As we discussed in Week 4 of Course 1, one should always have a solid grasp of the study design that was used to generate the data set being analyzed. When analyzing survey data, like the NHANES data that we have been analyzing this week, one should always ask the question of whether the underlying sample that was measured was selected using a complex sampling plan. Recall from Week 4 of Course 1 that such sampling plans may involve unequal probabilities of selection into the sample for different population units, stratification of the target population, and cluster sampling to save on the costs of data collection. If one wishes to make inference about a finite target population, these design features need to be accounted for as part of the inferential process.

The purpose of this document is to revisit the NHANES analysis examples that we have been discussing this week, and examine whether our inferences about the NHANES target population change when incorporating the complex sampling features of the NHANES (weighting, stratified sampling, and cluster sampling) into the analysis procedures.

First, we need to identify the additional variables in the NHANES data set containing the complex sampling features:

**WTINT2YR**: This variable contains the final survey weights for all sampled individuals given the medical history interview (NOTE: There is also a weight for those given a physical exam, but for purposes of this example, we will assume that this is the survey weight of interest). Roughly speaking, the weight indicates how many other people in the target population this survey respondent is representing (meaning that the sum of the weights provides an estimate of the size of the target population). If these weights are correlated with the measures of interest, we need to incorporate them in our estimation to compute representative finite population estimates. Recall also that the use of weights in estimation will tend to *increase* the standard errors of our estimates (the more variable the weights, the higher the standard errors).

**SDMVSTRA**: This variable contains codes representing the sampling strata used to select the NHANES sample. Recall from Week 4 in Course 1 that stratified sampling is expected to decrease the standard errors of survey estimates.

**SDMVPSU**: This variable contains unique codes for the primary sampling units (or clusters) that were randomly selected within each of the sampling strata. These codes generally refer to large geographic areas, like U.S. counties. Recall from Week 4 in Course 1, and our earlier lectures in this course, that cluster sampling is expected to *increase* the standard errors of survey estimates.

Given these variables, we will now revisit our inference examples that employed confidence intervals (the same basic ideas would apply to the hypothesis testing approaches as well.

First, let’s consider the example of estimating one proportion from Lecture 2. Table 1 below shows what happens to our estimate, its standard error, and the 95% confidence when accounting for different aspects of the complex sampling.

**Table 1: Revisiting estimation and inference for one proportion when accounting for the NHANES complex sample design.**

| **Approach** | **Estimate** | **SE** | **95% CI** |
|---|---|---|---|
| Ignoring Sample Design (assuming a simple random sample) | 0.410 | 0.015 | (0.381, 0.438) |
| Accounting for Weights Only | 0.374 | 0.015 | (0.345, 0.403) |
| Accounting for Stratification and Cluster Sampling Only | 0.410 | 0.022 | (0.364, 0.456) |
| Fully Accounting for Complex Sampling Features | 0.374 | 0.021 | (0.329, 0.418) |

In this example, we see that the weights made a non-negligible difference in our point estimate of the proportion of black adults with systolic blood pressure greater than 130, shifting it from 0.410 to 0.374. The sample design features also had a slight impact on the standard error of the estimate, increasing it by about 40%. Overall, if our objective was finite population inference, we would focus on the results in the final row of Table 1, where we fully accounted for the complex sampling features. Failing to account for the complex sampling in the analysis would lead us to overstate the proportion of black adults with high systolic blood pressure. The confidence intervals for the weighted and unweighted estimates do overlap, however, so our inferences would not change substantially.

As a second example, let's revisit our analysis comparing the mean systolic blood pressure among male African-American adults and female African-American adults. Table 2 presents our estimates of the mean difference, its standard error, and the 95% confidence interval for the mean difference following the alternative approaches.

### Table 2: Revisiting estimation and inference for the difference in means when accounting for the NHANES complex sample design.

| Approach | Estimate | SE | 95% CI |
|---|---|---|---|
| Ignoring Sample Design (assuming a simple random sample) | 5.221 | 1.181 | (2.906, 7.536) |
| Accounting for Weights Only | 4.249 | 1.131 | (2.032, 6.466) |
| Accounting for Stratification and Cluster Sampling Only | 5.221 | 1.443 | (2.393, 8.049) |
| Fully Accounting for Complex Sampling Features | 4.249 | 1.523 | (1.265, 7.233) |

We once again see that the weights are making a fairly substantial difference in the estimate of the mean difference: the estimate is nearly one unit lower when accounting for the survey weights. The complex sampling features are also once again increasing the variance of the point estimate. Despite the fact that the 95% confidence intervals overlap, we would likely report the results in the bottom row when making finite population inferences; for example, we would not rule out a difference of 2 in the means as being plausible when accounting for the complex sampling, but we would reject this hypothesis when ignoring the complex sampling features.