Slide 1 of 22

Title Slide: NHES Sample Design, Weights, Variance, and Missing Data

Slide 2 of 22

This module provides information about the weights and variables associated with estimates from the NHES:2012. It summarizes the NHES:2012 sample design as it relates to study weights and describe the weights that must be applied to assure data are representative of the target population(s). In addition, this module explains the procedures for calculating appropriate standard errors. It also describes the missing data codes used in the NHES and how to handle missing data to ensure accurate analysis.

Throughout this module, underlined blue screen text indicates a link to additional resources.

Slide 3 of 22

The NHES:2012 sample was a two-stage, stratified address-based sample covering the 50 states and the District of Columbia.

The first sampling stage selected residential addresses from the list of addresses provided by a vendor based on the United States Postal Service Computerized Delivery Sequence File.

The second sampling stage selected an eligible child from information provided on the household mail screener. The screener questionnaire was used to enumerate household members under 20 years of age in order to determine eligibility for the extended topical survey.

Both the screener and topical questionnaires were completed through self-administered paper-based questionnaires sent through the mail.

Slide 4 of 22

As in past NHES surveys, the NHES:2012 oversampled Black and Hispanic households using Census and sampling frame data. This oversampling is necessary to produce more reliable estimates for subdomains defined by race and ethnicity. In order to facilitate this oversampling, addresses were stratified by race/ethnicity into three mutually exclusive strata, defined by the following criteria:

The first stratum comprised Census tracts with 25 percent or more Black persons;

The second stratum comprised Census tracts with 40 percent or more persons of Hispanic origin and not 25 percent or more Black persons; and,

The third stratum comprised all other tracts.

Slide 5 of 22

The NHES:2012 sample contained all types of residential addresses in order to ensure the best possible coverage of households in the United States.

The initial sample of addresses was divided into two groups based on the address type: Group 1 contained any PO Boxes not flagged as a household's "only way to get mail", or *O-W-G-M*. Group 2 contained all other addresses which included street and city-style addresses, high rises, rural routes, OWGM PO Boxes, and addresses flagged as seasonal, educational, or vacant. The sample also included "drop points", where a single postal delivery point is used for multiple housing units, and "PO Box throwbacks", which are street addresses where the mail is delivered to a customer's PO box. Educational addresses, which are addresses identified as colleges, universities, dormitories, and apartment buildings occupied by students, were also included.

Addresses were then subsampled at different rates, depending on the address group. Undersampling occurred in group 1, the non-OWGM addresses to provide coverage of this address type, since some of these addresses may be occupied, non-duplicative addresses, while increasing the efficiency of the mailing.

Slide 6 of 22

In the NHES:2012, only one child per household was randomly selected for a Topical survey. In earlier iterations of the NHES, Computer Assisted Telephone Interviewing, or CATI, programming allowed for collection of data about multiple children per household, while limiting overall response burden. The mail format used in the redesign removed this flexibility.

Because ECPP-eligible children comprise a smaller portion of the population compared to PFI-eligible children, differential sampling was applied in households with children eligible for both Topical surveys, in order to ensure a sufficient sample size for the ECPP survey.

Each household in the sample was randomly pre-designated as either an "ECPP household" or "PFI household." This pre-designation was used only when a household had children eligible for both surveys. In any household with children in the eligible population for only <u>one</u> survey – that is, either the ECPP or PFI, but not both - one child was randomly selected for that survey. If a household had children eligible for both the ECPP and PFI surveys, one child was randomly selected in a pre-designated domain to receive the appropriate survey.

Slide 7 of 22

As was discussed in the Common Module “Analyzing NCES Complex Survey Data,” weights must be used in analyses of data from studies like the NHES in order to make estimates produced from the sample representative of the target population. Due to sample stratification and use of differential sampling rates at both sampling stages, the members of the NHES population did not all have the same chance of being selected into the study sample. Therefore, weights must be used to account for the study design. The weights also account for undercoverage and differential response rates to assure that the data are still representative of the population, even though some members of the sample did not respond to the survey. These weights are provided on the NHES data files.

Weights must be used to compute population estimates from the NHES. In the NHES:2012, the ECPP sampling weights are designed to make the sample representative of the population of children aged 0 to 6 and not yet enrolled in school. The PFI sampling weights are designed to make the sample representative of the population of children aged 20 or younger enrolled in kindergarten through 12th grade, or homeschooled for the equivalent grades.

Slide 8 of 22

Information from the first stage of sampling was used to create the household-level weights. Because the NHES:2012 is primarily concerned with information about eligible children, the household-level weights were calculated specifically as a basis for computing the person-level weights.

The **household base weight** represents the address’s probability of selection from the frame. It is the product of two factors:

- The weight associated with the differential sampling of addresses based on the race/ethnicity stratum of the address;
- And, differential sampling of Post Office box addresses that were designated as the only way to get mail.

The household-level base weight was then adjusted for screener nonresponse and address ineligibility, such as address vacancy or misclassification as residential.

The base weights of the nonrespondent cases, and a portion of the unknown eligible cases, were distributed among the base weights of the respondent cases within a nonresponse adjustment cell.

Chi-Square Automatic Interaction Detection analysis, or CHAID analysis, was used to identify characteristics most associated with screener nonresponse. The CHAID analysis is an algorithm-based multivariate analysis designed for categorical data that uses a decision-tree technique. CHAID first identifies the characteristic of the data that is the best predictor of response. Then, within the levels of that characteristic, CHAID identifies the next most likely response predictor or predictors, and so forth, until a tree is formed with all potential response predictors. The final result is a division of the entire dataset into cells by attempting to determine sequentially the cells that have the greatest discrimination with respect to the unit response rates.

Slide 9 of 22

The household-level weight was used as the base weight for each of the person-level weights for both the ECPP and PFI.

The person-level weight for each sampled person in the household is the product of the final household weight and several weight adjustment factors, including the following:

- The weight associated with sampling the person's domain, either ECPP or PFI, in the household;
- The weight associated with sampling the person from among all eligible persons in the given domain in the household;
- The weight associated with sampling a child in a joint custody arrangement at both parents' addresses; and,
- The weight associated with the ECPP or PFI topical questionnaire unit nonresponse.

The unit nonresponse adjustment at the topical level used characteristics for which information was available for both respondents and nonrespondents to form nonresponse adjustment cells. The adjustment cells were determined by a separate CHAID analysis for each topical survey.

The final step of person-level weighting was to rake the nonresponse-adjusted person-level weights to national control totals. The raking procedure typically improves the reliability of survey estimates and also corrects for the bias that results from households or persons not covered by the survey. The raking procedure for the NHES:2012 ECPP and PFI weights involved raking the nonresponse-adjusted person-level weights to national totals obtained using the number of children from the 2011 annual American Community Survey, or ACS (*A-C-S*) estimates.

Slide 10 of 22

The final ECPP and PFI person-level weights are available on the respective NHES data files.

The final weights to be used in all child-level analyses are named FEWT for the ECPP data file and FPWT for the PFI data file.

Slide 11 of 22

This table compares the unweighted sample counts and distributions and the weighted population counts and distributions of several key demographic variables for the NHES:2012 ECPP. For example, approximately 40 percent of the unweighted sample attends a center-based child care program. However, in the population, only about 34 percent of the children attend a center-based child care program. This suggests that parents of children in center care responded to the survey at a higher rate than parents of children not attending a center-based program. The table therefore illustrates the importance of weighting the data in order to produce accurate population estimates.

Slide 12 of 22

This table shows the unweighted sample counts and distributions and the weighted population counts and distributions of several key demographic variables for the NHES:2012 PFI. For example, in the unweighted sample, the percentages of children in elementary and high school are similar. However, in the population, about 49 percent of the students are in elementary school, compared to approximately 28 percent in high school. This suggests that parents of high school students responded to the survey at a higher rate than parents of elementary school students. Again, the table illustrates the importance of weighting the data in order to produce accurate population estimates.

Slide 13 of 22

NHES data users must always weight the data in order to accurately reflect the population of inference. Each data set for the ECPP and PFI has a final weight that is applicable for all cases in that data set. Unweighted data will not account for the sample design, nonresponse, or undercoverage and, therefore, will not represent the population.

Slide 14 of 22

In Common Module, “Statistical Analysis of NCES Datasets Employing a Complex Sample Design,” two standard error calculation procedures were discussed: Replication Techniques and Taylor Series linearization.

Replication is a method that calculates appropriate standard errors based on differences between estimates from the full sample and a series of created subsamples, or replicates.

If this is the method you will use to compute standard errors, you need to use the replicate weight variables that are associated with your sampling weight variable. For both the ECPP and the PFI, there are 80 replicate weight variables named FEWT1 through FEWT80 and FPWT1 through FPWT80 respectively.

The NHES:2012 replicate weights were created using a Jackknife replication method, specifically JK1. JK1 should be specified as the replication method within the statistical software used for analysis.

Slide 15 of 22

The second method for calculating standard errors is the Taylor Series Linearization Method. This method of standard error calculation uses PSU and strata identifiers to calculate the appropriate standard errors. If this is the method you will use to compute standard errors, you need to select the PSU and strata identifiers that are associated with your sampling weight variable. The PSU and stratum variables are named EPSU and ESTRATUM on the ECPP data file and PPSU and PSTRATUM on the PFI data file.

Slide 16 of 22

While the item response rates in the NHES:2012 are generally high, in the NHES:2012, as in most surveys, responses were not obtained for some items in the survey. It is important to understand **why** data are missing in order to appropriately recode missing data for your analysis.

There are numerous reasons for item nonresponse. Some respondents may not have known the answer to a question or simply did not wish to respond. Some respondents may have run out of time and left items at the end of the survey blank. Additionally, some respondents received short-form versions of the NHES questionnaires as a nonresponse follow-up strategy. Item nonresponse occurred for these respondents for items that were included on the full form, but not on the short-form versions of the questionnaires. Item nonresponse may have also occurred because a respondent's responses were not internally consistent, and this inconsistency was discovered during the editing stage of data processing, after the end of data collection. In many cases, items that were not internally consistent were set to "missing" during the editing stage. The NHES:2012 items that were set to missing during editing, that were missing due to nonresponse, or that were not asked on a short form, were all imputed.

Slide 17 of 22

Missing NHES questionnaire data are fully imputed. Three approaches to imputation were used in the NHES:2012:

- Unweighted sequential hot deck imputation, which was used for the majority of the missing data;
- Weighted random imputation, which was used for a small number of variables; and,
- Manual imputation, which was used in a very small number of cases for most variables.

These imputation methods are described in detail in the Data File Users' Manual.

For each data item for which any values were imputed, an imputation flag variable was created. These flags are named "F_" followed by the variable name. If the response for the item was not imputed, the imputation flag was set equal to 0. The imputation flag was set to 1 if the missing value was imputed using the standard hot deck approach. If an item was imputed manually, the flag was set to 2. The imputation flag was set to 3 for cases that were imputed using weighted random imputation.

The imputation flags were created to enable users to identify imputed values. Users can employ the imputation flag to delete the imputed values, use alternative imputation procedures, or account for the imputation in computations of the reliability of the estimates produced from the dataset.

Slide 18 of 22

There are four exceptions where missing data are not imputed.

First, open-ended responses to “other, specify” questions were not imputed. For these variables, missing values are set to -9.

Second, if a respondent reported his or her ethnicity as Hispanic but did not report a race, the race variable was not imputed. Instead, it was set to “Hispanic, race not reported.”

Third, variables appended to the NHES:2012 data files, such as sampling frame variables from the Common Core of Data or the Private School Universe Survey, and Census-derived variables were not imputed. Missing values on these variables are coded in the way that they are coded on the source data files.

Finally, some questions were intentionally skipped based on a respondent’s answer to a previous question. For example, an item asking about the number of hours per week a child spends in center-based nonparental care would be skipped if the respondent indicated that the child did not attend center-based child care. In these situations, the skipped question is treated as “not applicable” and is coded as -1 on the data file.

Each of these exceptions to imputation in the NHES:2012 data are described in the data file user’s manual.

Slide 19 of 22

As there were two PFI questionnaires used in the NHES:2012 data collection (PFI-enrolled and PFI-homeschool), researchers using PFI data should note that children whose parents responded to the PFI-enrolled questionnaire will have -1 (not applicable) codes for any variables that correspond to items that were only asked on the PFI-homeschool questionnaire. The reverse will be true for children whose parents responded to the PFI-homeschool questionnaire.

The variable PATH can be used to identify children whose parent completed the PFI-enrolled versus the PFI-homeschool questionnaire.

Slide 20 of 22

Again, it is important to understand **why** data are missing in order to appropriately recode missing data for your analysis. Analysts should review the questionnaire instruments to familiarize themselves with the skip patterns and decide for themselves how to treat missing data.

For example, depending on your particular research question, you may wish to recode some “not applicable” values to a valid code. In other situations, however, you may wish to recode these values to “system missing” so that they will be excluded in statistical analyses. For example, if you are interested in the average number of hours all children aged 0 to 6, not yet enrolled in kindergarten, spend in center-based care per week, you may wish to recode “not applicable” values to “zero” for those children who do not attend center-based care. However, if you are only interested in the average number of hours in center-based care for children who have a regularly scheduled center-based care arrangement, then you may wish to recode “not applicable” values to “system missing”.

Slide 21 of 22

This module provided information about the weights and variables associated with estimates from the NHES:2012. It summarized the NHES:2012 sample design as it relates to study weights and described the weights that must be applied to assure data are representative of the target population(s). In addition, this module explained the procedures for calculating appropriate standard errors. It also described the missing data codes used in the NHES and how to handle missing data to ensure accurate analysis.

Slide 22 of 22

Additionally, important resources that have been provided throughout the module summarized here along with the module’s objectives for your reference.

You may now proceed to the next module in the series, or click the exit button to return to the landing page.