



Sampling Design Options for Virginia Health Insurance Survey

Michael Davern, Ph.D.

The Virginia project team is interested in obtaining health insurance coverage estimates for a representative sample of people living in Virginia. The Virginia Health Insurance Survey team would like to have precise estimates for 5 health service regions, as well as a low-income over-sample. The low-income stratum is represented by area code plus telephone exchange groupings throughout the state with an estimated household Median income below \$32,000. The non-low-income portions of the health service areas are allocated to five separate strata. Our data to form the counts of households within the low-income stratum, each of the five health service area strata, and our data on the median income for each area code plus telephone exchange grouping in Virginia was based on data provided by Gensys Marketing System Group¹. The health services areas were mapped to area code plus exchange groupings using the zip code data given to us by the Virginia project team and the primary zip code associated with each area code plus exchange grouping data provided to us by Gensys Marketing Systems Group.

In order to determine the appropriate sample size, two pieces of information are needed. The first is the expected overall uninsurance rate for Virginia. The second is the desired amount of precision we would like to be able to make with each estimate. Because we have not conducted the Virginia Health Insurance Survey yet, we have to make our best guess as to what we expect the rate to be. The uninsurance rate in Virginia is was 14.9 percent according to a recent statewide survey. The rate for Health Service

¹ They use Census 2000 data to make their estimates of the number of households within each area code plus exchange grouping and to estimate the median household income within each area code plus exchange grouping.

Region 1 (Northern Virginia) was 11.0, the uninsurance rate for Health Service Region 2 (Eastern Virginia) is 13.1, the uninsurance rate for Health Service Region 3 (Central Virginia) was 17.2, the uninsurance rate for Health Service Region 4 (Northwest Virginia) was 14.1, and the uninsurance rate for Health Service Region 5 (Southwest Virginia) was 20.4. We will use these estimates to derive our minimum number of observations for the five Health Service Area strata. For the low-income stratum, we expect to find considerably higher levels of uninsurance. The most recent CPS figures show roughly 30 percent uninsurance rate for low-income people in Virginia.

To determine the desired amount of precision, we need to think about how large of confidence intervals we want around our key estimates. A 95 percent confidence interval of plus or minus 4 percent around an estimate gives a high level of precision. This means that our 95 percent confidence interval around our estimate of uninsurance would be 10.4 percent \pm 4 percent. The low income stratum would be 30 \pm 4 percent.

With the desired level of precision and the uninsurance rate we can then solve the following equation to determine the total sample size:

$$N = \frac{P * (1 - P) * DE}{SE^2}, \quad (1)$$

In formula 1 “N” is the number of respondents, and “P” is the rate of interest (e.g., the percent uninsured). “DE” is the design effect.² “SE” is the standard error and is computed by dividing the desired width of the confidence interval (either plus or minus 4 percent) by 1.96.³

The final piece we need to know is the design effect. The design effect is equal to the ratio of the variance taking the survey design features (stratification and clustering) under consideration, to the variance calculated under the assumption that all cases were sampled through a simple random sample procedure (Kish 1965). Because telephone surveys often involve three main adjustments to determine the probability of selection of each respondent, they are not simple random samples. The first adjustment is to account

² For CSCS surveys a design effect of 2 is expected. This can be increased, however, by having a very high degree of disproportionate sampling.

³ 1.96 is the value of a “z” distribution representing a 95 percent confidence interval.

for the fact that some households have more than one telephone line (and thus an increased probability of selection), the second adjustment is to account for the fact that some households have more people living in them than others (assuming one random person is selected from each household as the respondent), and the third adjustment is for the fact that some phone numbers are more likely to be selected due to disproportionate sampling; for example, sample a higher percentage of people from health service region 1 than in health service region 2.⁴

For RDD surveys using a fair amount of disproportionate sampling (for example like the Virginia Health Insurance Survey), that randomly chooses one person in each household to be the target, generally have a design effect equal to 2. A design effect of 2 means that the variance calculated taking into account the complex sample design features is 2 times as large as the variance calculated assuming that all the cases were selected through a simple random sample procedure. The square root of the design effect is multiplied by the standard error to obtain the actual standard error corrected for the survey design (1.41). (The square root of the design effect should be multiplied by the standard error calculated under the assumption of simple random sampling in order to obtain the appropriate standard error). Solving equation 1 with a design effect of 2 results in a minimum sample size for Health Service Area 1 (Northern Virginia) of 470, for Health Service Area 2 (Eastern Virginia) of 547, for Health Service Area 3 (Central Virginia) of 684, Health Service Area 4 (Northwest Virginia) of 602, and Health Service Area 5 (Southwest Virginia) of 780. Solving equation 1 for the low-income stratum we obtain about 1008 for a minimum sample size.

With the minimum cases allocated to the 5 health services areas and 1008 minimum cases allocated to the low-income stratum in Virginia we have a total sample

⁴ Disproportionate sampling (or stratified random sampling) can actually lead to a decrease in the design effect to the extent that the dependent variable is related to the strata (e.g., the 3 regions). However, many times the relationship is weak and the heterogeneity in the weights causes a higher design effect than would have been achieved without the disproportionate sampling. Furthermore, in telephone sampling our strata are often highly correlated with the actual strata (region of residence) but not everyone sampled in a particular telephone area code plus exchange combination actually lives in a region that the area code plus exchange was assigned to. Thus, in an analytical application we often have a higher design effect when using stratified sampling in a RDD survey than we would have had without the disproportionate sampling.

size of 3083 (the 1008 low-income cases are all also part of one of the five Health Service Areas). We will allocate these to the health service areas on the basis of the estimated number of households in each of the non-low income telephone area code plus exchange groupings in each health service area. This will lower the statewide design effect by reducing the variation in the basic household probabilities of selection. The basic probability of selection is the number of sampled households from each strata divided by the total number of households within the strata. The basic household weight is the inverse probability of selection. Table 1 shows the general breakdown for the state of Virginia.

Table 1: Virginia Sample Plan Summary

Stratum	Households	Minimum Sample Size Needed	Distribution of Low Income Sample In Health Services Areas	Additional Sample In Health Service Area to Reach Minimum	Basic Weight for Each Stratum at Minimum	Additional Sample to Smooth Out Weights*	Total Sample Size In Each Stratum	Average Household Weight Per Stratum
Health Service Region 1	367822	470	52	418	881	29	447	824
Health Service Region 2	692341	547	0	547	1266	293	840	824
Health Service Region 3	327759	684	596	88	3718	310	398	823
Health Service Region 4	494778	602	179	423	1169	177	600	824
Health Service Region 5	583536	780	181	599	974	108	707	825
Low Income **	377123	1008			374		1008	374
Total	2843359		1008	3083			4000	711

* The additional sample is used to smooth out the basic weights among the health services regions to drive down the design effect

** Low income cut is an area code plus exchange grouping with a median household income below \$32,000

Source: Median income, primary zip code and number of households for each area code plus exchange grouping in Virginia is provided by Gensys Marketing Systems Group and uses Census 2000 data; The primary zip code of each area code plus exchange grouping is matched to the appropriate health service region through zip code level data.

Conclusion:

SHADAC recommends that the Virginia project team pursue this recommended sample design. The design currently meets the minimum threshold within each strata for obtaining a reasonably precise measure of the uninsurance rate. The sampling plan also distributes the remaining cases in order to smooth out the probabilities of selection of the households included in the survey. This will make the state-wide estimates from the survey even more precise and drives down the design effect of the entire sample.