultants and best practices can provide detailed guidance for sampling design and question writing given the objectives and context. This need not be costly. Plenty of information is publicly available. For example, the Pew Research Center is a trove of best practice guidelines from decades of research, and much of it is published on their website.[2] Another good resource is Survey Insights.[3] Using the client needs and a good foundation in theory, piecing together input from various sources can stave off a large amount of error for just a little effort. During the implementation phase, clear, simple quality control and data integrity protocols are helpful to enforce standards among interviewing staff that might jeopardize inference.

Pre-testing and piloting questions should be and integral step of every survey, no matter the budget. Poorly written questions can invalidate inference before surveys are ever distributed if researchers do not carefully scrutinize questions for details that affect comprehension and cognitive biases. Pre-testing and piloting facilitate exploring the ways that

[2]https://www.pewresearch.org/
[3]https://surveyinsights.org/

different phrases and response options can elicit confusion and misalignment. Pre-testing a question or survey involves asking five, or so, individuals from the population of interest to answer one or more survey questions and eliciting their detailed feedback about their experience. Detailed feedback helps researchers uncover issues that cause respondents to misunderstand language in the question that might lead to inaccurate responses from ambiguity, confusion, fatigue, or simply measuring the wrong quantity. Piloting questions and surveys is similar with the important exception that more people are asked to respond and may be viewed more as a feasibility study in the sense that if piloting exposes serious flaws, the survey may need to be restarted. Depending upon the context and budget, one may go straight from pre-testing to distribution.

A good example of pre-testing comes from the Nielson Norman Group (NNG).[4] In 2019, NNG updated a 1998 survey question about the role of the internet in decision-making. The 1998 question was:

> Please try to recall a recent instance in which you found important information on the World Wide Web, information that led to a significant action or decision. Please describe that incident in enough detail so that we can visualize the situation.

NNG conducted three rounds of pre-testing and a final round of piloting. In the first round, NNG updated the phrase "World Wide Web" to "online." Their research showed that the former phrase has long since fallen out of fashion to be replaced by the latter. In the results from the first round, NNG learned that the word "significant" caused confusion because it was too vague. Respondents could not choose a single "significant" decision from among their online activities. To resolve the issue, NNG researchers gave some context for what could be considered significant. Just below the survey question, they wrote, "A significant action or decision can be any change in your plans, thoughts, or actions that you consider to be meaningful." After the second round of pre-testing, NNG discovered that this caused

[4]https://www.nngroup.com/articles/survey-questions-iterative-design/

even more confusion because it biased answers toward responses that indicated how online information changed respondents behavior. This was unintended since researchers did not want people to uniformly interpret the question as asking them to identify a change in behavior. Researchers removed the explanatory sentence and added a question before the main question that gave multiple responses to suggest different ways people could use the internet to inform their decisions:

> Which of the following online activities have you done in the past month? (Please select all that apply)
>
> - Bought something
> - Watched a TV show or movie
> - Planned a vacation
> - Sent an email
> - Posted on social media (for example, Facebook or Instagram)
> - Researched a topic

This resulted in a new problem: priming from the "last-response option," which cognitively biased people toward writing a response that reflects the last option they read in the multi-select question. In their final round, the researchers removed the multi-select question and added a simple contextual sentence at the end of the question to tell respondents they were free to choose from any instance that was significant to them:

> Please try to recall a recent instance in which you found important information online, information that led to a significant action or decision. Please describe that incident in enough detail so that we can visualize the situation. If you can recall several such instances, please describe the one that was the most important to you.

After successfully piloting this question, it became the final version used in their 2019 survey. Other examples of pre-testing the language and response types of questions like this abound on the internet. There is also a wealth of research publicly available to connect the various cognitive biases triggered by question and response formatting. Exploring those is a good example of using best practices to clarify what can go wrong and how to plan for it.

2.2 Non-Sampling Error

A discussion of non-sampling error forms the core of this paper. This is as much a personal as professional decision. Much of my time in statistics classes has been spent learning theory and computation – both invaluable tools. The benefit for me in in this project has consisted of connecting those tools to the verbal, practical task of writing questions and collecting data. To that end, I hope to convey that connection in what follows.

Recall that non-sampling errors have two broad consequences: the systemic error caused by repeated errors in the survey responses that lead to over- or underestimates of the population parameter of interest, and variable error caused by the cumulative error associated with a particular question. In the context of non-sampling error, researchers can connect these to broadly defined causes associated with individual questions and data processing, as well as fundamental flaws with the sampling process. These five causes are:

- Errors in specification
- Errors in the frame
- Non-response errors
- Measurement errors
- Data processing errors

I now discuss how each of these five sources of error contributes to total survey error with examples from the NRC's NCVS report.

Specification

When a survey question and its associated variable measure different things, then there is specification error. This is caused by failing to align the definitions of the quantities being measured with the way people will understand them as conveyed in the survey questions. Several possible factors can contribute to specification error. Poorly worded or irrelevant questions and responses may cause confusion and cognitive bias. Respondents may unwittingly draw upon erroneous information that does not reflect the quantity under inquiry. A rich literature in choice architecture studies how to order and frame questions to avoid cognitive biases that can contribute to specification error. The consequence of this error type is measuring the wrong parameter, thus contaminating inference by introducing both systemic and variable error. Including erroneous covariates or excluding important covariates may also result. The best solution to specification error is a thorough understanding of the research objective and the population parameter under inquiry and careful proofreading of every question. This is an instance wherein pre-testing and piloting questions yield high value for little effort in understanding how people interpret the language of the questions.

The example from the NCVS report underscores the importance of sharing a common definition of a sensitive and emotionally charged event. In their evaluation, the NRC notes that interviewers had very clear and specific definitions of terms like "sexual assault" when conducting the survey, but there is no indication in the wording of the question that the same clarity is conveyed to the respondent – indeed, there is no evidence the respondent even has the same definition. Since one goal of the NCVS is to measure incidence of sexual assault, it is critical to their research objective that respondents have an unequivocal understanding of the types of actions that could be called by that term. Instead, the NRC claims that the question wording was too ambiguous and oversimplified to capture the full spectrum of actions that could be considered sexual assault. They write:

> CONCLUSION 8-5 There is serious specification error in the National Crime Victimization Survey measurement of rape and sexual assault. Although the

> Bureau of Justice Statistics has developed clear definitions of the concepts, they are replaced in the omnibus screener by ambiguous wording that does not convey the multifaceted concepts to respondents.

The risk is failing to accurately measure these egregious offenses and dedicate appropriate resources to helping victims.

Measurement

Measurement error is the term given to the cumulative effects of misreporting, systemic and non-systemic, and is intimately linked with specification error. Even after pre-testing and careful scrutiny of the language in the questions, people may misinterpret terms, deliberately falsify their responses, mistakenly provide erroneous information, or simply feel uncomfortable about revealing the truth. Despite researchers' confidence they are capturing the research question and representative covariates in the survey language, measurement error can be introduced by respondents' incentives to misrepresent themselves or change their behavior when being surveyed.

Two important examples of mis-reporting are *acquiescence bias* and *social desirability bias*. Acquiescence bias occurs when respondents tend to answer in the affirmative even on contradictory survey items.[5] Respondents may want to be perceived as agreeable either because they want to please the interviewer or due to cognitive bias.[6] Social desirability bias occurs when respondents try to embellish or mask characteristics about themselves that make them appear more or less socially acceptable, respectively. The effect is that their responses reflect someone attempting to conform rather than who they really are.

Poorly equipped interviewers who do not enforce compliance with the survey protocols can also affect measurement error. In general, even though measurement error can be mit-

[5]The opposite may also occur where survey respondents tend to answer in the negative.

[6]An interesting contribution to measurement theory in this vein was made by Lee Cronbach in 1942 who postulated that a respondent's recall of an event or subject matter may not allow them to have negative memories of the topic in the survey question. For more, see his *Studies of acquiescence as a factor in the true-false set*, https://doi.org/10.1037/h0054677.

igated with good question writing and staff training, some degree is to be expected. One useful, simple technique for mitigating measurement error from cognitive bias is to randomize the sequence of some questions where ordering is not critical or to randomize the responses for certain questions. By doing this, researchers can add an extra layer of protection from systemic bias in the way questions are answered.

The NCVS example of measurement error elaborates on the specification error in the survey. The NRC notes in its assessment that the questions referring to sexual assault are visually and contextually polluted by references to specific weapons and acts of violence that may suggest to respondents only certain actions can be defined as sexual assault. They write:

> CONCLUSION 8-7 Questions about incidents of rape and sexual assault in the National Crime Victimization Survey are asked in the context of a criminal victimization survey and embedded within individual questions that describe other types of crimes. This context may inhibit reporting of incidents that the respondent does not think of as criminal, did not report to the police, or does not want to report to police.

The crucial and nuanced distinction between measurement and specification error in this instance is that the visual structure of the sequence of questions, rather than the specific language of the question itself, can cause respondents to fail to report sexual assault because it primes responses of a certain type, namely aggressive and potentially violent, even when assault may not be overt.

Frame

The term *frame* is used here to indicate the selection criteria and source material for sampling from the population of interest. For example, the criteria "hospital inpatients" and "hospital patients" are potentially two different frames. The researcher may be interested in hospital patients who are admitted for both inpatient and outpatient procedures. The terms "hospital

inpatients" and "hospital patients" risk identifying two different populations. A frame may be generated algorithmically, such as selecting an item from a list only if a random number generator lands on a pre-determined number.

Frame error is the result of several possible problems: mis-specifying the frame, lack of enforcement of random selection, erroneously including or excluding population elements, elements entering or leaving the frame in a non-random way during sampling, incomplete or outdated sources, just to name a few. A closely related term is "coverage error" where an inadequately[7] defined frame omits elements from the sample that are representative of the population. A frame that only captures inpatients with certain categories of diagnoses when the population of interest is all inpatients is an example of coverage error. The consequence of frame error is quite harmful to the survey: bias in the parameter estimates or measuring the wrong parameter entirely. This is especially problematic when population elements are excluded from the frame in a non-random way because inference can be invalidated. Controlling frame error lies in correctly specifying the frame before sampling is conducted and enforcing standards during sampling. Mathematical methods may be employed to correct for frame error when conducting the analysis (Biemer & Lyberg, 2003, 1st ed, p68-77).

The NRC underscores the gravity of frame error in the context of the NCVS. The NCVS uses a two-stage frame for sampling individuals. The first stage is borrowed from the methodology used by the Census Bureau in sampling for the decennial census for selecting counties and metropolitan areas. The second stage consists of a combination of address files also from the Census Bureau, specifically the Master Address File and a "non-institutionalized group quarter" file. The first file includes residential units, and the second includes places where particularly vulnerable populations may temporarily reside, for example substance abuse halfway houses and facilities for developmentally disabled and physically handicapped, as well as college residence halls. The NRC note that the Master Address File is potentially 15

[7]This inadequacy need not be the fault of the researchers. Legal definitions, for instance, may play a role. Consider Colorado microbreweries, who may produce beer for package sale or operate a restaurant, but not both. A surveyor unaware of this distinction may exclude some breweries from a representative sample.

years old when the NCVS uses it to conduct surveying activities because of the frequency of updating by the Census Bureau, and that the non-institutionalized group quarters file selection methodology does not sufficiently cover a representative sample of these facilities housing people at high risk for sexual assault. The NRC write:

> CONCLUSION 7-4 The frame for the ancillary listing of group quarters, which is an important part of the secondary sample for the National Crime Victimization Survey because their residents may be at higher risk for sexual violence, is seriously flawed in terms of both the building and enumeration of this secondary frame.

The consequence here is quite damaging: failing to capture the most vulnerable population individuals in the sample results in estimate of sexual assault and rape incidence that is biased toward finding no effect because the worst outcomes are not sampled.

Non-response

Non-response error occurs when there is an insufficient number of survey responses to conduct inference and takes on three forms: a unit may not respond to the survey, such as a whole household or an individual within a household; an item on the survey may not have a response; or an item on the survey may have only a partial response. Each of these has different consequences for survey and experimental design. Too many units not responding means the central limit theorem will not apply, and inference will be compromised. Too many non-responses to a particular item may eliminate a covariate from consideration in the analysis. If this covariate is simultaneously correlated with predictors and outcomes of interest, bias is likely to be a problem. Incomplete responses are problematic based on context and can have consequences for both precision and systemic error if non-responses are correlated in some way. Just as with frame error, if non-responsiveness is non-random, then bias in coefficient estimation is a concern. One simple example is that people who may not respond to questions about education may not do so in a way that is correlated with

their average level of education. When this happens, inferences made about education and related covariates in the sample will be compromised.

Anticipating and correcting for non-response error can be tricky because the researcher should be careful not to correct for non-responsiveness in a way that biases the sample. In the case where researchers have the opportunity to resolve refusals to respond, the researcher would like to enforce the random selection in the design by urging these individuals to respond. If the non-responses are random, then this is only an asymptotic issue, and the researchers simply need to ensure a sufficient number of responses to achieve the necessary sample size. Offering incentives or inducements to respond can be helpful in increasing the response rate, but they may contaminate the random selection by introducing spurious correlations from people who may not have otherwise responded. In that case, researchers may be measuring the response to the inducement and not the variables measured in the survey. Confidentiality can reduce non-response error if survey participants are concerned about their identities or data being revealed.

The NRC finding in the NVCS context is an example of potentially non-random non-response error. In 2009, the BJS contracted with an organization to audit their non-response correction methods in response to scrutiny that non-response rates seemed implausibly low for several years. The contractor's results, however, showed that "ignorable" non-response rates were disproportionately correlated with being male, black, and under 25 years old. Furthermore, the contractor's methodology assumed, without justification or testing, that individuals who respond at least once but not routinely are missing at random, a highly questionable assumption in the context of sexual crimes. The NRC report concludes:

> CONCLUSION 8-1 The overall unit response rates, as calculated, on the National Crime Victimization Survey are moderately high and have been reasonably stable over the past 10 years. Although an independent analysis concluded that the methods that the Bureau of Justice Statistics uses to adjust for non-response appear to provide a satisfactory correction for non-response bias at the unit

level, our panel has reservations about that analysis and remains concerned that there may be a non-response bias related to sexual victimization.

The non-response rates are likely correlated with sexual victimization, an outcome measured in the NCVS, and are, therefore, highly suspect.

Data Processing

Data processing error can occur whenever someone touches the data: recording, coding, transposing, editing, or tabulating. Data processing error can cause a multitude of problems affecting both systemic and variable errors. These errors are not always technical in nature. In other words, they are not always the result of computation or recording mistakes. They may also result from features and processes in the information technology that do not adequately facilitate recording the information at hand. I will show an example of this presently. Poorly equipped staff may make mistakes with coding and cleaning data, such as with open-ended, categorical, and missing responses. Computational errors with mathematical procedures that correct for probability weighting or coverage error are also sources of data processing error. The best remedies for this type of error are quality control, good training of survey personnel, and an adequate information system for the survey and analysis.

The NCVS example here puts the full consequence of mishandling data on display. The NCVS information system includes an automated process for classifying "serious criminal victimizations," a classification scheme which includes and "unclassified" option. However, the NCVS methodology provides no guidance on how often respondents are coded with this unclassified option, nor is there information regarding the overall error rate for this automated process. The NRC identify this as a lack of transparency in a process vital for users of the survey information:

> CONCLUSION 7-5 The Bureau of Justice Statistics does not provide public information on the edit process in the National Crime Victimization Survey...

The lack of transparency about these processes makes it difficult for data users to fully understand the survey's estimates.

This example shows that data processing errors can occur when the information system is not carefully checked to ensure it aligns with assumptions of the survey methodology.

Figure 3 is a helpful tool for organizing the influence of different errors in survey design. It has been my intention in describing the sources of various errors to convey that sometimes,

Figure 3

Survey Error Component	Systemic Risk	Variable Risk
Specification	-	-
Non-response	-	-
Frame	-	-
Measurement	-	-
Data Processing	-	-
Sampling Error	-	-

Organizing Total Survey Error Component Risks

errors are ambiguous as to their classification, and they may not fit neatly into one or the other category. The surveyor should not despair in this case. In the absence of experience, it often suffices to record a source of error and return it as new information is incorporated or refrain from classifying it entirely. Rather, simply addressing the source of error and attempting to accommodate it at the inference stage is more important that classifying it correctly. Remember, the total survey error framework is an organizational tool, not a rigorous mathematical authority.

3 Sampling for Surveys

Survey questions and inferences are only valid on the conditions that the sample represents the population of interest and that the sample is randomly selected. A non-representative sample invalidates inference because estimates do not capture the parameters under inquiry and, therefore, do not answer end users' questions. The statistical theory underpinning

parameter estimation rests on a foundation of random sampling. No discussion of survey error is complete without some reference to possible sampling designs and their advantages and disadvantages. This section explains several random sampling designs and alternatives when random sampling is infeasible. A couple good references for calculating estimators from a sample are Thompson (1992) and Lohr (2010). Before beginning, it is helpful to establish the difference between a population *element* and a population *unit*. An element is the most primitive or least rarefied entity in the population that can be sampled in a survey. A unit is the unit of observation relevant for inferential analysis, and it may or may not be the same as an element. The classic example is the household and its inhabitants. The people living in the house are the elements, and the household is often the unit of observation, though that need not be the case. Missing individuals from a household is less problematic than missing an entire household if indeed the household is the unit of observation.

Simple Random Sampling

Simple random sampling is the method familiar to every student of probability and statistics. Given a population of N elements, n are selected without replacement and with $n < N$. This type of sampling is often referred to as "probability sampling" because every possible sample of size n has an equal probability of selection in the sample frame. Random sampling is preferred because estimators, such as means and confidence intervals, can easily be calculated from statistical theory. Sampling variation is also relatively straightforward to describe and estimate. Estimators derived from a random sample are commonly shown in upper division and graduate level statistical theory courses to yield unbiased and minimum variance estimators for important quantities such as the population mean and the standard error of the sample mean.

Simple random sampling is rooted in the frequentist[8] view of probability and statistics which asserts that a random sample is one of a multitude of possible random samples that

[8] As contrasted with the Bayesian view.

could have been drawn. A strong assumption underlying asymptotic theorems in this view is that of *independent* samples, for it is only under the conditions of repeated independent, random samples that we get such celebrated results as the Central Limit Theorem. This has two implications for measuring survey error. In the presence of non-sampling errors listed above, a particular sample will likely not be randomly selected and may have a high degree of bias. That is, estimators may yield results that are, on average, significantly higher or lower than population parameters. Asymptotic theory assures us that an independent, random sample eliminates this problem because such samples, on average, yield estimates that are equal to the population parameters. This why researchers must strive to minimize non-sampling error and enforce random selection.

The second implication for a particular sample is that it may also exhibit a high amount of sampling variation, a term which I used previously and will describe in some detail here. Suppose, under the frequentist view, one were allowed to take 1,000 samples of the height of freshman at CU Denver. The *means* of these samples themselves take on a distribution, and this is called the *sampling distribution*, a term no doubt familiar to students of statistical theory. If these samples are all independently and randomly drawn, then this sampling distribution will be centered at the population mean, regardless of sample and population sizes. Furthermore, for very large sample sizes, this sampling distribution will have a low variance and standard deviation. This standard deviation is called the *standard error* of the mean to distinguish it from the more common use of standard deviation. Inferences made on independently, randomly chosen samples of sufficient size have low standard error, so estimates are precise – they do not drift very far apart from their center. As sample size decreases, this sampling distribution widens, lowering the standard error and making estimates less precise.[9]

[9]A somewhat confusing matter to point out given this explanation is that *non-sampling* error in a survey is a property of repeated sampling because it relates to the *average* outcomes that could over- or under-estimating some quantity of interest among all possible samples. *Sampling* error is a property of the specific sample drawn for the survey because lack of enforcement of randomization or insufficient sample size may result in large standard errors and imprecise measurements.

Non-sampling error can affect standard error. For example, poorly worded or ambiguous questions that cause confusion can lead to a wide variance in the estimate of the associated variable it attempts to measure, increasing the corresponding standard error. In this view, enforcing the random frame of sample selection is imperative because it minimizes mean squared error.

Stratified Random Sampling

Stratified random sampling is not much different from simple random sampling. The difference consists in randomly sampling exhaustive and mutually exclusive sub-groups of a population that exhibit sufficient within-group homogeneity rather than drawing a single sample containing element from all sub-groups. The sub-groups are chosen to reflect the proportions of those sub-groups in the total population. A very simple example is stratification by biological sex. Suppose a local town consists of 50% females and 50% males, but the country this town is in consists of 55% women and 45% men. A simple random sample that selects 50% each of women and men from this town may under-represent women and over-represent men in estimates of total population characteristics. By contrast, one could deliberately structure the sample frame so that it samples 55% women and 45% men making it more representative of the overall population. As one might notice, this approach can be useful if the survey objective includes obtaining results delineated along some pre-determined dimension, such as race, socio-economic status, cohort, set of behaviors, or the like.

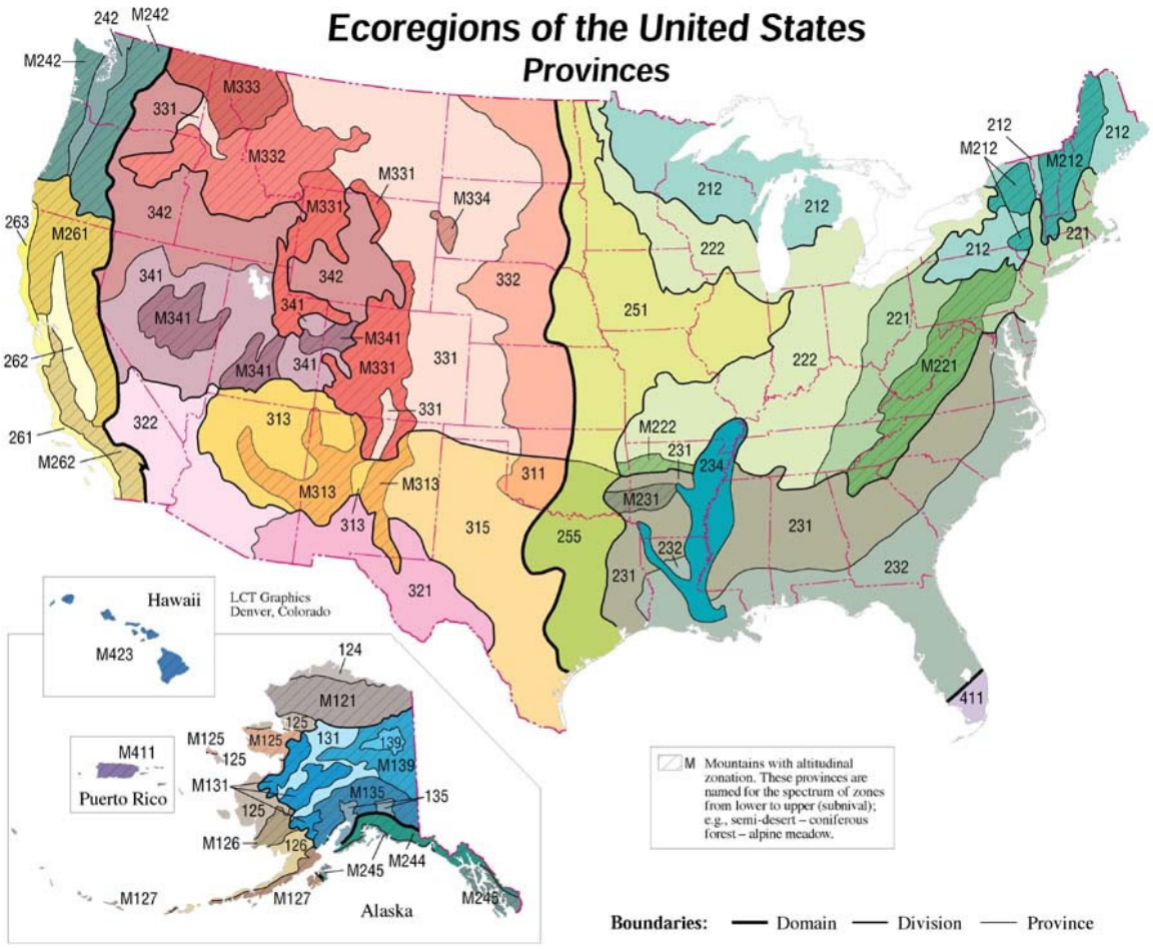
This sampling design affords the researcher lower sampling variation from independent, random samples of the overall population and can be useful to employ under the right circumstances. This is because the variance of an estimator in the stratified design is the sum of the variances of the estimators of the individual sub-groups. Estimators for a stratified sample are themselves weighted by the ratio of the sample variance to the sample size of each stratum. The result is greater precision than otherwise might be obtained from a simple random sample (Thompson, 1992). The sub-groups are more homogeneous than the

overall sample. This may have the added advantage of reducing the required sample size. Suppose the initial sample is determined to be 1,000 units to obtain evidence the effect under investigation, but sub-groups within the sample are small, leading to high standard errors for these sub-groups. Prior to sampling, if researchers can determine the required sub-group sample sizes to detect an effect, then researchers can sample directly from those sub-groups, and the overall sample size may decrease if the relative sizes of the sub-groups are small enough. This is because variance of the estimators is related to the *ratio* of the sub-group sample variance to the sub-group size, not simply the size (Thompson, 1992, 1st ed, p102-105).

I will introduce an example of stratified sampling to which I will return for cluster sampling. In 2002, the US Forestry Service introduced a sampling design for estimating camp site use among national parks (Zarnoch et al., 2002). There are 606 forest ranger districts spread across 192 million acres in the continuous 48 states in the National Forest System. The objective of the survey was to estimate the number of recreational (i.e. - hiking, camping, horseback riding) site visits in the forest system. The overall sampling design was a composite of three stages. The first stage stratified the forest system into ecoregions based on similar climatic, topographic, and biodiversity features. Some of these ecoregions extend into Canada and Mexico, and they are used by several US agencies such as the Geological Survey and Environmental Protection Agency.

As one might imagine, ranger districts are not uniformly distributed throughout the United States. There are larger densities along the Appalachian trail and adjacent regions in the east, throughout the mountainous regions of the west, and throughout the Pacific coast states. One may also expect that visits to each site exhibit a high degree of heterogeneity across the sites because of seasonal considerations. Considering these factors, researchers designing the sample deliberately elected to stratify the national forest land according to ecoregions and then perform a second stage of random sampling within each ecoregion. This ensured they had sufficient representation of the whole forest system and had the

Figure 4



Source: https://www.fs.fed.us/rm/ecoregions/images/maps/ecoregions-united-states.jpg

flexibility to change sampling strategies within each ecoregion if the circumstances warrant. After stratification, a random sample of 32 ranger districts was taken, with proportional representation of stations across all ecoregions.

Cluster Random Sampling

Cluster random sampling begins by defining groups of primary observation units, such as residential blocks, that contain secondary observation units, such as households. Ideally, these primary observation units are selected to exhibit the same proportion of heterogeneous

characteristics as the overall population of interest. For example, a residential block should reflect the same race, age, education, income, and gender diversity as the national population. Once the clusters of primary units are defined, a random sample of clusters is selected, then all secondary units are surveyed within each sampled cluster. Similar to stratification, cluster random sampling can make sampling easier to manage and more flexible for large sample sizes or when spacial orientation of observation units is a factor. It also ensures that the whole population is sampled in a representative way.

One major disadvantage is that the ideal is not always achieved in practice. Demographic characteristics are not uniformly distributed across towns and cities, and these are often correlated with socio-economic, racial, and other data. This is especially problematic if primary units were clustered based on stale information, and the clusters are found to be less representative than expected. Furthermore, clustering can result in wider sampling variation because clusters are sums of secondary units, and these units should have a high degree of heterogeneity to be representative. These factors can make choosing appropriate clusters challenging, and clustering is often deployed as part of a multi-stage random sampling strategy. Unbiased estimators can also be derived for cluster sampling by accounting for the primary and secondary units.

Continuing the US Forestry sampling design, recall the first stage in sampling was stratification followed by randomly sampling ranger districts. The ranger district was the primary unit. The second stage involved stratifying recreation sites according to site-days, the secondary unit, to accommodate the large amount of variation in seasons and activities across sites. There were five site types, two season types, and two day types, creating 20 possible strata for selecting the secondary unit. Once the stations were sampled, the researchers selected 70 site-days per district. The third stage of sampling was to conduct a two-minute survey with every car leaving the site for the last time during the site day. This example highlights the range of flexibility that can be utilized by combining basic sampling techniques to tackle a large amount of area and land use.

Non-Random Sampling

Often, random sampling in infeasible. This may be due to budget constraints, limited research resources, the scope of the survey objective, time constraints, or other factors. In these cases, researchers must resort to non-random sampling methods. These are often called "non-probability" sampling methods because not all possible samples have equal probabilities of being selected. The disadvantage to non-random sampling is that there is no way to guarantee that the sample is representative of the population of interest. I will briefly discuss a few non-random sampling options. These non-random samples should not be dismissed as not useful. Often, they provide insight in the formation of a survey for lines of inquiry or attributes to measure that researchers may not have otherwise considered.

Sample Matching

Sample matching can be used in contexts where comparing two groups to detect an effect is the research objective. In this case, treatment and control groups of individuals are selected such that the groups are matched along attributes, such as age, education, income, gender, and race. The goal of this matching is to obtain two groups of people who are alike in every way except the intervention that one group has received. However, there is no randomization, so there is no way to verify that all possible underlying factors are comparable and not correlated with the intervention and outcome of interest. For example, researchers can not verify that simply being willing to be sampled and surveyed is not correlated with some unobservable attribute that makes someone more or less likely to be in the treatment group and exhibit the outcome of interest.

Convenience

A convenience sample is a sample that a researcher selects because people are easy and convenient to access, such as people within one's own town, age cohort, or workplace. Convenience samples are non-random because people in certain groups who share common attributes or

interests likely share other traits, as well. This makes them a biased sample.

Voluntary or "Opt-in"

Voluntary samples are samples of individuals who choose to be surveyed. This suffers the same problem as convenience samples. People who opt into being surveyed may share some unobserved trait that is correlated with the effect the researchers hope to find. Voluntary samples are actually quite commonplace. Voter polls, for example, are voluntary samples. People who vote are more likely to be sampled for a voter issue survey than people who do not vote. Voluntary samples are useful for sampling via the internet. Designing and distributing an email or internet survey is helpful because people spend so much time online, and re-weighting techniques exist to adjust internet surveys to approximate a random sampling design, making inferences more accurate. Also, raking and propensity matching may be used to approximate random design. For more information, see the Pew Research Center page on weighting methodologies in the context of online samples.[10]

Quota

A quota sample shares similarities with stratification and sample matching. In a quota sample, the researcher determines a minimum number of people to survey from sub-groups that are mutually exclusive and exhaustive and that share some degree of homogeneity. Ideally, these sub-groups would match each other along all other dimensions, but that may not be enforceable. Once again, the problem is that, in the absence of random selection, there is no way to rule out other characteristics that may be correlated with covariates of interest. As with voluntary samples, techniques can be deployed to approximate random sampling.

[10] https://www.pewresearch.org/methods/2018/01/26/how-different-weighting-methods-work/

Probability Weighting

Probability weighting is a mathematical technique for changing the probabilities of elements' selection into the sample to more accurately reflect their proportion in the population of interest. This takes either or both of two forms. *Base weights*[11] are used before sampling is conducted. As an example, recall our fictional town comprising 50% men and 50% women. In an equal probability sample, men and women each have a 0.50 probability of being selected. If the national population is 45% men and 55% women, the researcher can adjust probabilities of selection to 0.45 for men and 0.55 for women. This adjustment should produce a sample of individuals more reflective of the national population by giving under-represented individuals a higher probability of selection, and over-represented individuals, a lower probability.

Adjustment weights[12] are used at the end of the survey process to decrease bias caused by non-sampling error due to coverage or non-response. These weights compensate for missing observations both at the unit level and the item level. Adjustment weighting can be a complex technique, and a full treatment is beyond the scope of this paper. However, to supply a simple example, we depart from our initial fictional town and find another fictional town where the proportion of women and men are 55% and 45%, respectively. We survey people in this town without using base weights because we believed these proportions to align with the national proportions, but we suddenly learn that this proportion has changed to 50% each women and men. In this case, our survey has been conducted, and we need to adjust for the fact that men are 5% under-represented in our sample, while women are 5% over-represented. One simple way to perform adjustment weighting is to use the proportion of the population weight relative to the sample weight. For women in our new sample, this is $\frac{0.50}{0.55}$ for women and $\frac{0.50}{0.45}$ for men. If we use base weights and adjustment weights in the same survey design, then the final weight applied is the product of the base and adjustment weights. An excellent resource on weighting is Lavallée and Beaumont (2015).

[11] In my research, I saw these go by the names design weights and estimation weights.

[12] These are also called post-stratification weights or auxiliary weights, although I am unsure how common these names are across categories of researchers.

4 Closing Remarks

The field of survey design is huge, and much ink has been spilled and pixelated over the question of how to achieve optimal survey design. Even the best survey design *ex ante* can be disrupted by practical considerations in the field. All this is to say that there is much I left out of this paper. For example, I have said nothing of Bayesian and non-parametric methods which can be constructed to suit all sorts of sampling issues. Nor did I cover sampling methods designed to suit elusive and rare events or adaptive methods. There are mountains of textbooks and terabytes of webpages available to anyone curious on these things. I chose to focus on the total survey error framework because it seems to me that, for students with a solid foundation in statistical theory, connecting the dots with the less statistical side, such as matching survey questions to sources of statistical error, is a useful supplement. I hope I have achieved that.

References

[1] Biemer, Paul P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *The Public Opinion Quarterly, 74*(5), p817-848. https://doi.org/10.1093/poq/nfq058.

[2] Biemer, Paul P., and Lars E. Lyberg. *Introduction to Survey Quality*. Hoboken, John Wiley & Sons, Inc., 2003.

[3] Lavallée, P. and Beaumont, J-F. (2015). Why We Should Put Some Weight on Weights. *Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach, Invited article*. Retrieved from http://surveyinsights.org/?p=6255

Lohr, Sharon L. *Sampling: Design and Analysis*. Boston, Brooks/Cole Cengage Learning, 2010.

[4] National Research Council. (2014). Estimating the Incidence of Rape and Sexual Assault. Washington, DC. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK202264/.

[5] Thompson, Steven K. *Sampling*. University Park, John Wiley & Sons, Inc., 1992.

[6] Zarnoch, S. J., Kocis, S. M., Cordell, H. K., English, D.B.K. (2002). A Pilot Sampling Design for Estimating Outdoor Recreation Site Visits on the National Forests. Res. Pap. SRS-29.Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 20p.