**11. The Area Sample Design One Solution**

The project in Chapter 8 requested that you design a sample of twenty-five census tracts and one block group per sample census tract (*m=*1). The desired total sample size is 1,000 persons which was split equally among five age groups. Thus, the requirement for an equal workload per block group (BG) leads to  persons in each BG. Table 11.1 shows the population counts from the 2000 census for the five age domains. Each domain was to receive a sample size of 200. The implied sampling rates range from about 0.12% for ages 25-44 to 0.51% for ages 18-24.

**Table 11.1** Population, sample size, and overall sampling rate for five age domains in Anne Arundel county, Maryland.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age domain  *d* | Population | Percent of population | Sample size | Domain sampling rate |
| 18 to 24 | 39,448 | 10.76% | 200 | 0.5070% |
| 25 to 44 | 160,940 | 43.92% | 200 | 0.1243% |
| 45 to 54 | 71,657 | 19.55% | 200 | 0.2791% |
| 55 to 64 | 45,637 | 12.45% | 200 | 0.4382% |
| 65+ | 48,765 | 13.31% | 200 | 0.4101% |
| Total | 366,447 | 100.00% | 1,000 | 0.2729% |

Since a self-weighting sample within each age group is desired along with the same workload in each PSU, the composite measure of size (MOS) method, described in section 10.7 can be used. In particular, the composite MOS for BG *j* in tract *i* is



where  is the number of persons in age group *d* in tract *i* and BG *j.* The MOS for tract *i* is then  where  is the set of all BG’s in tract *i.* The total MOS across all tracts and BG’s is . The project assignment asks you to select tracts and BG’s using Sampford’s procedure, which is one method of probability proportional to size selection in which joint selection probabilities can be computed. If we select a  sample of tracts followed by a  sample of 1 BG in each tract, then the selection probability of that BG is

**.

This is the same selection probability that would be obtained by selecting a sample of 25 BG’s directly from the frame of BG’s. However, notice that this sample design of selecting tracts first, followed by a single BG per tract, is not the same as selecting BG’s directly. If we selected BG’s directly using Sampford, all pairs of BG’s would have non-zero joint selection probabilities. Since we select tracts and then 1 BG per tract, the joint selection probability of any two BG’s in a given tract is zero.

The spreadsheet, *AnneArundel.MD.solution.xls*, shows the value of the composite MOS for each tract and BG, along with population counts by age group, and a variety of other calculations. Note that some ages are out-of-scope for this survey (0 to 5 years, 6 to 11, 12 to 17). These are excluded from the composite MOS.

A number of quality control checks must be made to determine whether sampling at the desired rates is possible in all BG’s. As outlined in section 10.7, the expected number of persons sampled in each domain in each SSU (BG) should be less than the population count in the SSU. Also, the sum of these expected counts in a BG across the domains must be less than the population in the BG. There are six BG’s that violate the requirement that  where  is the expected sample number in BG *ij* from domain *d.* The six are shown in Table 11.2. Each violates the sample size constraint in at least one age group. For example, tract 701400, block group 3 has a population of 16 in the 25-44 group, but the sampling algorithm requires an expected sample size of 16.4; the population is 7 in the 65+ group but the desired sample size is 23.6. A borderline case is tract 741100 where the population of 18-24 is 10 and the sample is to be 10.1.

Two other BG’s are shown in Table 11.2 that have no population in any of the in-scope age groups. These could be left in the frame in case some eligible people have moved in since the 2000 census. Or, if we are confident that the entire BG is out-of-scope, it could be dropped from the frame. In fact, inspection of the map in *Anne Arundel.blkgrps(streets).pdf* reveals that the tract.BG’s 740602.1 and 740603.1 are on a military reservation or in a wildlife preserve in the western part of the county. If the eligible universe covers only the non-institutional household population, it might be safe to drop these BG’s but we have combined them with BG 2 in their respective tracts for this exercise. The other deficient BGs were combined with other tracts.BG’s as shown in the table.

There is also one tract that has a relatively small selection probability based on the initial calculations. Tract 741100 has a selection probability of 0.005; the next smallest is 0.022. This smallest tract contains a single BG, which as shown in Table 11.2, was combined with tract.BG 740603.2. Tract 741100 is, thus, combined with tract 740603.

**Table 11.2** Block groups where the expected workload exceeds the population count.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TRACT | BG |  | 18 to 24 | 25 to 44 | 45 to 54 | 55 to 64 | 65+ | Action: combine with tract.BG |
| 701400 | 3 | Population | 0 | 16 | 0 | 0 | 7 | 701400.2 |
| 701400 | 3 | Workload | 0 | 16.4 | 0 | 0 | 23.6 |  |
| 740602 | 1 | Population | 0 | 0 | 0 | 0 | 0 | 740602.2 |
| 740602 | 1 | Workload | NA | NA | NA | NA | NA |  |
| 740603 | 1 | Population | 0 | 0 | 0 | 0 | 0 | 740603.2 |
| 740603 | 1 | Workload | NA | NA | NA | NA | NA |  |
| 740603 | 3 | Population | 5 | 101 | 0 | 0 | 0 | 740603.2 |
| 740603 | 3 | Workload | 6.7 | 33.3 | 0 | 0 | 0 |  |
| 741100 | 1 | Population | 10 | 42 | 16 | 12 | 0 | 740603.2 |
| 741100 | 1 | Workload | 10.1 | 10.4 | 8.9 | 10.5 | 0 |  |
| 750600 | 1 | Population | 0 | 0 | 45 | 0 | 7 | 750600.2 |
| 750600 | 1 | Workload | 0 | 0 | 32.6 | 0 | 7.4 |  |
| 750700 | 2 | Population | 0 | 4 | 0 | 0 | 0 | 750700.1 |
| 750700 | 2 | Workload | 0 | 40 | 0 | 0 | 0 |  |
| 750801 | 5 | Population | 0 | 21 | 30 | 8 | 0 | 750801.4 |
| 750801 | 5 | Workload | 0 | 7.2 | 23.1 | 9.7 | 0 |  |

NA = not applicable

As this example shows, tracts that are geographically adjacent may not have consecutive identification numbers. Figure 11.1 is a schematic map of the tracts in the county. Consulting a map like this may be necessary to make reasonable combinations. Alternatively, longitude-latitude centroids for tracts are available from the Census Bureau. These can be used to calculate the distance between the centers of the tracts to determine which are geographically near each other. This approach will permit tracts to be combined via a computer algorithm without manual intervention. This is particularly useful when the frame of tracts is large.

**Table 11.3**  Summaries of tract and BG selection probabilities and weights after combining small units.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Probability or weight | Min. | 1st Quartile | Median | Mean | 3rd Quartile | Max. |
|  | 0.0225 | 0.1828 | 0.2673 | 0.2660 | 0.3329 | 0.5920 |
|  | 0.0015 | 0.0489 | 0.0783 | 0.0828 | 0.1030 | 0.4385 |
| 1/ | 1.69 | 3.00 | 3.75 | 5.25 | 5.48 | 44.44 |
| 1/ | 2.28 | 9.71 | 12.78 | 22.08 | 20.45 | 684.90 |

After these combinations are made, the selection probabilities for tracts and BG’s are summarized in Table 11.3. The range of selection probabilities for BG’s is 0.0015 and 0.4385 while the range of weights for BG’s is 2.28 to 684.90. Although the range of BG probabilities is substantial, a self-weighting sample or persons can still be selected from each domain since there are no deficient BG’s after combining. The Sampford method was used to select a sample of 25 tracts and then 1 BG per sample tract. The code for combining BG’s and tracts is in the file *Anne Arundel.MD.analysis.R*.

The selected sample tracts and BG’s are listed in Table 11.4 and shaded in Figure 11.2. The expected workloads in each BG are also shown in the table. The workloads are not integers. This means that when the samples of persons within a sample BGs are selected, the sampling will be done using fixed rates not fixed sample sizes. For example, tract.BG 701102.2 has a population of 76 in age group 18-24 and the sample size is 6.3 in Table 11.4. Persons in that age group and BG would be sampled at rate .

**Quality Control Checks**

Checking the correctness of your work is always important. In this case, there are some simple assessments that will help determine whether computations and sample selections are correct. The weight for a sample BG is . These can be used to make population estimates which we can compare to frame numbers. There are two conditions that should hold exactly for any sample that has been selected. First, define

.

The -estimator of the total for this variable across the domains in a BG is



since . In this case . Thus, the estimator of the total of  in any block group is the same constant, 40. The estimator of the population total of  is



which can be verified using the fact that  and the definition of . Since  is the total sample size, this -estimator must be 1,000. The population totals of the numbers of persons in each domain and across all domains can also be computed as

, *d*=1, …, 5

.

These do not necessarily equal the population counts, but serve more as a reasonableness check. If the estimates are far from the frame counts, then further checking is warranted to decide whether errors have occurred. For this sample, we have =40 for each domain,  = 1,000,  = (38,011.38, 173,593.95, 63,811.75, 45,011.18, 52,714.43), and  = 373,142.7. The estimates  and  are reasonably near the population counts in Table 11.1. These checks can also be found in *Arundel.MD.analysis.R.*

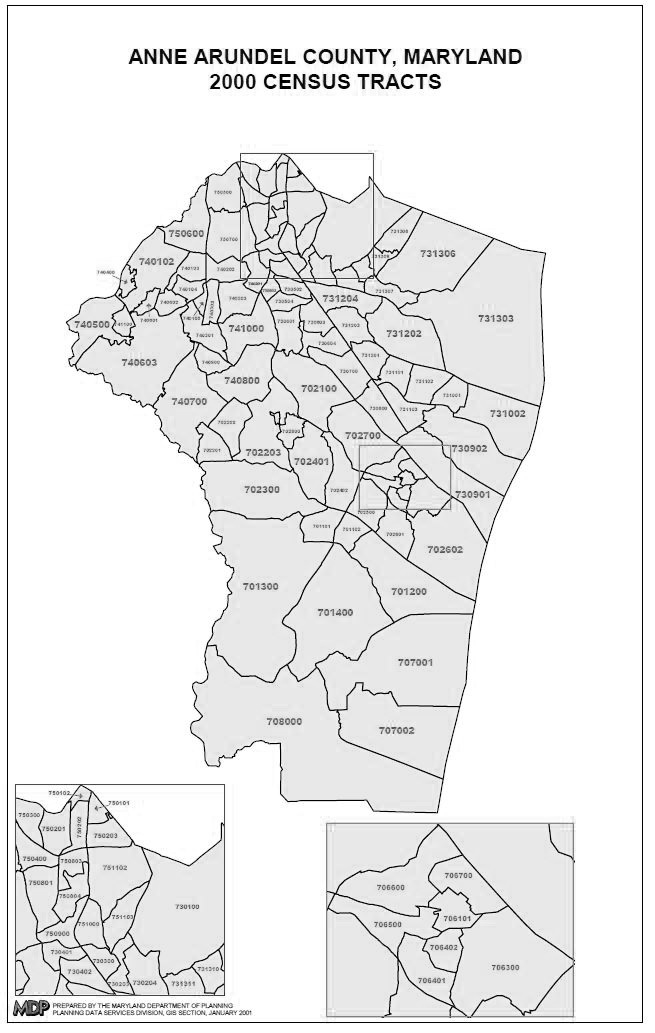
**Additional Considerations**

We spent some time above worrying about the effects of tracts and block groups with small composite measures of size. One of the paradoxes of designing samples is that a significant amount of time is spent considering events that may not happen. We may not select one of the BG’s with an extremely small MOS, but if we do, its size may not support the desired sample sizes for domains. In addition, its weight will be large and can unnecessarily increase variances. This issue will be addressed again in Chapter 15.

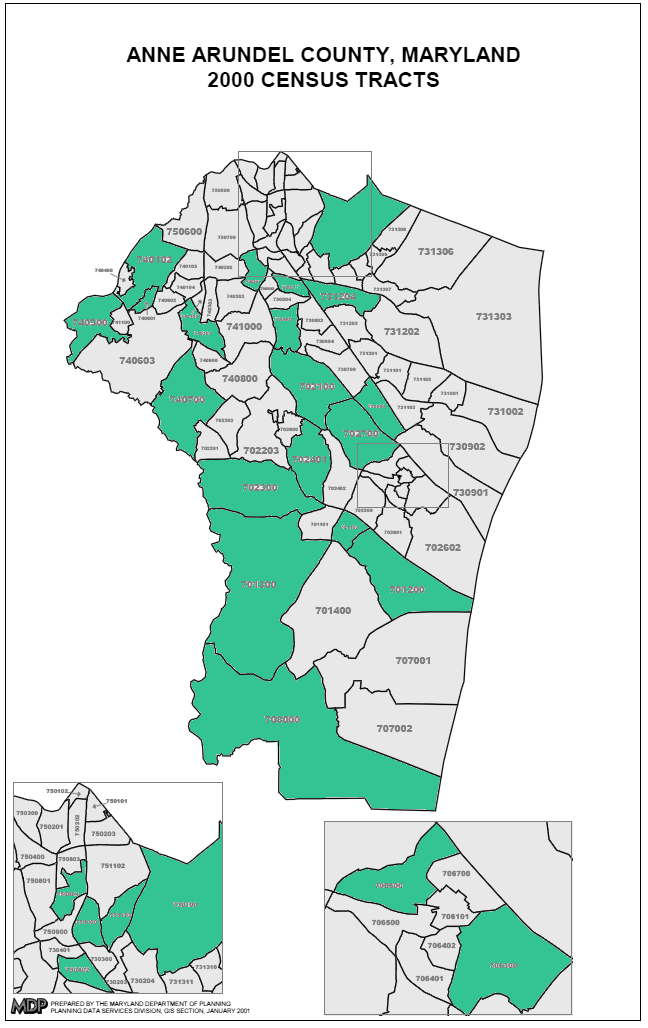
The fact that only 1 BG is selected per tract might raise the question of whether variances can be estimated with this design. We can still estimate design-variances because the number of first-stage units is 25, the number of sample tracts. However, there are alternative designs that might be worth considering. If the residents of different areas of the county were known to have different characteristics, it would be advisable to stratify by sub-county geography in some way. The BG numbers assigned by the Census Bureau can be used to sort the BG’s in a more-or-less geographic order and strata created from the sorted list. A BG map should be consulted to be sure whether numeric sorting will achieve your stratification goals. A BG map for Anne Arundel county is in the file, *Anne Arundel.blkgrps(streets).pdf*, on the website for this book.

**Table 11.4** Sample tracts and block groups within tracts with expected workloads in each BG.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Workloads | | | | | |
| Tract | Block group | 18 to 24 | 25 to 44 | 45 to 54 | 55 to 64 | 65+ | Total workload |
| 701102 | 701102.2 | 6.3 | 6.1 | 7.5 | 7.8 | 12.3 | 40 |
| 701200 | 701200.3 | 5.6 | 9.3 | 6.5 | 8.3 | 10.3 | 40 |
| 701300 | 701300.2 | 5.8 | 6.1 | 10.2 | 8.6 | 9.3 | 40 |
| 702100 | 702100.4 | 7.7 | 7.8 | 12 | 7.2 | 5.2 | 40 |
| 702300 | 702300.4 | 4.3 | 4.8 | 8.1 | 15 | 7.8 | 40 |
| 702401 | 702401.2 | 1.3 | 0.6 | 2 | 5.4 | 30.6 | 40 |
| 702700 | 702700.3 | 5.7 | 10.2 | 6.9 | 9.2 | 8 | 40 |
| 706300 | 706300.2 | 2.8 | 6.6 | 7.6 | 8.5 | 14.4 | 40 |
| 706600 | 706600.5 | 19.3 | 7.8 | 6.2 | 4.6 | 2.2 | 40 |
| 708000 | 708000.1 | 7.2 | 7 | 6.6 | 10 | 9.2 | 40 |
| 730100 | 730100.3 | 6.8 | 10.3 | 5.2 | 8 | 9.7 | 40 |
| 730402 | 730402.2 | 7.4 | 6.8 | 6.2 | 8.5 | 11.1 | 40 |
| 730502 | 730502.2 | 11.9 | 7.7 | 7.3 | 6.3 | 6.8 | 40 |
| 730601 | 730601.4 | 6.4 | 5.8 | 12.3 | 10.3 | 5.2 | 40 |
| 730800 | 730800.2 | 2 | 3.9 | 9.4 | 12.2 | 12.5 | 40 |
| 731204 | 731204.1 | 8.2 | 7.8 | 6.4 | 7 | 10.5 | 40 |
| 740102 | 740102.1 | 8.3 | 5.3 | 8.7 | 9.7 | 8 | 40 |
| 740201 | 740201.4 | 8.8 | 12.6 | 9.1 | 5.5 | 4 | 40 |
| 740301 | 740301.2 | 10.1 | 15.4 | 8 | 4.7 | 1.8 | 40 |
| 740500 | 740500.1 | 9.1 | 14.6 | 8.6 | 5.4 | 2.3 | 40 |
| 740601 | 740601.3 | 17.9 | 21.4 | 0.7 | 0 | 0 | 40 |
| 740700 | 740700.2 | 9.3 | 15.6 | 6.4 | 4.2 | 4.4 | 40 |
| 750804 | 750804.1 | 8 | 10.4 | 4.8 | 6 | 10.8 | 40 |
| 751000 | 751000.1 | 6.9 | 5.9 | 6.7 | 8.2 | 12.3 | 40 |
| 751103 | 751103.2 | 5.6 | 5.9 | 4.7 | 16.5 | 7.3 | 40 |



**Figure 11.1**  Tract map for Anne Arundel County Maryland. Source: Maryland Department of Planning, Planning Data Services Division, January 2001.



**Figure 11.2** Selected tracts in Anne Arundel county.