Wiley StatsRef:
Statistics Reference Online

Stratified Sampling

By *Van L. Parsons*

Keywords: *sample survey, sampling frame, sample allocation, systematic sampling, implicit stratification, complex survey, health survey, sample design, multistage sampling, cluster sampling*

Abstract: Stratified sampling is a probability sampling method that is implemented in sample surveys. The target population's elements are divided into distinct groups or strata where within each stratum the elements are similar to each other with respect to select characteristics of importance to the survey. Stratification is also used to increase the efficiency of a sample design with respect to survey costs and estimator precision. In this article, the foundations of stratified sampling are discussed in the framework of simple random sampling. Topics include the forming of the strata and optimal sample allocation among the strata. Practical implementation issues for stratified sampling are discussed and include systematic sampling, implicit stratification, and the construction of strata using modern software. The importance of using stratified sampling in practice is demonstrated by its usage in five major large-scale health surveys conducted in the United States and the United Kingdom. For these surveys, details of the stratification and sampling methods are provided. Topics include multistage cluster sampling within strata and the use of systematic and probability proportional to size sampling.

Stratified sampling is a **probability sampling** method that is frequently implemented in sample surveys. In general, a population's elements (or in practice the **sampling frame** members) are divided into distinct groups or strata where within each stratum the elements are similar to each other with respect to select characteristics of importance to the survey. Typically, each stratum is independently sampled using a method for which an **unbiased** estimator of stratum total or stratum mean can be computed. These estimated stratum totals may then be added to obtain an estimator for the population total. Similarly, a stratum-weighted average of the estimated stratum means may be computed to estimate the overall population mean. **Stratification** is used to increase the efficiency of a sample design with respect to cost and estimator precision.

In this article, we discuss the theoretical foundations of stratified sampling. For simplicity, we focus on the most elementary sampling structure, *stratified random sampling*. Here, the population elements are sampled by **simple random sampling**. "Real-life" surveys tend to use stratified multistage cluster sampling (*see* **Cluster Sampling**; **Multistage Sampling**), but most of the elements presented here can be extended to these more complicated structures.

National Center for Health Statistics, Hyattsville, MD, USA

DOI: 10.1002/9781118445112.stat05999.pub2

Two simple examples of stratified populations are as follows:

1. A population of physicians is stratified by state of practice and specialty (e.g., cardiology or neurology). One such stratum is New York cardiologists.
2. A population of hospital-discharged patients in the United States is stratified by region and the size of hospital classified by bed size. A typical stratum could be hospitals in the South with 500 or more beds.

After we establish some basic groundwork on stratification, some actual surveys will be discussed.

1 Basic Foundations for Stratification

The planners of a sample survey usually start by having a set of analytic survey objectives, including domains of study and precision requirements for target estimators. In addition, costs and administrative constraints are an integral part of the survey design process. **Cochran** [1, Section 5.1] discusses four principal reasons why planners consider a stratified design:

1. The population contains subpopulations which are of primary interest to the survey planners. When distinct estimates of known precision are needed for these selected subpopulations, it is advisable to treat each subpopulation as a "population" in its own right. Sample sizes within designated subpopulations may be increased to meet target precision levels.
2. It may be administratively convenient to stratify. An agency conducting a survey may have field offices that may be used to stratify a population. Each field office may be responsible for the survey administration of its part.
3. Distinct groups within a population may differ to such a degree that different sampling procedures are required. In addition, as the population of study may be partitioned by multiple frames with different operational characteristics, a stratified sample may be the only workable option.
4. Stratification may improve the precision of sample estimators of the entire population. It may be possible to divide a heterogeneous population into subpopulations, each of which is internally homogeneous. Sampling variability within each stratum should be much smaller than sampling variability over the population as a whole. Precise estimates of the means of each stratum can be obtained by targeted sample sizes, and then the estimates combined to form an estimate for the entire population. With an appropriate allocation of sample, this estimator will be more precise than one created from the same size sample but without stratification.

A survey's analytic objectives may result in conflicting requirements for stratification. As an example, for most US federally sponsored health care surveys (*see* **Health Services Data Sources in the US**), the primary objective is to assess the overall health and health care needs of the Nation, which suggests planning a stratified design consistent with item 4 above. However, a survey's secondary objectives may target specific subpopulations, for example, geographical areas such as states or population demographic groups, which suggest planning a stratified design consistent with item 1 above. Typically, survey planners implement a mixture of both design methodologies focusing on acceptable precision standards for all survey objectives subject to fixed resources.

1.1 Simple Random Samples and Stratification

The gaining of precision through stratification, as outlined in reason 4 above, can be justified theoretically for many commonly used sampling methods. For simplicity, the mathematical theory presented here will

Wiley StatsRef: Statistics Reference Online, © 2014–2017 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat05999.pub2

focus on simple random sample (SRS) methods and linear estimators; for example, totals and means having a known population total as its denominator.

Suppose that a population of N elements has already been divided into L strata of known sizes N_h, $h = 1, 2, \ldots, L$, and that stratum h contains population elements Y_{hi}, $i = 1, 2, \ldots, N_h$. The true stratum population means and variances are defined as $\overline{Y}_h = \sum_{i=1}^{N_h} Y_{ih}/N_h$ and $S_h^2 = \sum_{i=1}^{N_h} (Y_{hi} - \overline{Y}_h)^2/(N_h - 1)$, respectively, and the true total population mean and variance may be expressed as $\overline{Y} = {}_h \sum_{h=1}^{L} \sum_{i=1}^{N_h} Y_{hi}/N = \sum_{h=1}^{L}(N_h/N)\overline{Y}$, and $S^2 = \sum_{h=1}^{L} \sum_{i=1}^{N_h} (Y_{ih} - \overline{Y})^2/(N-1)$, respectively. Now, if an SRS of size n is taken from the entire population, the typical estimator of $\overline{Y}$ is the sample mean, $\overline{y}_{\text{srs}} = \sum_{j=1}^{n} y_j/n$. For stratified random sampling over L strata, a random sample of size n_h is taken independently within each group. The sample stratum mean, $\overline{y}_h = \sum_{i=1}^{n_h} y_{hi}/n_h$, is calculated and weighted by relative stratum size to define the stratified estimator, $\overline{y}_{\text{str}} = \sum_{h=1}^{L}(N_h/N)\overline{y}_h$. Both estimators are unbiased for the true population mean – that is, $E(\overline{y}_{\text{srs}}) = E(\overline{y}_{\text{str}}) = \overline{Y}$ – and they have respective variances:

$$\text{var}(\overline{y}_{\text{srs}}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} \quad \text{and} \quad \text{var}(\overline{y}_{\text{str}}) = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)$$

Note that for an estimator of population total, one just multiplies the above mean estimators by N and the variances by N^2.

The impact of stratification on survey costs and estimator precision depends on how the sample of n units is allocated to the strata. If the unit cost of sampling varies by some naturally defined strata, few survey planners would consider an unstratified sample. Most surveys are conducted under the constraint of a fixed budget, and the survey planners must have some ability to control sample sizes by stratum costs. For example, before 2016, the National Health Interview Survey (NHIS), a survey which monitors the health of the United States, was based for the most part on an area frame sample (*see* **Area Sampling**). Before household sampling and in-person interviewing, a field representative had to canvas a targeted area and list all dwelling units. This listing process was very expensive. As a cost-saving measure, starting in 2016, the NHIS stratifies targeted sampled areas into those which have high-quality vendor unit-address lists available and into those which still require on-site dwelling listing. This stratification eliminates about 85% of the on-site listing. See **Stratified Sampling, Allocation in** for additional topics on optimal sample allocations with respect to cost and precision requirements.

For surveys in which sample unit costs are identical over strata, both stratified and nonstratified sampling can readily be compared for various allocations of a fixed total sample size n. First, under some mild conditions, it can be shown that the optimal allocation of the total sample size n into L independent samples, n_h, is defined by

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}$$

with S_h the stratum population standard deviation. This is referred to as the **Neyman Allocation**. This is optimal in the sense that $\text{var}(\overline{y}_{\text{Neyman}}) \leq \text{var}(\overline{y}_{\text{str}})$ for any other stratified sample allocation. Intuitively, the largest strata are the most important in estimating the population mean, and more sample should be allocated to the large strata. Moreover, the strata with the largest dispersions require more sample for precise estimation. The proportionality factor $N_h S_h$ can be thought of as a combined measure of these two intuitive concepts. In practice, a true Neyman allocation is rarely implemented because the true S_h are usually unknown. Instead, survey planners use a variable that has a close relationship with the variable of interest. For example, in planning a survey of hospital-discharged patients, a sampling frame hospital is stratified in part by geographic location, type of hospital, and bed size. Bed size is positively correlated with the number of discharges that a hospital produces and can be used as a proxy for the number of discharges.

DOI: 10.1002/9781118445112.stat05999.pub2

If population means for patient discharges are the focus of estimation, the strata hospital counts and bed size standard deviations can be used to define a Neyman allocation.

Another commonly used stratified sampling method is proportional sampling, where $n_h = n(N_h/N)$. In practice, this method is easy to implement provided that the true sizes, N_h, of the strata are known. If the S_h do not deviate much by strata, then proportional sampling will be close to the Neyman optimal. If the strata sizes N_h are large, then the following relationships among the variances for Neyman, proportional, and unstratified SRS sampling can be established:

$$\text{var}(\overline{y}_{\text{srs}}) = \text{var}(\overline{y}_{\text{prop}}) + \frac{(1 - n/N)}{n} \times \sum_{i=1}^{L} \left(\frac{N_h}{N}\right) (\overline{Y}_h - \overline{Y})^2$$

with

$$\text{var}(\overline{y}_{\text{prop}}) = \left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right) \sum_{h=1}^{L} \left(\frac{N_h}{N}\right) S_h^2$$

and

$$\text{var}(\overline{y}_{\text{prop}}) = \text{var}(\overline{y}_{\text{Neyman}}) + \frac{1}{n} \sum_{i=1}^{L} \left(\frac{N_h}{N}\right) (\overline{S}_h - \overline{S})^2$$

From the abovementioned two equations, one sees that the proportional allocation will provide a sampling variance at least as small as an unstratified SRS. The greatest sampling efficiencies occur when the true stratum means vary to a large degree. Neyman allocation will reduce the variance by a factor proportional to the variance of the stratum standard deviation. It should be noted that for a fixed proportion allocation on the same strata, both $\text{var}(\overline{y}_{\text{prop}}) \leq \text{var}(\overline{y}_{\text{srs}})$ and $\text{var}(\overline{x}_{\text{prop}}) \leq \text{var}(\overline{x}_{\text{srs}})$ for different characteristics y and x. This is an attractive property of stratified proportional sampling. For a Neyman allocation determined by a variable y, but also used for an other unrelated survey variable x, it is possible that $\text{var}(\overline{x}_{\text{Neyman}[y]}) \geq \text{var}(\overline{x}_{\text{srs}})$. Such a phenomenon may occur if $S_h(x)$ and $S_h(y)$ are negatively correlated. For example, in a population stratified by income level, a Neyman allocation targeted to estimate occupational work-loss days would probably be quite inefficient for estimating health insurance coverage for the unemployed. Thus, while proportional sampling may not be optimal, it would be a safe strategy to use when several different variables are to be estimated using the same sample. The above discussion has assumed that a mean or total for an entire population is the target of estimation.

While stratified proportional sampling reduces the variance relative to unstratified SRS, Kish [2, Section 3.4] points out that, in practice, the relative gains may be only small or moderate. This is because survey planners do not have population variables available to define a highly efficient stratification. A special case of interest is the impact of stratification on estimating a population proportion, p. Here, the population variance is $p\,(1-p)$, and the stratum variance becomes $S_h^2 = p_h(1 - p_h)$, with p_h the stratum proportion. The nature of this variance makes it somewhat insensitive to stratification if the resulting strata p_hs are in the central range 0.20–0.80. However, for stratified cluster sampling, typical in major surveys, the efficiency gains for proportional sampling are greater than the element sampling discussed here.

If the strata themselves are of interest and comparisons are to be made among strata, then the individual stratum estimates, $\overline{y}_h$, and not the aggregate, $\overline{y}_{\text{str}}$, are most important in meeting precision requirements. Equal allocation of sample to strata, $n_h = n/L$, could be used. In such cases, especially when the strata sizes, N_h, vary greatly in size, $\text{var}(\overline{y}_{\text{str}})$ may exceed $\text{var}(\overline{y}_{\text{srs}})$. For several computationally detailed examples of the stratified random sampling method, the reader is referred to Levy and Lemeshow [3, Chapters 5, 6].

Also important is the subdomain or subpopulation estimator where the subdomain does not necessarily coincide with the stratum. Frequently, subdomains dictate the main precision requirements of a survey. The estimation of the mean of a population subdomain is a special case of ratio estimation (*see* **Ratio and**

Wiley StatsRef: Statistics Reference Online, © 2014–2017 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat05999.pub2

Regression Estimates). Here, the target population parameter is

$$\overline{Y}_{\mathrm{D}} = \frac{\sum_{i=1}^{N} Y_i \delta_i}{\sum_{i=1}^{N} \delta_i}$$

where $\delta_i = 1$ if unit i is in subdomain D and 0 if not. A combined ratio estimator is $\overline{y}_{\mathrm{D,strc}} = (\overline{yd})_{\mathrm{str}} / \overline{d}_{\mathrm{str}}$, where the $(\overline{yd})_{\mathrm{str}}$ and $\overline{d}_{\mathrm{str}}$ are the stratified estimators of the numerator and denominator of $\overline{Y}_{\mathrm{D}}$. For this **ratio estimator**, the denominator may be random, and the **linearization method** may be used to derive approximate variance formulas.

1.2 Departures from Simple Random Samples and Direct Stratification

In many large-scale surveys, especially those that use multistage cluster sampling with probability proportional to size (PPS) sampling (*see* **Probability Proportional to Size (PPS) Sampling**; **Sampling With Probability Proportional to Size**), the process of creating multiple levels of stratification and the use of PPS samples can become quite involved. Instead, a process called *implicit stratification* along with **systematic sampling** is frequently used [4, Chapter 2] (*see* **Systematic Sampling**; **Sampling in Developing Countries**; **Systematic Sampling Methods**). For example, government agencies and data vendors often maintain large sampling frames containing household addresses where each address is identified by hierarchical levels of geography, for example, US county and block, and associated social–economic–demographic (SED) information corresponding to those levels, for example, block-level urban status, minority population proportion, and **median** income. A survey planned for a metropolitan area covered by this frame can sort the frame hierarchically by the SED variables. This leads to a three-layer implicit stratification. A systematic sample, say 1 of every 50 households, can be taken from the ordered frame to obtain a sample of households. This method is very flexible, easy to implement and has frequently characterized sampling in many National Center for Health Statistics (NCHS) surveys. The expected systematic sample sizes are proportional to the total size within any sort level. For variance estimation purposes, coarse levels containing large samples are often treated as stratified proportional samples. This facilitates analysis using conventional software (*see* **Software for Sample Survey Data**).

2 Real Examples of Stratified Sampling for Health Surveys

Following are four examples taken from the NCHS family of surveys and an example of a United Kingdom national survey. They are typical of the large-scale surveys conducted by government agencies to produce official statistics. The somewhat involved design structures of the NCHS surveys, as with most large-scale surveys, are here somewhat simplified to emphasize the fundamentals, in particular, the importance of stratified sampling in major health surveys. The reader should refer to the NCHS references[5] to get more thorough descriptions and timely updates for any specific survey.

2.1 Example 1: National Maternal and Infant Health Survey (NMIHS)

While this survey was last conducted in 1988, its data are available, and the survey provides an example of sampling administrative records from strata. One component of the National Maternal and Infant Health Survey (NMIHS) targets mothers who have recently given live birth or experienced fetal or infant death. Sample mothers are contacted by mail or telephone whenever possible. With this mode of data collection,

DOI: 10.1002/9781118445112.stat05999.pub2

certificates of live births, certificates of infant death, and official reports of fetal death are used as sampling frames. These frames are stratified and sampled as follows. As birth and death records are collected and processed within a registration area, typically a state, these areas define an imposed administrative stratification. Mothers of low-**birthweight** infants and black infants are specific targets of investigation. To meet these two study objectives, the live-birth certificates are partitioned by black and nonblack statuses and then cross-classified by the infant's birthweight to form race–birthweight strata within each state. SRSs are taken within each stratum, but to meet precision goals for low-birthweight and/or black infants, these particular strata are oversampled, that is, $n_B/n > N_B/N$, where B is any oversampled stratum. The reader is directed to Ref. 6 for additional discussion.

2.2 Example 2: National Ambulatory Medical Care Survey (NAMCS)

For the National Ambulatory Medical Care Survey (NAMCS) design, the target population consists of physician–patient visits to nonfederal offices of physicians engaged in office-based practice in a given year. (The NAMCS also includes a community health center visit component that will not be discussed in this article.) For the 2012 main sample, in addition to the Nation, the 34 largest states and physician specialties are targeted for precise estimation. The NAMCS focuses on meeting the stratification goals of items 1 and 4 listed in Section 1. Physicians are stratified by 9 Census divisions, the 34 most populous states as nested within divisions, and primary care status, that is, whether or not the physician is in a primary care specialty such as internal medicine or pediatrics.

Physician metropolitan statistical area location status and 18 broad specialty groups and then individual specialties within the groups are used as implicit stratification variables, and systematic sampling is used to select physicians. The physician sample is randomly divided into 52 weekly samples. This stratification will allow for precise state-level estimation for the 34 largest states and the 9 Census divisions of the United States. The mode of data collection for this survey involves an interviewer conducting a computer-assisted personal interview with the physician, selecting systematic random samples from chronological lists of visits made to the physician, and abstracting data from patient visit records. Small practices have all patient visits sampled, and systematic samples of patient visits are taken from larger practices[7].

2.3 Example 3: National Health Interview Survey (NHIS)

Since 1957, the NHIS has been a major health survey conducted annually to produce health-related statistics for the US noninstitutionalized population. As of this writing, the 2006–2015 NHIS is the most recent source of data and is the focus of the following discussion. The primary objective of the NHIS is to produce statistics for the Nation, while secondary objectives are producing statistics for age–race/ethnic–sex domains and tertiary objectives are producing statistics for most of the states. The NHIS design focuses on meeting the stratification goals of items 1 and 4 listed in Section 1. The survey contains about 37 000 interviewed households and about 96 000 persons, and it targets numerous health variables on age–race/ethnic–sex domains for specified precision levels. The mode of data collection involves a **face-to-face** interview that requires a stratified multistage cluster sample to be cost effective. The target population is not directly available as a sampling frame but is covered for the most part by two distinct frames. An area sample frame consists of relatively compact geographical units, US Census-defined blocks, which contain dwelling units covered by the most recent US Decennial Census; this frame includes most of the target population. To keep the coverage current, a frame consisting of places where new residential housing has been constructed since the last Decennial Census is also created. The former frame contains Decennial Census SED information, but the latter frame contains

Wiley StatsRef: Statistics Reference Online, © 2014–2017 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat05999.pub2

only location. These two frames are strata as discussed in item 3 of Section 1. At a coarse geographical level, the United States is partitioned into primary sampling units (PSUs), counties (or equivalents), which contain components of the two major frames. The PSUs are stratified within each state by metropolitan status and population size to form primary strata. Within each PSU, aggregates of Census-defined area frame blocks, called *segments*, are classified into minority population substrata defined by the density of Hispanic, black and Asian populations; the segments within substrata have the same expected number of dwelling units. A substratum containing the new construction frame is also created within the PSU.

At the first stage of sampling, the very large population PSUs – for example, New York City and Los Angeles – are treated as self-representing primary strata. For non-self-representing strata one or two PSUs are selected from each stratum, with probability **proportional to population size**. At the second stage of sampling, a systematic sample of segments is taken from each minority population density substratum within a sampled PSU. At the third stage, all minority households are sampled within each segment, while a systematic sample of the complement households is taken. The end result is that the total minority population elements are in a much greater proportion in the sample than in the population. This oversample allows for greater precision for estimation on small minority age–sex subdomains[8].

2.4 Example 4: National Health and Nutrition Examination Survey (NHANES)

The National Health and Nutrition Examination Survey (NHANES) series of surveys conducted since 1971 assess the health of the noninstitutionalized population. The collected data are much more in-depth than the NHIS in that subjects may be given a complete physical examination at a specially designed mobile examination center. As of this writing, the survey is conducted in a 4-year cycle, each year having a sample of 15 locations. Each annual location will be visited by one of the three mobile examination centers that travel across the country to sampled locations. The NHANES has an annual target of examining 5000 persons. Data are released at the 2-year time period and 2-year cycles can be combined to increase precision. NHANES planners have the objectives of measuring hundreds of health-related variables for both the Nation and a large number of targeted different age–sex–race/ethnic-income subdomains of the U.S. population. Objectives for the 2011–2014 design led to 87 targeted subdomains with an emphasis on Hispanic and non-Hispanic black and Asian populations. Thus, the NHANES design focuses on meeting the stratification goals of items 1 and 4 listed in Section 1.

As the sampling plan requires mobile examination centers to be at 15 fixed locations for a 9-week period and for survey respondents to travel to the examination center, the geographical unit of "US County" is used as a cost-effective and respondent-convenient PSU definition. The PSUs are assigned measures of size based on weighted averages of estimated race-ethnic-income population sizes. To cover 4 years of sample, 60 locations for visits by a mobile examination center are assigned to PSUs sampled from the universe via PPS sampling. In the 2011–2014 design, six of the largest PSUs are in sample with certainty. Five of these PSUs are in sample for 1 of the 4 years, while the largest certainly PSU supports three distinct locations and is in sample for 3 of the 4 years. Overall, the certainty PSUs contain eight locations for a mobile examination center. The balance of PSUs are noncertainty and are stratified into five noncontiguous state groups by state health rankings derived from death rate, infant mortality rate, percentage of adults with high blood pressure, percentage of adults overweight or obese, percentage of adults with poor nutrition, and percentage of adults who smoke. Within each of the five strata, substrata are formed by census region and by percentage of population living in rural areas. These stratification procedures define 13 major strata for PSU sampling for the 4-year NHANES cycle. The major strata are then substratified by demographics to form a total of 52 minor strata. One PSU is sampled PPS from each minor stratum. The four minor strata per major strata are allocated across the 4-year design. One minor stratum from each major stratum is designated as a mobile examination location for about a 9-week period each year. Within PSUs, all the

DOI: 10.1002/9781118445112.stat05999.pub2

geographical areas are partitioned into clusters of census-defined blocks called *segments*. These segments are stratified implicitly by minority density and geography and then sampled systematically. Procedures for sampling dwelling units within segments and then persons within dwelling units are somewhat involved, and the reader is referred to the main NHANES document[9].

2.5 Example 5: Children's Dental Health Survey (CDH)

The Children's Dental Health (CDH) is a system of national surveys in the United Kingdom to monitor the dental health and dental health care for children. This survey has been fielded every 10 years since 1973, and in 2013, the survey targeted children of ages 5, 8, 12, and 15 years being educated in mainstream state and independent schools in England, Wales, and Northern Ireland. These age cohorts are of analytical interest because the dentition and dental health for children in these cohorts are expected to differ. The major goals of the CDH are to produce reliable statistics for each targeted age cohort within each of the three countries along with a specific goal of making comparisons between low-income school children and others in the same age cohort. The survey mode of data collection includes a personal dental examination, thus requiring a geographically clustered sample to keep field operations and costs manageable. A sample of about 10 000 dental exams is planned with about 2500 dental exams in each of the four age cohorts.

For both sampling and operational efficiency, the CDH requires a multistage cluster sampling design. England, Wales, and Northern Ireland are explicit strata with England and Wales being partitioned into 9 and 3 substrata, respectively, by Region. Each Region consists of Local Authority Districts (LADs) (England) or Unitary Authorities (Wales). For the CDH, these units will be referred to as LADs. (As design features, the LADs have a similar survey role as does the geographical PSU in the U.S.'s NHIS and NHANES surveys.) Each LAD has a measure of size determined by the number of school pupils in the targeted age cohorts adjusted by a "deprivation" (schools associated with lower socioeconomic status pupils) weighting factor. This factor inflates the sizes of schools with designated higher deprivation scores and allows for the oversampling of low-income children. Using a PPS systematic sample along with an implicit stratification of Region and LADs within each country, 81 sample LADs are selected from England and 27 from Wales. Within each sampled LAD, the schools are stratified into groups (clusters) of primary schools and groups of secondary schools with "deprived" schools grouped together. In a manner similar to the LAD assignment of total size, each group of schools within its substratum is given a measure of size determined by the number of school pupils in the targeted age cohorts adjusted by a "deprivation" weighting factor. As in the previous selection stage, PPS samples within each LAD's primary school and secondary school strata are taken, resulting in 81 sampled primary school groups and 81 sampled secondary school groups in England; Wales has corresponding sample counts of 27 school groups. Within each school group, a sample of schools is selected using simple random sampling. At the last stage, within each school, pupils are selected using systematic random sampling. In Northern Ireland, the schools are less geographically dispersed and additional clustering is not needed. Schools are explicitly stratified by primary and secondary status, and the "deprived" schools are given oversampling weighing factors. Schools are implicitly stratified and sampled using PPS systematic sampling. Pupils are selected by systematic sampling. The reader is referred to the web page[10] for comprehensive documentation on the CDH.

3 Defining Strata

The discussion so far has considered sampling methods while treating the strata as already given, but the construction of the strata themselves is also important. Some basic issues involve the selection of

Wiley StatsRef: Statistics Reference Online, © 2014–2017 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat05999.pub2

stratification variables, the boundaries between strata, and the number of strata to use [11, Chapter 12]. Most optimality results cannot be implemented in practice, owing to limited population information, conflicting design objectives, cost considerations, and administrative restrictions. Frequently, survey planners consider several stratified sampling options with respect to cost and precision to determine a design that will perform well over a wide spectrum of target variables; the selected design may not be optimal for any given variable.

Some stratification boundary defining rules have been studied under theoretic conditions. The best known is the *cum* $\sqrt{f}$ *rule*. This rule states that if y is a continuous variable, and $f(y)$ is the density function of y with support $[A_L, A_U]$ then an approximate optimal stratification of the population into H units with boundaries $A_L = a_0 < a_1 < a_2 < \ldots < a_{H-1} < a_H = A_U$, each to be sampled by the Neyman allocation, would result from creating the H strata in such a way that

$$\int_{a_{h-1}}^{a_h} [f_Y(y)]^{1/2} \mathrm{d}y = \frac{1}{H} \int_{A_L}^{A_U} [f_Y(y)]^{1/2} \mathrm{d}y$$

That is, each stratum accounts for $1/H$ of the total integral of $\sqrt{f}$. In practice, a variable related to y would be used. Using a histogram approach, Cochran [1, Section 5.A.7] provides a computational example of this method.

For element sampling, Kish [2, Section 3.6I] and Cochran [1, Section 5.A.8] suggest keeping the number of strata modest in size. This suggestion is based in part on some simple theoretic structures. Let y be a **linear regression** on x, with ρ the correlation between y and x in the unstratified population. If the population is partitioned into L strata defined by the variable x, using the optimal strata boundaries along with sample sizes of n/L in each stratum, then

$$\frac{\mathrm{var}(\bar{y}_{\mathrm{str}})}{\mathrm{var}(\bar{y}_{\mathrm{srs}})} \geq \left[\frac{\rho^2}{L^2} + \left(1 - \rho^2\right)\right]$$

As L becomes large, the lower bound tends to $(1 - \rho^2)$, and a point of little return in variance reduction can be established. For $\rho < 0.95$, little reduction occurs for more than six strata. This argument would assume that, total population estimators rather than individual stratum estimators are important.

Advancements in computer hardware and software have made the construction of optimal strata and sample allocation a less formidable task than it was in the past. For univariate stratification, Baillargeon and Riverst[12] discuss recently developed algorithms and also provide an R-package (*see* **R**). For multivariate stratification and target variables, Ballin and Barcaroli[13] discuss the use of a "genetic algorithm" (*see* **Genetic Algorithms**) to obtain optimal stratification and sample allocation. An R-package is also provided.

4 Other Stratification Issues

4.1 Stratification After Sampling

Frequently, variables well suited for partitioning the population into strata do not exist before sampling. In this case, an estimation technique, called **poststratification**, can be used. Here, the sampled data are stratified after sampling and then an estimator for population mean or total is created as if the sample had come from a presample stratification. For example, in the case that an SRS of size n is taken from an unstratified population, the sample is first poststratified into H strata, with sample stratum means $\bar{y}_g$, for $g = 1, 2, \ldots, H$, and then a poststratified mean estimator is defined: $\bar{y}_{\mathrm{pstr}} = \sum_{g=1}^{H} (N_g/N)\bar{y}_g$, where N_g/N are the population totals of the poststratification classes (see Ref. 14, Section 11.6).

DOI: 10.1002/9781118445112.stat05999.pub2

While the functional forms of the stratified and poststratified mean appear identical, there are some important distinctions as to implementation and statistical properties. First, selected poststratification classes must have known population sizes (or independently known accurate estimates of size). For example, in large-scale surveys, age–race–sex classes are frequently used for poststratification, as the US Bureau of the Census produces very accurate national tabulations, which it updates quarterly. However, a poststratification on a health status variable would be difficult, as the true class totals could not be obtained. Second, the sample sizes, n_g, observed within the poststratification cells are themselves random variables. If these sample sizes are reasonably large, perhaps having $n_g > 20$ in each class, then this method is almost as precise as the proportional stratified sampling discussed earlier. This method can be extended to more complicated sampling schemes.

4.2 Stratification and Variance Estimation

Well-designed surveys plan a method to compute approximately unbiased estimators of variance for the basic total and mean estimators (see Ref. 15, Chapter 2). In large-scale multistage cluster samples, some strata may have only one sampled cluster. This is problematic as most traditional variance estimators require at least two clusters per stratum. For such cases, collapsed strata may be created for variance estimation (see Ref. 15, Chapter 5). Original strata may be collapsed by combining strata with similar stratum characteristics. The sampled clusters within a collapsed stratum are treated as having been sampled with replacement (*see* **Sampling With and Without Replacement**). This step is typically done by survey planners, but modern statistical software for complex surveys (*see* **Software for Sample Survey Data**) frequently has corrective procedures for such situations.

Acknowledgments

The author thanks Iris Shimizu, Ryne Paulose, Jennifer Parker, and Te-Ching Chen for the valuable discussions on the NAMCS and NHANES designs.

Related Articles

Multistratified Sampling; **Stratified Multistage Sampling**; **Agricultural Surveys**; **Percentage**; **Sampling Agricultural Resources**.

References

[1] Cochran, W.G. (1976) *Sampling Techniques*, 3rd edn, John Wiley & Sons, Inc., New York.
[2] Kish, L. (1965) *Sampling Techniques*, 3rd edn, John Wiley & Sons, Inc., New York.
[3] Levy, P.S. and Lemeshow, S. (2008) *Sampling of Populations: Methods and Applications*, 4th edn, John Wiley & Sons, Inc., Hoboken, NJ.
[4] Lehtonen, R. and Pahkinen, E. (2003) *Practical Methods for Design and Analysis of Complex Surveys*, 2nd edn, John Wiley & Sons, Ltd, Chichester.
[5] NCHS (2016) *http://www.cdc.gov/nchs/surveys.htm* (retrieved 5 July 2016).
[6] NMIHS (2015) *http://www.cdc.gov/nchs/nvss/nmihs.htm* (retrieved 5 July 2016).
[7] NAMC (2015) *2012 NAMCs Micro-data file documentation*. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NAMCS/doc2012.pdf* (retrieved 5 July 2016).

Wiley StatsRef: Statistics Reference Online, © 2014–2017 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat05999.pub2

[8] Parsons V. L., Moriarity C., Jonas K., Moore T., Davis K. E., Tompkins L., 2014, Design and estimation for the National Health Interview Survey, 2006–2015, *Vital Health Stat.* **2**(165), *http://www.cdc.gov/nchs/data/series/sr_02/sr02_165.pdf*.
[9] Johnson C. L., Dohrmann S. M., Burt V. L., Mohadjer L. K., 2014, National Health and Nutrition Examination Survey: sample design, 2011–2014, *Vital Health Stat.* **2**(162), *http://www.cdc.gov/nchs/data/series/sr_02/sr02_162.pdf*.
[10] CDH (2015) *http://content.digital.nhs.uk/catalogue/PUB17137* (retrieved 15 September 2016).
[11] Särndal, C.E., Swensson, B., and Wretnman, J.H. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York.
[12] Baillargeon, S. and Rivest, L.-P. (2011) The construction of stratified designs in R with the package *stratification*. *Surv. Methodol.*, **37**, 53–65.
[13] Ballin, M. and Barcaroli, G. (2013) Joint determination of optimal stratification and sample allocation using genetic algorithm. *Surv. Methodol.*, **39**, 369–393.
[14] Thompson, S.K. (2012) *Sampling*, 3rd edn, John Wiley & Sons, Inc., Hoboken, NJ.
[15] Korn, E.L. and Graubard, B.I. (1999) *Analysis of Health Surveys*, John Wiley & Sons, Inc., Hoboken, NJ.

DOI: 10.1002/9781118445112.stat05999.pub2