Paper 3274-2015

Statistical Analysis of Publically Released Survey Data with SAS/STAT® Software SURVEY Procedures

Donna Brogan, Emory University

ABSTRACT

Several U.S. federal agencies conduct national surveys to monitor health status of residents. Many of these agencies release their survey data to the public. Investigators might be able to address their research objectives by conducting secondary statistical analyses with these available data sources. This paper describes the steps in using the SAS/STAT® SURVEY procedures to analyze publicly released data from surveys that use probability sampling to make statistical inference to a carefully defined population of elements (the target population).

INTRODUCTION

Federal and state governments in the United States conduct annual or periodic probability sample surveys for a wide variety of purposes, e.g. to assess characteristics of the human population, the economy, society's institutions, and the environment. Within the past 25 years data and accompanying documentation from some of these surveys have been released to the public, giving researchers the opportunity to conduct secondary data analyses. These datasets are a rich resource of free or inexpensive data, often called "complex survey data", which can be used to answer a wide array of research questions or test research hypotheses.

However, some researchers may not be aware of unique complexities associated with using this resource. First, making statistical inference to a finite population uses a different conceptual framework than is used in standard statistical methods, impacting analysis objectives and interpretation of results. Second, reading survey documentation requires familiarity with technical terminology related to probability sampling.

Third, standard statistical procedures used by researchers were developed for data obtained via simple random sampling, a sampling plan virtually never used by real probability surveys. Thus, specialized statistical procedures, such as the SAS/STAT® SURVEY procedures, are used for analysis of complex survey data.

Fourth, skip patterns are common in survey data collection instruments, requiring the frequent use of subpopulation analyses and careful variance estimation. Additional complexities likely already familiar to researchers include use of detailed coding manuals, substantial data management tasks, and merging data files.

This paper explains the statistical context for analysis of a complex sample survey dataset, the steps involved in locating and analyzing a survey dataset, and interpretation of the results of analyses. The points of the paper are illustrated by using the "sample adult" dataset from the 2013 National Health Interview Survey or NHIS (National Center for Health Statistics, 2015A). The National Center for Health Statistics (NCHS) of Centers for Disease Control and Prevention (CDC) conducts the annual NHIS, and the "sample adult" dataset is one of many datasets available from the 2013 NHIS (National Center for Health Statistics 2015B).

STATISTICAL CONTEXT OF PROBABILITY SAMPLING

Secondary data analysts of complex survey data work within the context of sample survey methodology (Groves et al 2009), an interdisciplinary field, and, in particular, within the context of probability sampling (Lohr 2010), a subspecialty of statistics. Researchers are familiar with standard statistical methods and with specialized statistical methods used in their subject matter fields such as biostatistics or epidemiology. Often, however, they are not trained in general survey methodology or in specialized statistical methods for complex survey data.

Standard statistical methods typically assume a probability distribution that produces values of a random variable, theoretically infinite in number, where some observed values of the random variable (the data) are used to estimate parameters of the probability distribution. Generally, standard statistical methods assume that the data are independently and identically distributed in order to estimate parameters.

The paradigm for sample surveys is different. A given and *finite target population* of elements (members) is of interest, and it is desired to know the value of *parameters* of this target population. For example, the target population for the 2013 NHIS "sample adult" dataset is civilian noninstitutionalized adults (age 18+ years) who reside in a household (occupants of a housing or dwelling unit) within the U.S. in 2013. The size of the target population is 237 million, a number which is a default output from any of the SAS/STAT SURVEY procedures used to analyze this dataset.

Two target population parameters of interest are number and percentage of adults in the target population who have arthritis, the latter quantity being the population prevalence of arthritis. It is useful to give the mathematical definition of these two parameters to compare with later formulas that show how these parameters are estimated. Let y_i be the value of the dichotomous variable arthritis, coded as 1= yes and 0 = no, for element i in the target population, i = 1, 2,…,N, where N is approximately 270 million. The target population parameter Y is the number of adults in the target population who have arthritis, defined as:

$$Y = \sum_{i=1}^{i=N} y_i \qquad (1)$$

The target population parameter P is the percentage of adults in the target population who have arthritis, defined as:

$$P = \frac{(100)Y}{N} = \frac{100}{N} \sum_{i=1}^{i=N} y_i \qquad (2)$$

Taking a census and measuring all 237 million elements in the target population to determine the values of Y and P above is not feasible. Thus, a probability sample of elements is selected, using one of many possible sampling techniques. *Probability sampling* means that each element in the target population has a nonzero and known (to the sampler) probability of being selected into the sample. The sampled elements are measured on the characteristic of interest, e.g. arthritis. The collected data are weighted, survey design variables are constructed, and the data are analyzed with survey-specific statistical procedures to produce point estimates of identified population parameters.

To derive formulas for estimators of population parameters, as well as formulas for the estimated variance and standard error for each estimator, the characteristic y (arthritis here) is assumed to be a *constant* for each element i. This assumption differs from standard statistical methods that consider random variables. The randomness in the finite population paradigm results from the idea that the one sample that is obtained is just one of many possible samples that could have been obtained. Expectation of estimators and variance estimation for estimators is done by taking the expectation over all possible samples, i.e. conceptualizing the implementation of the given sampling plan many times (never done in reality).

CHOOSE A COMPLEX SURVEY DATASET

If you plan to use one or more publically released survey datasets, first define your target population, research questions and research hypotheses. Then begin the search for survey datasets that cover your target population and include the variables in your research questions and hypotheses. Depending upon what survey datasets you locate, you may revise your research questions or hypotheses and your target population definition.

Once you find possible survey datasets that include your target population and specified variables, read the survey documentation to determine the following:

- Are the variables (or items) of interest to you defined and measured adequately in the survey?

- Are the item nonresponse rates low enough for items of interest to you?
- Is the survey response rate adequate and correctly reported?
- Were data collectors trained adequately and were quality control procedures in place for data collection?
- Do the various sampling frames have adequate coverage of your target population?
- Does the sampling plan seem reasonable, and was it implemented satisfactorily?
- Does the weighting procedure seem reasonable, and does it include satisfactory nonresponse adjustments and post-stratification procedures?
- Does the documentation explain which weighting variables to use for which analyses?
- Does the documentation include guidance on how to describe the sampling plan to survey software? This information typically is in a section called “variance estimation”.
- Are the datasets provided as ASCII files, SAS datasets, or in some other format?
- If the datasets are provided in ASCII format, are electronic files provided to convert the ASCII files into a SAS dataset, along with format information?

Some of the questions above may be difficult to answer without background in survey methodology, but later sections of this paper may provide some guidance. Just because a survey dataset is publically available on the WEB does not automatically mean that it is a good quality dataset to use for your research.

For researchers interested in health, there are several CDC WEB sites, as well as other sites, that provide complex survey datasets. NCHS of CDC conducts many health surveys, some annually and some periodically.

OVERVIEW OF PROTOTYPE PROBABILITY SAMPLING PLANS

Secondary survey data analysts should be familiar with prototype sampling plans for probability surveys and with typical survey design variables in complex survey datasets. This knowledge helps you describe to the survey software the dataset’s sampling plan so that the software uses correct formulas for point estimates and estimated variances. This section gives an overview of prototype sampling plans used today. More detail can be found in Lohr (2010).

SIMPLE RANDOM SAMPLING

Simple random sampling is the simplest method of probability sampling. In order to select a simple random sample of n elements from a target population of N elements, the required *sampling frame* is a list of all N elements in the population, individually identified so that contact can be made with each element selected into the sample. A simple random sample is selected in such a way that each possible sample of n elements from the target population of N elements has an equal chance of being the one sample actually selected. A corollary of simple random sampling is that each element in the population has an equal chance of being selected into the sample, and that probability is n/N.

Simple random sampling rarely is used because generally it is not possible to construct the required sampling frame. Even if a sampling frame could be constructed, simple random sampling may spread out the sample over such a large geographic area that data collection costs would be prohibitive, especially if in-person interviews are conducted with elements (persons) as in the NHIS and in NHANES, the National Health and Nutrition Examination Survey (National Center for Health Statistics 2015C).

STRATIFIED RANDOM SAMPLING

The required sampling frame for stratified random sampling is the same as for simple random sampling. Before sampling begins, however, each of the N elements on the sampling frame is classified into one of two or more strata (groups) based on known or approximate values of one or more characteristics. Typically a simple random sample or a systematic random sample is selected within each stratum.

Stratified random sampling often is used when it is desired to *oversample* elements in one or more of the strata in order to increase sample size in that stratum. Elements that are oversampled are given a higher probability of being selected into the sample, compared to other elements.

PRAMS (Pregnancy Risk Assessment Monitoring System), a statewide survey sponsored by CDC that is conducted annually in about 40 states, uses stratified random sampling (Centers for Disease Control and Prevention 2015A). The target population is pregnancies that ended in a live birth during the calendar year to a woman resident in the state. Birth certificates are used to construct a monthly sampling frame which is stratified by one or more variables on the birth certificate. Women who delivered a low birth weight infant typically are oversampled. Systematic random sampling is used within each stratum

Note that *both* simple and stratified random sampling take only one stage of sampling to get to the elements selected into the sample. Each element selected into the sample is called a *primary sampling unit* **(***PSU)* because it is selected into the sample at the *first* stage of sampling.

STRATIFIED ONE STAGE CLUSTER SAMPLING

Although this sampling plan is not used often in practice, it is useful to define because sampling plans that are more complicated frequently are approximated as stratified one stage cluster sampling for the purpose of variance estimation.

In one stage cluster sampling the target population of N elements is naturally grouped together in M different clusters, where the number of elements in each cluster is 1 or more. An example target population is all household residents of a particular town, and the residents are grouped or clustered together in M different housing or dwelling units. The first stage sampling frame is the list of all M clusters in the target population, i.e. housing units in the town, and these PSUs generally are stratified prior to sampling. Within each stratum, typically a simple random sample or systematic random sample of clusters is selected. For each cluster that is selected into the sample, *all* elements of that cluster come into the sample. Thus, there is no second stage of sampling for elements, for a total of one stage of sampling. The PSU is the cluster, and sampled elements are clustered within their PSU.

STRATIFIED MULTI-STAGE CLUSTER SAMPLING

Sometimes the target population is defined as persons in the United States who are resident in or affiliated with particular institutions during a given time frame such as a calendar year. Example institutions are nursing homes, prisons, elementary schools and hospitals. Often the first stage sampling frame contains geographic areas such as counties (the PSUs) that cover the U.S. Frequently these PSUs are stratified by region of country and other characteristics before first stage sampling. PSU's within each first stage stratum typically are selected into the sample with *probability proportional to the estimated size (ppes)* of the target population within each PSU.

Once a PSU comes into the sample, a sampling frame of institutions within that PSU is constructed, where institutions are the second stage sampling units (SSUs). The institutions may be stratified before second stage sampling. Typically institutions are selected into the sample with ppes of the target population within each SSU.

Once an institution is selected into the sample, a typical third stage sampling frame may contain all persons within the institution. This frame also may be stratified before elements are selected into the sample. In this scheme three stages of sampling are used to get to selected elements for the sample: county, institution and element (person).

There are many variations on this prototype sampling plan, including the number of stages of sampling. Elements selected into the sample via multi-stage sampling are clustered, in this example within institution and also within the PSU (e.g. county). Many National Health Care Surveys conducted by NCHS use stratified multi-stage cluster sampling (National Center for Health Statistics 2015D).

AREA PROBABILITY SAMPLING

Area probability sampling (APS) is a special case of stratified multi-stage cluster sampling discussed above and is treated separately because it is used frequently. The target population typically is defined as persons who live in a given geographic area in a household, where household usually is defined as all

residents of a dwelling or housing unit. The PSU is a geographic area such as a county, and the first stage sampling frame almost always is stratified. Second and third stages of sampling proceed to smaller geographic areas like census tracts or block groups. When small enough geographic areas, sometimes called segments, are selected into the sample, field workers identify (count) all housing units within that sampling unit and construct the next stage sampling frame by listing the housing unit addresses. This "counting and listing" phase is moving toward being more automated to reduce expense. Typically a systematic random sample of addresses is selected from the housing unit frame. Generally ppes sampling is used for sampling units at all stages except for the last stage of selecting housing unit addresses.

Interviewers visit each sampled address to determine if members of the target population reside in the housing unit. Sometimes subsampling is done among eligible residents of the housing unit. Data collection typically is done in the home via a personal interview for persons selected into the sample.

There may be five or six stages of sampling before an element is selected into the sample. Observations in the dataset are clustered together in their neighborhood where they live and, ultimately, within the PSU (first stage sampling unit) from which they are selected.

Frequently this prototype sampling plan is called a household survey, and there are many variations on the description given here. Example NCHS surveys that use area probability sampling are NHIS and NHANES.

TELEPHONE SAMPLING

The target population is defined as persons (perhaps in a restricted age range) living in households within a given geographic area, e.g. the United States. The objective of telephone sampling is to call randomly selected telephone numbers to reach the target population.

Telephone RDD (random digit dialing) methodology using landline phone numbers became a popular sampling method in the late 1980's when only about 5% of U.S. households did not have a landline telephone (Blumberg et al 2008). The sampling frame of all possible landline telephone numbers is generated by listing all area code and prefix combinations that operate for landline telephones in the given geographic area, and then generating all possible last four digits of the 10-digit telephone number, 0000 to 9999, for each area code and prefix combination. This method ensures that unlisted landline telephone numbers are on the sampling frame. Methodological details have changed over the past thirty years to improve efficiency. Current landline methods, usually some variation of list-assisted RDD (Dillman et al 2014), remove from the sampling frame businesses and phone numbers that are judged unlikely to go to residences. The remaining telephone numbers on the frame are stratified by geography and by likelihood of being a residential phone, followed by stratified random sampling with oversampling of numbers more likely to be residential.

All sampled landline numbers are called. When a sampled number reaches a residence, screening questions determine if any household members belong to the target population. If yes, then one or more of those eligible is selected for a telephone interview. Today elements (persons) come into the landline sample after one or two stages of sampling.

By 2010 almost all telephone sampling methods had switched to a *dual frame* approach because of substantial undercoverage of the landline only sampling frame. For example, for the time period January-June of 2014, an estimated 44% of U.S. households are wireless only, and an estimated 43% of adults and 52% of children live in wireless only households (Blumberg and Luke 2014). Dual frame RDD surveys use a landline frame as described above, but supplemented with a cell phone frame that contains all possible cell telephone numbers for the defined geographic area.

Although dual frame RDD may sound like an easy fix to the current major problems of landline only RDD, challenges remain:

- A landline phone goes to a housing or dwelling unit, like APS, whereas a cell phone goes to a person; this difference impacts how selection probabilities for sampled elements are calculated.
- Automatic dialers are used for landline numbers but current law prohibits automatic dialers being used for cell phone numbers.
- A completed interview for a cell phone number is more expensive than for a landline number.
- Response rates for the cell phone frame typically are lower than for the landline frame, but both frames suffer from declining response rates.
- For a statewide survey with the target population being household state residents, generating the cell phone frame from all area code and prefix combinations within the state will not cover persons residing in the state who had their cell phone number issued to them in a different state. This is very problematic for certain areas of the United States.

A national survey that uses dual frame RDD is the National Immunization Survey (NIS) conducted by NCHS where the target population is children aged 18-35 months residing in households in the United States (National Center for Health Statistics, 2015E). A statewide survey that uses dual frame RDD is BRFSS (Behavioral Risk Factor Surveillance System), sponsored by CDC, where the target population is adults who reside in households within the state (Centers for Disease Control, 2015B). However, since all states, plus the District of Columbia (DC), conduct this survey and ask many of the same questions each year, states swap completed interviews with each other when one state calls a cell number with that state's area code and prefix, but the selected person now resides in a different state.

For dual frame RDD, the PSU is a telephone number, and it takes either one or two stages of sampling to get to the element.

RESIDENTIAL ADDRESS BASED SAMPLING (ABS)

Over the past decade the use of residential address based sampling in the United States has increased dramatically (Dillman et al 2014). The target population typically is persons living in households within a given geographic area (nation, state, or local area). The sampling frame is a list of residential addresses for the given geographic area. This list is based on the CDSF (computerized delivery sequence file) that the U.S. Bureau of the Census develops and uses for the mail portion of the decennial census of housing and population. Additional information for the sampling frame is provided by the emergency 911 system that has been assigning addresses and GPS coordinates to rural housing units over the past several years. The first stage sampling frame of addresses typically is stratified by geography and perhaps by other characteristics based on census data.

Once an address is selected into the sample, contact with the housing unit typically is by mail or by a landline telephone number found in a reverse directory. Sampling within the household to select one or more persons into the sample can be challenging via mail, but is being operationalized in several surveys. Data collection for persons selected into the sample is via a self-administered questionnaire, telephone interview or perhaps a WEB site.

In ABS the PSU is an address, and it takes either one or two stages of sampling to get to the element (person).

OVERVIEW OF STATISTICAL PROCEDURES FOR COMPLEX SURVEY DATA

Standard statistical procedures typically assume that the data were obtained via simple random sampling and, hence, do not take into account three common characteristics of complex survey data:

- Unequal probability selection of sampled elements from the target population, partially due to oversampling of specific domains.
- Clustering or statistical dependence of some observations, often due to multi-stage sampling.
- Stratification of the first stage, as well as possibly subsequent, sampling frame(s) prior to sampling.

Many papers have shown that analyzing complex survey data using standard statistical procedures yields biased estimators of target population parameters and underestimated standard errors (e.g. Brogan, 2015); this paper does not replicate these findings.

ESTIMATION OF TARGET POPULATION PARAMETERS

In terms of estimating target population parameters, unequal selection probabilities is addressed easily in survey software by conducting a weighted analysis. Estimators are a function of values of variables for observations in the dataset, weighted by an appropriate sampling weight variable. The value of the sampling weight variable for a given observation is the number of elements in the target population represented by that observation. Sometimes the terms expansion or inflation sampling weight variable are used.

For example, if it is desired to estimate the number of adults in the target population who have arthritis, i.e. Y as defined earlier in Equation (1), using the "sample adult" dataset from 2013 NHIS, then the estimator formula used by the survey software (assuming no item nonresponse) is

$$\hat{Y} = \sum_{k=1}^{k=34557} w_k y_k \qquad (3)$$

where:

- $y_k = 1$ (or 0) if observation k in the dataset has (or does not have) arthritis.
- w_k is the value of the sampling weight variable for observation k.
- The summation is over the 34557 observations in the "sample adult" dataset.

If it is desired to estimate the percentage of adults in the target population who have arthritis, i.e. P as defined earlier in Equation (2), then the estimation formula used by the survey software (assuming no item nonresponse) is

$$\hat{P} = 100\left(\sum_{k=1}^{k=34557} w_k y_k\right) / \left(\sum_{k=1}^{k=34557} w_k\right) \qquad (4)$$

Note that analyzing complex survey data without weighting by the appropriate sampling weight variable yields biased estimators of target population parameters.

The initial or base value of the sampling weight variable for a given observation is the inverse of its selection probability as determined by features of the sampling plan such as stratification, oversampling, and clustering. The base value typically is modified by one or more adjustments for nonresponse at possibly multiple stages of sampling. The final adjustment is post-stratification to a known distribution of demographic variables in the target population, based on the U.S. decennial census of population and housing, supplemented with postcensal estimates. Only one sampling weight variable can appear on the WEIGHT statement in any of the SAS/STAT SURVEY procedures.

The calculation of values for one or more sampling weight variables typically is done by the agency and/or contractor that releases the complex survey dataset(s) to the public. The survey documentation should describe how the weighting was done. There are a few instances where the data analyst may need to do minor calculations to obtain the value of an appropriate sampling weight variable, e.g. for NHANES where survey years may be combined to increase sample size and/or survey ddf (denominator degrees of freedom).

The survey ddf for a given complex sample survey is defined as (number of PSUs in the sample - # of first stage strata in the sample). For the 2013 NHIS the survey ddf is 300, i.e. exactly 2 PSUs within each of 300 first stage strata. The generally recommended minimum value for the survey ddf is 30.

If only point estimates of target population parameters are desired, with no measure of sampling variability for the point estimates, then SAS/STAT standard statistical software such as the FREQ

procedure or the MEANS procedure can be used by adding a WEIGHT statement to the program. However, any calculations of estimated variance for point estimates, or quantities that rely on estimated variance, will not be correct because PROC FREQ and PROC MEANS do not account for potential clustering of observations and for stratification in the sampling plan.

ESTIMATED VARIANCE FOR ESTIMATORS

Variance estimation for estimators based on complex survey data is not straightforward for two reasons:

- When the sampling plan is more complicated than one stage sampling plans like stratified random sampling or stratified one-stage cluster sampling, it is difficult or impossible to obtain a closed form algebraic expression for the estimated variance.
- Many estimators of target population parameters are ratio estimators where both the numerator and denominator have sampling variability.

To address the first issue, i.e. multi-stage sampling, variance estimation typically makes no attempt to keep track of all stages of the sampling plan. Rather, the actual sampling plan is approximated by the simpler plan of stratified one-stage cluster sampling, but *only* for the purpose of variance estimation. That is, all of the observations in the dataset that are within a given PSU within a given first stage stratum are assumed to have been selected via one-stage cluster sampling rather than from the multi-stage sampling plan that actually was used. This approximation, known as the UCVE or ultimate cluster variance estimate (Brogan 2005), has been shown over many years to be acceptable with perhaps a slight overestimate.

Some software packages for complex survey data analysis allow the data analyst to describe the multiple stages of stratification and sampling in stratified multi-stage sampling; SAS/STAT is not one of these. However, the utility of this high-end feature is limited since most public release datasets do not include the many stage-specific survey design variables that would be needed to use this feature—for confidentiality reasons.

To address the second issue, i.e. ratio estimators, two common approaches in statistics are used: Taylor Series Linearization (TSL) and replication methods such as balanced repeated replication (BRR) and jackknife (Brogan 2005). Whether TSL or replication methods are used for approximate variance estimation, the estimated variance formulas are impacted by clustering, stratification, and weighting. In general, these different variance estimation approximations for ratio estimators yield comparable results.

The SAS/STAT SURVEY procedures have options for TSL, BRR or jackknife variance estimation. Secondary data analysts should note in the survey documentation whether the publically released dataset is set up for TSL, BRR or jackknife variance estimation. Most publically released complex survey datasets are set up for TSL. Unless there are compelling reasons to do otherwise, the variance estimation method recommended by the survey documentation should be used.

DESCRIBE SAMPLING PLAN AND CHOOSE VARIANCE ESTIMATION METHOD

Each SAS/STAT SURVEY procedure needs to know the actual or approximate sampling plan for the input dataset so that it knows what formulas to use for estimators and for variance estimation. The sampling plan is described by options on the PROC statement and additional statements like STRATA, CLUSTER, WEIGHT and REPWEIGHTS.

TAYLOR SERIES LINEARIZATION

If the complex survey dataset is set up for TSL variance estimation, then the option VARMETHOD = TAYLOR goes on the PROC statement. In addition, the following survey design variables should be explained in the documentation and appear in the dataset:

- First stage stratification variable used for the PSUs, e.g. a variable named STRATVAR. This variable goes on the STRATA statement. Although rare, the first stage stratification variable could be defined as a cross-classification of two variables, e.g. named STRATV1 and STRATV2. In this case, these two variables go on the STRATA statement.

- A PSU variable if elements within the sample are clustered within primary sampling units, e.g. named PSUVAR. This variable goes on the CLUSTER statement. SAS always reads cluster variable codes for each observation *within* the first stage stratum code(s) for that observation. Although rare, the primary sampling units also could be defined as a cross-classification of two variables, e.g. named PSUV1 and PSUV2. In this case these two variables go on the CLUSTER statement.
- One or more sampling weight variables. There may be more than one sampling weight variable in the dataset to use, depending upon what variables are being analyzed (NHANES is an example), but only **one** sampling weight variable is used on the WEIGHT statement in a given SAS/STAT SURVEY procedure.

For the 2013 NHIS "sample adult" dataset, the sampling plan and TSL variance estimation is described to the SAS SURVEY procs as follows:

```
proc surveyXXXX  varmethod = taylor  data = nhis      ;
strata   strat_p  ;     cluster  psu_p   ;    weight   WTFA_SA  ;
```

The survey documentation or exploration of the 2013 NHIS sample adult dataset shows that the variable strat_p is coded from 1 through 300 (300 first stage strata) with the variable psu_p coded as 1 or 2 within each stratum. The letter p which is part of the stratum and PSU variable names stands for "pseudo". In fact, there are 300 "pseudo" first stage strata and two "pseudo" PSUs within each pseudo stratum. The word "pseudo" is used because the variables strat_p and psu_p may not correspond exactly with the actual strata and PSUs used in the sampling plan, but the variables strat_p and psu_p are designed and coded for TSL variance estimation purposes.

The syntax above for using TSL with the 2013 NHIS is a common syntax for many surveys that use TSL, with different variable names of course for the survey design variables.

There is a rarely used option on the PROC statement to incorporate a finite population correction (fpc) factor into variance estimation when first stage sampling fractions are high; see the SAS/STAT documentation for details.

There are two sampling weight variables in the "sample adult" dataset: WTFA_SA and WTIA_SA. WTFA_SA is the *final* (all adjustments done, including post-stratification) annual sampling weight variable recommended for data analysts to use. WTIA_SA is the *interim* value of the sampling weight variable, after adjustments have been done for nonresponse but before the post-stratification adjustment. The sampling weight variable WTIA_SA may be useful for experienced data analysts who wish to re-post-stratify or estimate variance using methods that differ from what is discussed in this paper.

REPLICATION METHODS: BRR AND JACKKNIFE

The PROC statement for any SAS/STAT SURVEY procedure includes the option VARMETHOD = BRR or the option VARMETHOD = JACKKNIFE (or JK). If the dataset is set up for BRR or JACKKNIFE variance estimation, then the following survey design variables should be explained in the documentation and appear in the dataset:

- A sampling weight variable to be used to calculate point estimates of specified target population parameters. One sampling weight variable goes on the WEIGHT statement.
- A set of replicate weight variables that define the several replicates to be used for variance estimation. All of the replicate weight variables go on the REPWEIGHTS statement.

For example, 52 replicate weight variables for BRR might be named BRRWgt01-BRRWgt52 or 48 replicate weight variables for JACKKNIFE might be named JKKWgt01-JKKWgt48. Consider the replicate weight variable BRRWgt01; it defines how replicate 01 is constructed. All observations in the dataset that are *not* included in replicate 01 have the value zero for the variable BRRWgt01. All remaining observations that are in replicate 01 have a positive value for the variable BRRWgt01, each positive value being a reweighting so that these remaining observations make inference to the entire target population. The software estimates the target population parameter(s) using BRRWgt01, yielding one or more point estimates from replicate 01. The process is repeated for replicate 02, and so on.

There are additional options that likely need to be specified in the SAS/STAT SURVEY PROC for either BRR or JACKKNIFE; see the survey documentation and the SAS documentation for details.

SOME CODING ISSUES

Many surveys top code or bottom code certain items or variables to protect confidentiality of responses. In the 2013 NHIS "sample adult" dataset, the variable AGE_P is coded in years for 18 through 84. However, 1014 observations are top coded as AGE_P = 85, meaning 85 or older. This limits using age as a continuous variable unless assumptions are made and also limits analyses by age for subpopulations such as adults aged 80 years or older. In other instances a value of a variable for an observation may be coded as "not available", meaning that the agency chose to not release the value to the public, again to protect confidentiality. The NCHS has a Research Data Center (RDC) where researchers can apply for access to data not released to the public.

Item nonresponse typically is coded numerically in public release datasets and frequently for the reason, e.g. refuse or don't know. A different code, sometimes numeric, often is used to indicate that the item was *skipped* because of the skip pattern; this is *not* item nonresponse and is sometimes called *planned missing.*

Data analysts should code the numeric codes for item nonresponse, not available, and skip to something like .D (don't know). .R (refuse), .S (skip) and .N (not available) prior to analysis. However, there may be variables where you wish to consider "don't know" or "refuse" as a valid answer to the item among other answers such as yes or no.

CHOOSE HOW TO HANDLE ITEM NONRESPONSE IN ANALYSES

Even if you analyze only one variable, it likely has item nonresponse coded by you as missing using the SAS convention. If you do not choose how to deal with item nonresponse in your analyses, SAS chooses for you. Consider the arthritis variable in the NHIS sample adult dataset. It was asked of all sample elements, and it is desired to estimate the prevalence of arthritis in the target population.

The SAS choice, which could also be your choice, is to assume that item nonresponse is MCAR (missing completely at random). That is, each observation who chose not to respond to the arthritis item did so statistically independently of what the answer would have been if the observation had chosen to answer. With this assumption, SAS deletes the item nonresponse observations from the input dataset and analyses the remaining observations, all of whom have a valid value for the arthritis variable. The MCAR assumption allows the estimated arthritis prevalence to make inference back to the target population.

A common method to deal with item nonresponse in other software packages that analyze complex survey data is to define a subpopulation as all elements in the target population who would answer the item, if asked. Then the software does a *subpopulation analysis* and estimates arthritis prevalence for this defined subpopulation. Although the point estimate of the parameter of interest, arthritis prevalence, is the same with both approaches, i.e. MCAR assumption and subpopulation analysis, the two approaches differ on:

- The inference population (target population versus subpopulation who would respond to item)
- Estimated standard error of the point estimate

The estimated standard errors differ for the two approaches because, with the subpopulation approach, observations who are *not* in the subpopulation (i.e. did not respond to the item when they should have), contribute to variance estimation. In the MCAR approach, these observations do *not* contribute to variance estimation because they are no longer in the dataset that is being analyzed. In general, the estimated variance seems to be somewhat larger with the subpopulation approach compared to the MCAR approach, although the difference usually is not detectable for analyzed items with very small item nonresponse rates.

SAS offers both of the methods discussed above. To choose the default MCAR, do nothing. To choose the subpopulation approach, add the option NOMCAR to the PROC statement.

A third method for handling item nonresponse is to do imputation or multiple imputation for item nonresponse. These methods are not discussed here. As an aside, the variable income, no matter how asked, has a consistently high item nonresponse rate over many different surveys. NCHS offers NHIS datasets with multiple imputation for income and instructions for their use (National Center for Health Statistics, 2015F).

The author has developed her own method for handling item nonresponse when analyzing one or more variables together. First, always use NOMCAR. Second, if the item nonresponse rate is low on all variables in the analysis, and, in addition, there is no obvious reason to question the MCAR assumption for each variable, then use the point estimate obtained from the subpopulation approach to make inference to the target population. Otherwise, state the inference population as those in the target population who would answer all of the items in the analysis, if asked. If the item nonresponse rate for a particular item is quite high, it may be better to avoid analyzing that item.

The issue of item nonresponse may become increasingly difficult as the number of variables in the analysis increases. An example is fitting a logistic regression model to complex survey data using the SURVEYLOGISTIC procedure with several independent variables, using NOMCAR. The only observations that are used in fitting the model are those who have a nonmissing value for every variable in the analysis. Some researchers have mentioned losing half of the sample for their chosen model, although it is crucial to check out whether *all* of this sample loss is due to item nonresponse or whether some of it is due to skip patterns or planned missing. In general, the inference population when fitting statistical models with several variables may need to be restricted to elements in the target population (or appropriate subpopulation) who would answer all of the items in the analysis, if asked.

DOMAIN AND SUBPOPULATION ANALYSES

Domain variables are used when, as in epidemiology, you wish to do an analysis “stratified by” another variable. For example, to estimate the arthritis prevalence for males and females in the target population, the analysis of the dichotomous arthritis variable is “stratified” by the domain variable sex, which forms the two domains males and females. The SAS/STAT SURVEY procedures do an analysis of the arthritis variable, first for the subpopulation of males only and then for the subpopulation of females only. The domain variable, such as sex, goes on the DOMAIN statement in all SAS/STAT SURVEY procedures for data analysis except PROC SURVEYFREQ where it goes on the TABLES statement.

The word *subpopulation* means a subset of elements from the target population, and data analysts often are interested in subpopulations. For example, questions about PSA testing are analyzed for the subpopulation of older males only, and questions regarding pregnancy are analyzed for the subpopulation of reproductive age females only. Note the difference here between the subpopulation of interest being older males only and the above situation where both subpopulations, males and females, were of interest when sex was a domain variable.

It may seem intuitive to subset the dataset to older males for PSA analyses and to reproductive age females for analysis of pregnancy questions. However, warnings abound not to do this. The correct method is to conduct a subpopulation analysis so that the dataset observations that do *not* belong to the subpopulation contribute to the estimated variance for the point estimate that is based on subpopulation observations only. Subsetting the dataset to observations only in the subpopulation often leads to an underestimated variance for a subpopulation point estimate.

Most software packages for complex survey data analysis have a subpopulation statement that allows the data analyst to define a subpopulation of interest with syntax like sex = female and age > 15 and age < 50. The SAS/STAT SURVEY procedures do not offer this capability to users, although they certainly have the capability to conduct subpopulation analyses as indicated by how they conduct domain analyses.

The method for conducting subpopulation analyses, as recommended by SAS, is as follows:

- Form an indicator variable SUBIND coded 1 if observation belongs to subpopulation, coded 0 if observation does not belong, and coded .M if cannot discern whether observation belongs to subpopulation

- Conduct an analysis in one of the SAS/STAT SURVEY procedures with SUBIND as a DOMAIN variable or SUBIND on the TABLES statement in PROC SURVEYFREQ, i.e. do the analysis for both levels of the variable SUBIND
- Ignore the analysis results for SUBIND = 0; the results of interest are for SUBIND = 1.
- Use NOMCAR on the PROC statement.

Although this method works for subpopulation analyses, it is less than desirable because you may get two to three times the output that you wish.

EXAMPLE 1: USE PROC SURVEYFREQ TO ESTIMATE NUMBER AND PERCENTAGE OF ADULTS IN 2013 NHIS TARGET POPULATION WHO HAVE ARTHRITIS

The PROC SURVEYFREQ program below estimates the two population parameters specified in equations (1) and (2) for the 2013 NHIS target population, i.e. number and percentage of adults with arthritis. All 34557 observations in the "sample adult" dataset were asked the arthritis question, the arthritis variable arth1db is coded 1=yes and 0=no, and 51 observations have a missing value for arth1db, for an item nonresponse rate of 0.15% (very low). The option CL on the TABLES statement below requests default 95% Wald confidence limits (symmetrical around the point estimate) for the percentage parameter:

```
PROC SURVEYFREQ   data = nhis   varmethod = taylor   NOMCAR  ;
strata  strat_p   ;    cluster  psu_p   ;   weight WtFa_SA ;
tables    arth1db  /   cl   ;
title "Example 1 SURVEYFREQ  arth1DB  2013 NHIS  Sample Adult" ;
run ;
```

Data Summary	
Number of Strata	300
Number of Clusters	600
Number of Observations	34557
Sum of Weights	237394354

Variance Estimation	
Method	Taylor Series
Missing Values	NOMCAR

Output 1. Data Summary Table and Variance Estimation Table for Example 1

Output 1 above gives two default tables for the SURVEYFREQ program above for Example 1. The Data Summary table indicates that SURVEYFREQ found 300 first stage strata (i.e. 300 different values for the variable strat_p), a total of 600 different PSUs or clusters *within* these 300 strata, and 34557 observations. SURVEYFREQ adds up the value of the sampling weight variable WtFa_SA over all observations in the dataset and obtains the number 237,394,354, which is the estimated number of elements in the 2013 NHIS target population, based on the post-stratification procedure. You should know all of the information in the Data Summary from preliminary exploration of the dataset before analysis and confirm that the figures in the Data Summary are correct. The Variance Estimation table reminds you that you chose TSL for variance estimation and that you requested the option NOMCAR to deal with item nonresponse. Every time that a SAS/STAT SURVEY procedure reads in this same sample

adult dataset with the same specifications for variance estimation and item nonresponse, the Data Summary and Variance Estimation tables will be exactly as seen above.

DXArth							
arth1DB	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	95% Confidence Limits for Percent	
0 No	25962	183160367	1752834	77.2689	0.3169	76.6453	77.8926
1 Yes	8544	53882301	912050	22.7311	0.3169	22.1074	23.3547
Total	34506	237042668	2103121	100.000			
Frequency Missing = 51							

Output 2. Arthritis Analysis Results Table for Example 1

Output 2 above presents the arthritis analysis results for Example 1. PROC SURVEYFREQ finds that 51 observations do not have a value for the variable arth1DB, which you also should already know from preliminary exploration of the dataset. PROC SURVEYFREQ finds 8544 observations in the dataset who have arthritis (Frequency column). When PROC SURVEYFREQ adds up the value of the sampling weight variable WtFa_SA for these 8544 observations, it obtains the number 53,882,301 (Weighted Frequency column). In essence, the 8544 observations in the sample who have arthritis represent an estimated 53.9 million adults in the target population who have arthritis. The estimated standard error for the point estimate 53.9 million is 912,050 (Std Dev of Wgt Freq column). PROC SURVEYFREQ will provide corresponding confidence limits by specifying the option CLWT on the TABLES statement.

The 25962 observations in the sample who do not have arthritis represent an estimated 183,160,367 adults in the target population who do not have arthritis. The 34506 observations in the results table above represent an estimated 237,042,668 adults in the target population. Note that this number 237,042,668 is slightly less than the 237,394,354 figure in the Data Summary table because 51 observations in the sample are not in the table of Output 2.

From Output 2 above, the estimated percentage of adults who have arthritis is 22.73% (Percent column) with estimated standard error of 0.3169% (Std Err of Percent column). Confidence limits (default 95% and default Wald) are (22.11%, 23.35%).

How does PROC SURVEYFREQ calculate the estimated 22.73% in Output 2? The numerator is the estimated 53,882,301 adults in the target population who have arthritis, and the denominator is the estimated number of adults in the target population, i.e. 237,042,668. Data analysts should check percentage calculations in output tables to make sure that they understand how estimated percentages are calculated by PROC SURVEYFREQ.

Now, to what population do the Example 1 results in Output 2 make inference? The answer to this question determines how you present your results to an audience or in your journal article.

Since I used the option NOMCAR in PROC SURVEYFREQ, the Example 1 results technically make inference to a *subpopulation* of the 2013 NHIS target population, where the subpopulation is defined as: civilian noninstitutionalized adults (18+) residing in a dwelling unit in the United States in 2013 *who would answer an item about having arthritis, if asked.* The estimated size of this subpopulation is 237,042,668 adults from Output 1 above. Thus, my results are stated as follows:

- The estimated number of adults with arthritis in the defined subpopulation immediately above is 53.9 million with estimated standard error of 912,000.
- The estimated prevalence of arthritis in this same subpopulation is 22.73%, with estimated standard error of 0.3169% and a 95% Wald confidence interval of (22.11%, 23.35%).

However, if now I make the assumption of MCAR, since the item nonresponse rate for the arthritis variable is so low, then my inference population becomes the 2013 NHIS target population and my results can be stated as follows:

- The estimated number of adults with arthritis in the 2013 NHIS target population is 53.9 million with estimated standard error of 912,000.
- The estimated prevalence of arthritis in the 2013 NHIS target population is 22.73%, with estimated standard error of 0.3169% and a 95% Wald confidence interval of (22.11%, 23.35%).

So, why don't I just assume MCAR from the beginning for this analysis and forget about using NOMCAR? The answer is that I, and other sample survey statisticians as well, prefer the variance estimation calculations given by the NOMCAR option because in some instances they are slightly more conservative. In Example 1 here, though, the MCAR and the NOMCAR options both give essentially the same estimated variances because the item nonresponse rate is so low.

For making inference to the 2013 NHIS target population, note that the point estimate 53,882,301 may be an underestimate of the parameter Y defined in Equation (1) because 51 observations have a missing value for the variable arth1DB. If any of these 51 observations actually has arthritis, then the point estimate of Y would be larger than 53,882,301. This issue is of no practical concern here for two reasons:

- The item nonresponse rate is quite small, 0.15%, and any increase in the point estimate is unlikely to be substantial
- The point estimate will be rounded off anyway to something like 53.9 million or 54 million rather than reporting the point estimate to the nearest person, likely resulting in the same or a similar point estimate that would have been obtained with no item nonresponse

If you are bothered by this possible underestimation, you can obtain an alternate point estimate by multiplying the target population size of 237,394,354 by the estimated proportion 0.227311.

PROC SURVEYFREQ has several options for confidence limits on population percentages, including the logit method and alternate methods for population percentages that are assumed to be small or large. See the PROC SURVEYFREQ documentation for details.

HOW NOT TO STATE THE RESULTS OF EXAMPLE 1

Below are some incorrect statements I have seen in published journal articles that report analyses of complex survey data. Do not make statements like:

- Because the sample was biased, we had to fix (repair) it by doing a weighted analysis.
- The prevalence of arthritis in the target population is 22.7%.
- The prevalence of arthritis among sample respondents is 22.7%.
- The prevalence of arthritis in the weighted sample is 22.7%.

EXAMPLE 2: USE PROC SURVEYMEANS TO ESTIMATE NUMBER AND PROPORTION OF ADULTS IN 2013 NHIS TARGET POPULATION WHO HAVE ARTHRITIS

PROC SURVEYMEANS can be used to do the same analysis done by PROC SURVEYFREQ in Example 1. The variable arth1DB can be analyzed as either continuous or categorical in SURVEYMEANS since it is coded 1=yes and 0=no. The target population parameters of interest are the number and proportion of adults in the target population who have arthritis. There are many options on the PROC SURVEYMEANS statement for statistics to be output; see the PROC SURVEYMEANS documentation.

The PROC SURVEYMEANS syntax for a comparable analysis to Example 1, using the variable arth1db as continuous, is:

```
proc SURVEYMEANS   data = nhis   varmethod = taylor   NoMcar
      nobs   nmiss    mean  clm   sum   plots = none  ;
strata  strat_p   ;    cluster  psu_p   ;   weight WtFa_SA ;
```

```
var   arth1db  ;  /* considered continuous */
/* Suppress default plot for continuous variable on PROC statement since
arth1db not really continuous */
title "Example 2A SURVEYMEANS Arth1DB(Continuous) 2013 NHIS Sample Adult" ;
run ;
```

The SURVEYMEANS syntax for a comparable analysis to Example 1, using the variable arth1db as categorical, is:

```
proc SURVEYMEANS   data = nhis   varmethod = taylor   NoMcar
      nobs     mean  clm   sum  ;
strata  strat_p   ;    cluster  psu_p   ;   weight WtFa_SA ;
var   arth1db  ;
class  arth1db  ;  /* consider variable on VAR statement as categorical */
title "Example 2B SURVEYMEANS Arth1DB(Categorical) 2013 NHIS Sample Adult";
run ;
```

The analysis results for Example 2 from PROC SURVEYMEANS, not presented here, are exactly the same as in Example 1 but organized differently from the PROC SURVEYFREQ output. The previous discussion in Example 1 regarding inference population to use applies to the Example 2 results as well.

EXAMPLE 3: USE PROC SURVEYFREQ TO ESTIMATE NUMBER AND PERCENTAGE OF ADULTS IN 2013 NHIS TARGET POPULATION WHO HAVE ARTHRITIS, BY SEX

The variable sex (coded 1 = male and 2 = female) is a domain variable in this analysis since results are needed for each of the two domains males and females. The domain variable sex goes on the TABLES statement in the PROC SURVEYFREQ program below and is cross-tabulated with the dependent variable Arth1db. By default PROC SURVEYFREQ outputs estimated cell percents, i.e. estimated percentage of adults in target population who are in each of the 4 cells of the 2-way table sex * arth1db. Cell percents are not of interest in this analysis. The option NOCELLPERCENT on the TABLES statement below stops PROC SURVEYFREQ from outputting cell percents.

Since sex is the *row* variable on the TABLES statement below, row percents are of interest, and they are requested as an option on the TABLES statement. The item nonresponse rate in this example is the same as in Examples 1 and 2, i.e. 0.15% for the variable arth1db, because the variable sex has no missing values.

```
proc SURVEYFREQ   data = nhis   varmethod = taylor   NoMcar;
strata  strat_p   ;    cluster  psu_p   ;   weight WtFa_SA ;
tables  sex *  arth1db  /   row  nocellpercent   ;
title "Example 3 SURVEYFREQ  Sex * Arth1DB  2013 NHIS  Sample Adult" ;
run ;
```

Output 3 below is from the Example 3 PROC SURVEYFREQ program above. The Data Summary and Variance Estimation boxes in the Example 3 output are the same as in Output 1 for Example 1 and are not repeated here. Based on comments above for Example 1, I interpret the results in Output 3 to make inference to the 2013 NHIS target population even though I used NOMCAR on the PROC statement.

The number of sample males in the analysis is 15417, of whom 3137 have arthritis. These 3137 arthritic males in the sample represent an estimated 21,464,362 arthritic males in the target population. That is, the estimated number of males in the target population who have arthritis is 21,464,362 (perhaps a slight underestimate as discussed earlier in Example 1), with estimated standard error of 521,940. The estimated prevalence of arthritis among males in the target population is 18.81% (calculated as 100 x [21,464,362] / [114,140,403]), with estimated standard error of 0.4044%. The interpretation of the results for the female domain in the target population is similar, with an estimated arthritis prevalence of 26.38% for females in the target population.

Table of SEX by arth1DB						
SEX	arth1DB	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Row Percent	Std Err of Row Percent
1 Male	0 No	12280	92676041	1208139	81.1948	0.4044
	1 Yes	**3137**	**21464362**	**521940**	**18.8052**	**0.4044**
	Total	15417	114140403	1360531	100.000	
2 Female	0 No	13682	90484326	1098019	73.6230	0.4338
	1 Yes	**5407**	**32417939**	**676110**	**26.3770**	**0.4338**
	Total	19089	122902265	1381003	100.000	
Total	0 No	25962	183160367	1752834		
	1 Yes	8544	53882301	912050		
	Total	34506	237042668	2103121		
Frequency Missing = 51						

Output 3. Arthritis Analysis Results Table, by Sex, for Example 3

EXAMPLE 4: USE PROC SURVEYMEANS TO ESTIMATE NUMBER AND PERCENTAGE OF ADULTS IN 2013 NHIS TARGET POPULATION WHO HAVE ARTHRITIS, BY SEX

The PROC SURVEYMEANS program below for Example 4 considers the variable arth1db to be continuous. The DOMAIN statement below requests that the analysis be done for each of the two domains males and females. The output for this program is not shown here because the results are comparable to Output 3 above where PROC SURVEYFREQ was used. The comparable PROC SURVEYMEANS program for treating arth1db as categorical is straightforward and not included here.

```
proc SURVEYMEANS   data = nhis   varmethod = taylor   NoMcar
      nobs   nmiss    mean  clm   sum   plots = none  ;
strata  strat_p   ;    cluster  psu_p   ;   weight WtFa_SA ;
var   arth1db  ;  /* considered continuous */
domain   sex   ;
/* Suppress default plot for continuous variable on PROC statement since
arth1db not really continuous */
title "Example 4 SURVEYMEANS Arth1DB(Continuous) By Sex  2013 NHIS" ;
run ;
```

EXAMPLE 5: USE PROC SURVEYFREQ TO DETERMINE IF TWO CATEGORICAL VARIABLES ARE STATISTICALLY INDEPENDENT

Examples 1 through 4 above illustrate *descriptive* analyses of complex survey data. Descriptive analyses of health survey data provide continuing surveillance of the health status of U.S. residents and their use of health care facilities and providers. Annual health surveys make it possible to investigate trends over time in health status and health care utilization.

Survey data also can be used to investigate *relationships* between variables, e.g. to determine whether two categorical variables are statistically independent or not. The Pearson chi-square test (observed – expected) is a favorite statistical procedure for this purpose, but it is for data obtained via SRS. Sample survey statisticians modified the Pearson chi-square test, and similar tests, for use with complex survey data. Because all modifications are approximations, and because different researchers suggest different approximations, there are up to 8 options in PROC SURVEYFREQ to test the null hypothesis that two

categorical variables are statistically independent. Read the PROC SURVEYFREQ documentation for details. The most commonly used option seems to be CHISQ, which Rao and Scott developed by using a weighted Pearson test with a design correction because complex survey data are used.

In the PROC SURVEYFREQ program below there are two TABLES statements. The first TABLES statement is for sex * arth1db; the CHISQ test is requested as an option and the usual table results will *not* be printed because these results were obtained earlier in Output 3 above. The second TABLES statement is for the race-ethnicity variable racethdb2 crossed with arth1db.

```
proc SURVEYFREQ   data = nhis   varmethod = taylor   NoMcar  ;
strata  strat_p   ;    cluster  psu_p   ;   weight WtFa_SA ;
tables   sex * arth1db  /  chisq  noprint  ;
tables  racethdb2 * arth1db  /  row  nocellpercent  chisq    ;
 title "Example 5 SURVEYFREQ CHISQ Test  2013 NHIS Sample Adult" ;
 run ;
```

Output 4 below presents the Rao-Scott Chi-Square Test results for the 2-way table sex * arth1db. The weighted Pearson chi-square calculated statistic is 281.1980. This figure, though, is too large because no recognition is taken of the complex survey design. The design correction calculation is 1.5399, which is divided into the 281.1980, resulting in the Rao-Scott chi-square statistic of 182.6105. This statistic approximately follows the chi-square distribution with 1 df (based on the 2 x 2 table sex * arth1db). Assuming that the null hypothesis of statistical independence is true, the probability (Pr) of obtaining a value of the chi-square statistic as large as 182.6105 or larger is <.0001.

The chi-square statistic of 182.6105 is converted into an F-statistic by dividing the chi-square statistic by the df for the chi-square distribution, i.e. 1.0. The value of the F-statistic is 182.6105, and it approximately follows the F-distribution with numerator df = 1 and denominator df = 300 (the survey ddf). Assuming that the null hypothesis of statistical independence is true, the probability (Pr) of obtaining a value of the F-statistic as large as 182.6105 or larger is < .0001.

Most sample survey statisticians use the F-statistic to decide whether or not to reject the null hypothesis because it is more conservative than using the Rao-Scott chi-square statistic. Using the F-statistic is especially important when the survey ddf is smaller (but hopefully still at least 30).

Rao-Scott Chi-Square Test	
Pearson Chi-Square	281.1980
Design Correction	1.5399
Rao-Scott Chi-Square	182.6105
DF	1
Pr > ChiSq	<.0001
F Value	182.6105
Num DF	1
Den DF	300
Pr > F	<.0001
Sample Size = 34557	

Output 4. CHISQ Results for the 2-Way Table sex*arth1db in Example 5

Based on the F-statistic, the null hypothesis of statistical independence between sex and arthritis is rejected. What is the conclusion? First, my conclusion makes inference to the 2013 NHIS target population for reasons explained earlier. Second, rejecting the null hypothesis means that males and females in this target population differ on the prevalence of arthritis. Third, since there are only two sex domains, and since the estimated prevalence of arthritis is 18.80% for males and 26.38% for females, I further conclude that, in the target population, females have a higher prevalence of arthritis than do males.

The output for the relationship between race/ethnicity and arthritis is not presented here, but is briefly summarized. First, there are 259 observations missing from the 2-way table racethdb2 * arth1db, most of whom have a missing value for the variable racethdb2 and 51 of whom have a missing value for the variable arth1db. The item nonresponse rate for this 2-way table is 259/34557 = 0.75%, still quite small. Thus, I use the option NOMCAR but make inference to the 2013 NHIS target population, as explained earlier.

Second, the estimated prevalence of arthritis for the four race/ethnicity domains in the target population is as follows: 26.3% for nonHispanic whites, 21.6% for nonHispanic blacks, 13.7% for nonHispanic others, and 11.8% for Hispanics. The requested CHISQ test has a p-value < .0001 for the F-statistic. The null hypothesis of statistical independence of race/ethnicity and arthritis is rejected.

Third, so what is the conclusion? Rejection of the null hypothesis allows me to say that the four race/ethnicity domains in the target population do *not* have the same prevalence of arthritis. However, the CHISQ test, or any other of the chi-square test options in SURVEYFREQ, do not allow me to say anything about which domains differ from other domains following rejection of the null hypothesis.

FOLLOWUP TO A CHI-SQUARE TEST THAT REJECTS STATISTICAL INDEPENDENCE

If your 2-way table for testing statistical independence is larger than 2 x 2, then the null hypothesis of statistical independence can be false in different ways. The race/ethnicity and arthritis relationship of Example 5 is used here as an example.

If you wish to compare the four race/ethnicity domains on prevalence of arthritis, e.g. via specific paired comparisons or more general linear contrasts of race/ethnicity domains on prevalence, you can use PROC SURVEYREG. Recall that the variable arth1db is coded as 1=yes and 0 = no. Fit a cell means model (no intercept) with arth1db as the dependent variable (assumed continuous) and racethdb2 as the independent categorical variable at four levels. The four estimated regression coefficients from PROC SURVEYREG are the estimated *proportion* of adults who have arthritis for each of the four race/ethnicity domains in the target population. Use the CONTRAST or ESTIMATE statements in PROC SURVEYREG to form linear combinations of the four estimated proportions of interest to you. This procedure implemented in PROC SURVEYREG is preferred by most statisticians over the procedure of comparing arthritis prevalence confidence intervals among the four race/ethnicity domains to see if the confidence intervals overlap.

EPIDEMIOLOGICAL MEASURES OF RELATIONSHIP STRENGTH IN PROC SURVEYFREQ

Epidemiologists frequently work with dichotomous outcome variables, i.e. have disease versus not. Further, they often use 2 x 2 tables such as risk factor exposure (yes, no) by disease (yes, no). PROC SURVEYFREQ has several options for 2 x 2 tables to quantify the strength of the relationship between dichotomous exposure and outcome variables. They are: disease prevalence difference between exposed and not exposed domains, ratio of disease prevalence (exposed to not exposed), and ratio of disease odds (exposed to not exposed). Point estimates of these target population parameters are provided, along with confidence limits.

For cross-sectional survey data, both the prevalence ratio and the odds ratio in the target population can be estimated. In this case I recommend that the prevalence ratio be presented rather than the odds ratio. The calculated odds ratio always is larger than the calculated prevalence ratio, and as the overall prevalence of the disease increases, this discrepancy between odds ratio and prevalence ratio widens. Presentation of the odds ratio, especially for prevalent conditions like obesity or overweight, in my opinion may overstate the strength of the relationship between the exposure and outcome variables.

HOW TO DECIDE IF POINT ESTIMATES ARE PRECISE ENOUGH TO REPORT

The examples above show that estimated standard errors for point estimates are small and confidence intervals for population parameters are narrow when inference is to the target population or to some demographic domains of interest. However, it may be of interest to estimate parameters for a domain or subpopulation that is a small percentage of the target population and for which the sample size is not large. NCHS suggests that researchers use the coefficient of variation (CV) of a point estimate to decide whether or not to report any point estimate. A point estimate that is too variable, i.e. unstable, should not be reported.

The estimated CV for a given point estimate is defined as the estimated standard error of the point estimate divided by the value of the point estimate. Sometimes the CV is called relative standard error (RSE) since the size of the estimated standard error is judged *relative to* the size of the point estimate. NCHS recommends that any reported point estimate have an estimated CV less than 0.30 or 30% and that any point estimate with a larger CV not be reported.

PROC SURVEYFREQ and PROC SURVEYMEANS have the option to request the estimated coefficient of variation for point estimates of means, percentages and totals or sums (e.g. number of persons in target population with arthritis).

Some surveys offer their own guidelines for reporting point estimates in terms of sample sizes or in terms of sample sizes combined with the use of the CV.

CONCLUSION

It is important for secondary data analysts of complex survey data to understand the statistical context in which they work. Understanding the paradigm of a finite population helps you to define the target population parameters you wish to estimate and helps you specify null hypotheses about target population parameters that you wish to test. Basic knowledge of prototype sampling plans and different approaches to variance estimation helps you to correctly describe the dataset's sampling plan and variance estimation method to the survey software. Understanding the correct procedure for variance estimation for subpopulations prevents you from underestimating variance in this situation. Knowing how to estimate simple target population parameters helps you to read correctly the output from the survey software. Understanding the statistical inference process helps you to state correctly the analytical results in your paper presentation or journal article. Although it is important to know how to code the available SAS/STAT SURVEY procedures, coding expertise alone, without understanding the statistical context, may lead to flawed research.

The SAS/STAT SURVEY procedures provide excellent capability for analysis of complex survey data, with some of these capabilities illustrated in this paper. It is anticipated that SAS will add to their capabilities for complex survey data analysis as new mathematical statistical research is done to convert existing standard statistical procedures for SRS data for use with complex survey data.

My comments in this paper are consistent with the *design-based* approach to analysis of complex survey data, and it is for this approach that most survey software, such as procedures included in SAS/STAT, has been written.

REFERENCES

Blumberg, S. J. and J.V. Luke. 2014. "Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, January-June, 2014." National Center for Health Statistics. Accessed March 26, 2015. http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201412.pdf

Blumberg, S. J., J. V. Luke, M.L. Cynamon and M. R. Frankel. 2008. "Recent Trends in Household Telephone Coverage in the United States." *Advances in Telephone Survey Methodology,* edited by J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Japec, P. J. Lavrakas, M. W. Link and R. L. Sangster, 56-86. Hoboken, NJ: John Wiley.

Brogan, Donna. 2005. "Sampling Error Estimation for Survey Data." *Household Sample Surveys in Developing and Transition Countries*, 447-490. New York, NY: United Nations.

Brogan, Donna. 2015. "Analysis of Complex Survey Data, Misuse of Standard Statistical Procedures", to be published online in *Wiley Stats Ref*, http://onlinelibrary.wiley.com/book/10.1002/9781118445112

Centers for Disease Control and Prevention, 2015A. "PRAMS." Accessed March 26, 2015. http://www.cdc.gov/prams/

Centers for Disease Control and Prevention, 2015B. "Behavioral Risk Factor Surveillance System, BRFSS". Accessed March 26, 2015. http://www.cdc.gov/brfss/

Dillman, Don A., J. D. Smyth and L. M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: the Tailored Design Method.* 4th ed. Hoboken, NJ: John Wiley.

Groves, Robert M., F. J. Fowler Jr., M. P. Couper, J. M. Lepkowsi, E. Singer and R. Tourangeau. 2009. *Survey Methodology.* 2nd ed. Hoboken, NJ: John Wiley.

Lohr, Sharon. 2010. *Sampling: Design and Analysis*. 2nd ed. Boston, MA: Brooks/Cole, Cengage Learning.

National Center for Health Statistics, 2015A. "National Health Interview Survey." Accessed March 26, 2015. http://www.cdc.gov/nchs/nhis.htm

National Center for Health Statistics, 2015B. "National Health Interview Survey: 2013 Data Release." Accessed March 26, 2015. http://www.cdc.gov/nchs/nhis/nhis_2013_data_release.htm

National Center for Health Statistics, 2015C. "National Health and Nutrition Examination Survey." Accessed March 26, 2015. http://www.cdc.gov/nchs/nhanes.htm

National Center for Health Statistics, 2015D. "National Health Care Surveys." Accessed March 26, 2015. http://www.cdc.gov/nchs/dhcs.htm

National Center for Health Statistics, 2015E. "National Immunization Survey." Accessed March 26, 2015. http://www.cdc.gov/nchs/nis.htm

National Center for Health Statistics, 2015F. "National Health Interview Survey: 2013 Imputed Family Income/Personal Earnings Files." Accessed March 26, 2015. http://www.cdc.gov/nchs/nhis/2013imputedincome.htm

RECOMMENDED READING

- *Applied Survey Data Analysis,* by Steven G. Heeringa, Brady T. West and Patricia A. Berglund. 2010. Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton, FL.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Donna Brogan, Ph.D.
Professor Emerita of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University, Atlanta, GA 30322
dbrogan@emory.edu