Chapter 4
A Practical Approach to Sampling

The major strength of probability sampling is that the probability selection mechanism permits the development of statistical theory to examine the properties of sample estimators. [. . .] The weakness of all nonprobability methods is that no such theoretical development is possible; as a consequence, nonprobability samples can be assessed only by subjective valuation.

—Graham Kalton,
Introduction to Survey Sampling

If proper sampling procedures are followed, in a matter of days, an opinion poll conducted on approximately 1,000 individuals can be reasonably taken as a measure of public opinion for a population as large as China's. This is the power of sampling, its ability to approximate from a small group the characteristics of the whole population within a know margin of error.

Different methods of respondents' selection can be employed when conducting a survey, that is, interviewing experts, the typical respondent, or a group of respondents. Because only part of the population is sampled, the estimated parameters are subject to a sampling error. Sampling error is a measure of "how closely we can reproduce from a sample the results that would be obtained if we should take a complete count or census" (Hansen, Hurwitz, and Madow 1953, 10).

The ability to estimate and reduce this error depends on how the sample is selected. If the researcher knows what chances each population member has to be included in the sample, he can use statistical theory to estimate the properties of the survey statistics. On the contrary, when the selection of respondents is based on personal judgment, it is not possible to have an objective measure of the reliability of the sample results (Kalton 1983). This is not to say that there is no room for subjective judgment in probability sampling. Rather, subjective judgment plays an important role in sample design as long as the final selection of the sample elements is left to a random process (Hansen, Hurwitz, and Madow 1953).

Volumes have been written on probability sampling. Hence, far from being a discussion on sampling techniques, what follows is a short review of how to determine the sample size using four of the most commonly used sampling procedures: simple random sampling (SRS), stratified random sampling, systematic sampling, and probability proportional to size (PPS) sampling. Particular attention is dedicated to how to deal with frame problems and how to perform weight adjustments. Finally, a practical case of stratified random sampling of manufacturing establishments is presented.[1]

Determining the Sample Size in Simple Random Sampling

The simplest form of probability sampling is SRS. With this method, every possible sample of equal size—as well as each individual element[2] in the population (see box 4.1)—has the same non-zero probability of being selected: the *epsem* (equal probability of selection) design.[3]

The sample size in SRS depends on three factors:

- The population size
- The variability of the parameter we wish to estimate
- The desired level of precision and confidence we wish to reach.

If we are interested in estimating the population mean of the parameter $\overline{Y}$ with precision e_0 and confidence α, the minimum sample size is then determined by the following formula:

$$n = \frac{z_{\frac{\alpha}{2}}^2 S^2}{e_0^2 + z_{\frac{\alpha}{2}}^2 \frac{S^2}{N}}$$

Similarly, if we are interested in estimating the population proportion $\overline{P}$ for a given characteristic with precision e_0 and confidence α, the minimum sample size is calculated as follows:

$$n = \frac{z_{\frac{\alpha}{2}}^2 P(1-P)}{e_0^2 + z_{\frac{\alpha}{2}}^2 \frac{P(1-P)}{N}}$$

where:

N = population size
n = sample size
S^2 = population variance of Y (assumed to be known)

[1] I am grateful to Dr. Mohammed Yusuf from Survey and Research Systems (Dhaka, Bangladesh) for comments and suggestions on this chapter.

[2] Given the peculiarities of business activities, an important problem in business surveys is the identification of the sample element (see box 4.1).

[3] As commonly done in the literature, we refer here to SRS without replacement.

Box 4.1

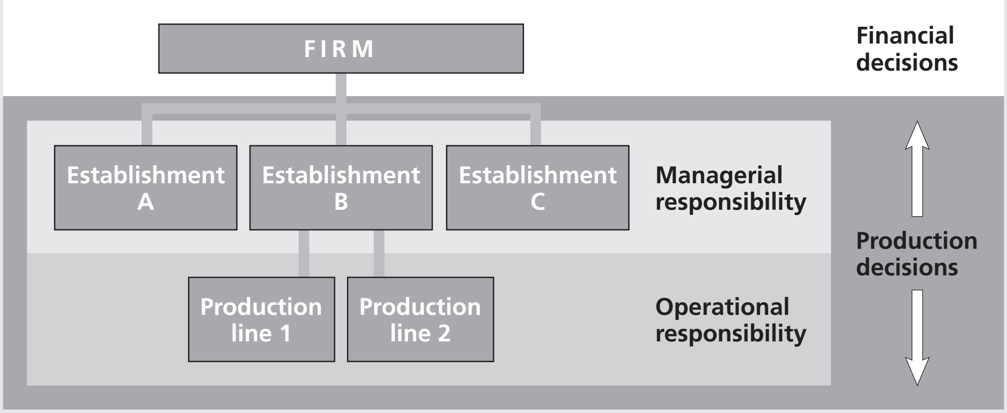
The Sampling Unit in Business Surveys

Contrary to household surveys in which the identification of the sampling unit is easier (that is, husband, wife, and so on), in surveys of business activities this task is complicated by the fact that there is an array of business forms, from small family-owned stores to large international corporations. So, when designing a sample of formal manufacturing activities, what is the best unit of analysis? Is it the establishment, the factory, the plant, the company, the enterprise, or the firm? Theoretical motivations related to the purpose of the study and the desired level of homogeneity, as well as practical considerations on the availability and accuracy of the data, will dictate the answer.

While this identification is not an issue for small entities, large businesses often include different legal structures, operational structures, and ownership structures. They produce different goods, in different locations, at different scales. The resulting heterogeneity in the structure of each sampling unit makes it impossible to have an internationally recognized standardization. Nonetheless, from a theoretical point of view, two types of statistical units are generally identified. One corresponds to the level where financial decisions are made and the other corresponds to the level where production decisions are taken. If the primary interest of the researcher is the behavior related to resource allocation, then the former level is the appropriate unit of investigation (the *firm* in box figure 4.1.1).

Box Figure 4.1.1.

Unit of Analysis in a Business Survey



Source: Author's creation.

(*continued*)

Box 4.1 (continued)

When, on the contrary, decisions on the purchase of factors of production are of analytical interest, then the latter level of analysis should be adopted. Furthermore, within this level, two sublevels can be identified. The level at which managerial decisions are made regarding the whole production process and the level of single product operation. When the object of analysis is the overall managerial responsibility, then the *establishment* is the appropriate level of investigation. If the technological characteristics of a single production process are of interest, then the individual *production line* is the appropriate unit of reference (Nijhowne 1995).

From a practical point of view, while it is relatively easy to acquire firm data to analyze financial decisions, it is much harder to collect information on individual product lines. Although records are kept to support management decisions, it is quite unlikely to find the same level of detail at all levels of production lines in any business unit. The only way to gather these data is for the researcher to reconstruct them from aggregate values, a process that results in an inevitable loss of accuracy (Colledge 1995).

P = population proportion of Y (assumed to be known)
e_0 = desired level of precision
α = desired level of confidence (that is, for instance, 95%)
$z_{\alpha/2}$ = z distribution corresponding to α level of confidence[4]

With SRS, the population mean and population proportion as well as the corresponding variances are estimated as follows:

Parameter Estimated	*Sample Mean*	*Sample Variance*
Population Mean		
$\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$	$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{n}$	$\operatorname{var}(\bar{y}) = \frac{(1-f)}{n} s^2$
Population Proportion[5]		
$\bar{P} = \frac{\sum_{i=1}^{N} y_i}{n}$	$\bar{p} = \frac{\sum_{i=1}^{n} y_i}{n}$	$\operatorname{var}(\bar{p}) = \frac{(1-f)}{n-1} p(1-p)$

[4] See appendix 4 for values of the $z_{\alpha/2}$ corresponding to different levels of confidence.
[5] Where $y_i = 1$ or $y_i = 0$ if the ith element has the desired characteristic or not, respectively.

where the variance of the sample elements is given by the following:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

and $f = \frac{n}{N}$ is the sampling fraction and, in SRS, the probability of inclusion.

Finally, the weight of each element is as follows:

$$\omega_i = f^{-1} = \frac{N}{n}$$

An example of SRS is presented in Box 4.2.

Determining the Sample Size in Stratified Sampling

The efficiency of the sample design can be improved by exploiting any available information on the population under study. In particular, when some population characteristics related to the variable estimated are known, this information can be used to divide the whole population into groups or strata, each sampled separately.[6] This process of stratification increases the efficiency of the design the greater the homogeneity of the elements belonging to the same group. This homogeneity, however, must refer to the characteristic that is being estimated, not to the variable used to identify the strata (Hansen, Hurwitz, and Madow 1953).

In stratified sampling, three different methods can be followed (Kish 1965; Sukhatme, Sukhatme, and Sukhatme 1984).

Method 1. Equal Allocation

The sample size is allocated equally among strata:

$$n_h = \frac{n_o}{H}$$

where:

H = number of strata
h = stratum, with $h = 1,2,3, \ldots h \ldots H$
n_h = stratum sample size
n_o = desired sample size

[6] See the *Productivity and Investment Climate Surveys (PICS): Implementation Manual* (World Bank 2003, 18–22) for information about how to identify strata in Investment Climate Surveys.

Box 4.2

Advising a Mayor

Suppose you are an advisor to a mayor. An opinion poll conducted by a prominent newspaper of 400 residents shows that 54 percent of residents are in favor of a new development program. Because elections are approaching and the mayor wishes to run for reelection, would you advise him to support the project in his political campaign?

The mean result from the poll (54%) is not sufficient to ensure that the majority of residents support this project. We need to determine the implied level of confidence and precision of the poll results. Basically, we need to work backward from the equation used to determine the sample size with proportions under SRS, calculating the implied error e_0 for a given α.

$$n = \frac{z^2_{\frac{\alpha}{2}} P(1-P)}{e_0^2 + z^2_{\frac{\alpha}{2}} \frac{P(1-P)}{N}}$$

Because we want to be certain of our results, we can assume that the level of confidence is 95 percent. The implied error is then given by the following:

$$e = \sqrt{\frac{(N-n) \times \left[z^2_{\frac{\alpha}{2}} \times P \times (1-P) \right]}{n \times N}}$$

Assuming the population of resident voters is 650,000, the implied error is 6.4 percent. This means that we cannot be sure that the majority of the residents support this project, because the true proportion of the residents in favor of the project falls between 49.1 percent and 58.9 percent.

Suppose the mayor wants to know whether he should put this project on his reelection campaign. He asks you to conduct an opinion poll to determine the true proportion of the city population for or against it. How would you arrive at the answer?

We need three pieces of information to determine the sample size. First, we need the target population, which we know is 650,000 resident voters. Second, we need the level of error and desired confidence. Because the major wants to be reasonably certain, we set the error at +/– 3 percent, and the level of confidence at 95 percent. Finally, we need the variance of the true proportion. To be on the safe side, we can assume that the residents are equally split between supporters and opponents. This implies the maximum variance possible. In other words, we assume the worst possible scenario. This implies that we will select the highest sample size for the desired level of precision and confidence.

Box 4.2 (continued)

The required sample size is then given by the equation on proportions. That is

$$n = \frac{z^2_{\alpha/2} \times P \times (1-P)}{e_0^2 + z^2_{\alpha/2} \times \frac{P \times (1-P)}{N}} = \frac{1.96^2 \times 0.5 \times (1-0.5)}{0.03^2 + 1.96^2 \times \frac{0.5 \times (1-0.5)}{650000}} = \frac{3.8416 \times 0.25}{0.009 + 3.8416 \times \frac{0.25}{650000}} = 1065$$

Hence, you need to conduct a survey of 1,065 residents to be able to determine with 95 percent confidence what they think about this project within a 3 percent margin of error. If the survey results show that at least 54 percent of respondents are in favor of the project, we can be sure that most residents are in favor of the project and we can advise the major to support it during his campaign.

Method 2. Proportionate Allocation

The sample size is allocated proportionally to the size of each stratum:

$$n_h = n_o \frac{N_h}{N}$$

where:

N_h population size of the *h*th stratum and $\sum_{h-1}^{H} N_h = N$

Method 3. Optimum Allocation

The sample size is allocated among strata to reach the desired precision at the minimum cost or to reach the maximum precision for a given cost. In this case, depending on the level of precision desired or amount of resource available, four different cases can be envisaged:

Case 1. Given a desired sample size, n_o, and assuming that the unit cost across strata is the same,[7] the optimum allocation n_h across strata is determined by the following formula:

$$n_h = n_o \frac{W_h S_h}{\sum_{h=1}^{H} W_h S_h}$$

[7] Cost per unit in the *h*th stratum c_h is constant across strata:

$$C = \sum_{h=1}^{H} c_h n_h = c \sum_{h=1}^{H} n_h = cn.$$

known as “Neyman” allocation, where:

$$W_h = \frac{N_h}{N} \text{ stratum weight}$$

$$S_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left(y_{ih} - \bar{y}_h\right)^2 \text{stratum variance}$$

Case 2. Given the desired level of precision, e_o, and assuming that the unit cost across strata is the same, the optimum sample size for each stratum is determined by the following formula:

$$n_h = \frac{W_h S_h \sum_{h=1}^{H} W_h S_h}{V_0 + \frac{1}{N}\sum_{h=1}^{H} W_h S_h^2}$$

and the total minimum sample size required is given by:

$$n_{\min} = \frac{\left(\sum_{h=1}^{H} W_h S_h\right)^2}{V_0 + \frac{1}{N}\sum_{h=1}^{H} W_h S_h^2}$$

where:

$$V_0 = \left(\frac{e_o}{z_{\alpha/2}}\right)^2$$

Case 3. Given the amount of resources available to conduct the survey, C_o, and assuming that the unit cost across strata is variable,[8] the optimum allocation of each n_h is determined by the following formula:

$$n_h = \frac{C_o \frac{W_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{H} W_h S_h \sqrt{c_h}}$$

[8] Cost per unit in the *h*th stratum c_h is different across strata:

$$C = \sum_{h=1}^{H} c_h n_h$$

and the total sample size is as follows:

$$n = \frac{C_o \sum_{h=1}^{H} \frac{W_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{H} W_h S_h \sqrt{c_h}}$$

Case 4. Given the desired level of precision, e_o, and assuming that the unit cost across strata is variable, the required sample size in each stratum is determined by the following formula:

$$n_h = \frac{W_h S_h}{\sqrt{c_h}} \cdot \frac{\sum W_h S_h \sqrt{c_h}}{V_o + \frac{1}{N} \sum_{h=1}^{H} W_h S_h^2}$$

and the total minimum sample size required is as follows:

$$n_{\min} = \frac{\sum_{h=1}^{H} \frac{W_h S_h}{\sqrt{c_h}} \cdot \sum_{h=1}^{H} W_h S_h \sqrt{c_h}}{V_o + \frac{1}{N} \sum W_h S_h^2}.$$

With stratification, the mean and variance are estimated as follows:

Parameter Estimated	*Sample Mean*	*Sample Variance*
Population mean	$\bar{y} = \sum_{h=1}^{H} W_h \bar{y}_h$	$\text{var}(\bar{y}) = \sum_{h=1}^{H} W_h^2 \text{ var}(\bar{y}_h)$
Population proportion	$(\bar{p}) = \sum_{h=1}^{H} W_h \bar{p}_h$	$\text{var}(\bar{p}) = \sum_{h=1}^{H} W_h^2 \text{var}(\bar{p}_h)$

where the exact formula for the variance depends on how the elements within the strata are sampled. If SRS is used within each stratum, then the variance formula becomes (1) for the sample mean:

$$\text{var}(\bar{y}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1 - f_h}{n_h} \right) s_h^2$$

where:

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{ih} - \bar{y}_h)^2.$$

or (2) for the sample proportion:

$$\text{var}(\bar{p}) = \sum W_h^2 \left(\frac{1-f_h}{n_h-1}\right) p_h(1-p_h)$$

where:

$p_h = \sum_{i=1}^{n_h} \frac{y_{ih}}{n_h}$, and $y = 1$ or $y = 0$ if the sample unit has the characteristic of interest or not.

In stratified sampling, the probability of inclusion depends on the allocation method adopted:

Method 1. In Equal Allocation

$$f_h = \frac{n/H}{N_h}$$

Method 2. In Proportionate Allocation

$$f_h = \frac{n}{N}$$

Method 3. In Optimum Allocation

$$f_h = \frac{n_h}{N_h}$$

and the weights are calculated as usual as $\omega = f^{-1}$.

How to Carry Out Systematic Sampling

Systematic sampling consists of selecting units at fixed intervals throughout the frame (or stratum) after a random start. Given a population size, *N,* and a desired sample size, *n,* systematic sampling consists of (1) determining the sample interval $k = \frac{N}{n}$; (2) selecting a random number (*RN*) from 1 to k^{9}; and (3) choosing all possible elements belonging to positions *RN, RN* + *k, RN* + 2*k,* . . . , *RN* + (*n* – 1)*k.*

[9] A table of random numbers is reproduced in appendix 5. This table can be used from any point in any direction (vertically, horizontally, diagonally) to get the series of random numbers needed.

Conceptually, with systematic sampling, we subdivide the population in an $n \times k$ matrix:

	1	2	3	. . .	k
1	y_1	y_2	y_3	. . .	y_k
2	y_{k+1}	y_{k+2}	y_{k+3}	. . .	y_{2k}
. . .	. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .	. . .
n	$y_{(n-1)k+1}$	$y_{(n-1)k+2}$	$y_{(n-1)k+3}$	. . .	y_{nk}

and then select the sample corresponding to the column number equal to the drawn *RN*.

Steps in Systematic Sampling

Suppose we have the following population frame:

Obs.	1	2	3	4	5	6	7	8	9
Y	5	9	12	10	7	5	9	11	10

Step 1. Given $N = 9$ and $n = 3$, determine $k = \dfrac{N}{n} = \dfrac{9}{3} = 3$.

Step 2. Divide the observations in the population in a $n \times k$ matrix.

$$n\left\{\overbrace{\begin{matrix}1 & 2 & 3\\ 4 & 5 & 6\\ 7 & 8 & 9\end{matrix}}^{k}\right.$$

Step 3. Select a *RN* between 1 and $k = 3$, suppose $RN = 2 = k$.

Step 4. Select the sample corresponding to column two and include population elements in position 2, 5, and 8, hence obtaining the sample.

obs	Y
2	9
5	7
8	11

The order in which units (elements) are listed is critical in systematic sampling. If the population includes homogeneous elements, the sampler should position homogeneous units across rows (not columns) in the $n \times k$ matrix described before. This creates strata composed of n rows, and systematic sampling will be equivalent to a proportional stratified sampling. One can also observe that forming the $n \times k$ matrix is equivalent to dividing the population into k clusters represented by the columns of the matrix. Hence, while listing the elements in the population, a sampler should attempt to distribute each patch or type of homogeneous units as far as possible uniformly among the clusters, that is, the columns of the matrix.

It is possible that the population size, N, is not an exact multiple of n, so that $N = n \times k + r$ where r is less than k. There are two possible ways to handle the situation. First, select r sample units at random from the population and drop them before creating the $n \times k$ matrix and then proceed from step 3 onward. Second, a more precise method would be to form the matrix with all $n \times k + r$ units so that r columns within the $n \times k$ matrix would have $n + 1$ units, while the remaining $n - r$ columns would have n units. Next, select one column at random, as in step 3, and proceed onward as usual. A similar procedure can be used in situations in which $N = n \times k - r$. The step described is equivalent to considering the list of $N = n \times k + r$ or $N = n \times k - r$ units as circular and after a random number (RN) is selected elements in position RN, 2 RN, 3 RN, and so on are selected achieving a sample of n, $n + 1$ or $n - 1$ units.[10]

In systematic sampling, the mean is calculated as follows:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

because it is an *epsem* design.

Estimating the variance in systematic sampling poses challenging problems. In practice, assuming the order of elements is random,[11] the variance is computed as follows:

$$\operatorname{var}(\bar{y}) = \frac{(1-f)}{n} s^2$$

where s^2 is the variance as calculated in SRS.

The variance in stratified systematic sampling is computed as follows:

$$\operatorname{var}(\bar{y}) = \frac{(1-f)}{n} \sum_{h=1}^{H} W_h s_h^2$$

[10] Kish (1965) presents a number of other methods to deal with this problem. We refrain from addressing them here because, in general, one additional observation in a sample of an Investment Climate Survey presents no implementation problem.

[11] Kish (1965) presents alternative formulas when the SRS requirement is not met.

The probability of inclusion for each element is as follows:

$$f = \frac{1}{k}$$

and the weight is:

$w = f^{-1} = k$ and $w_h = f_h^{-1} = k_h$ in stratified SRS.

How to Carry Out the Probability Proportional to Size Selection Method

The PPS selection method is used when we follow a two-stage sampling procedure and are faced with clusters of unequal size, and if we wish to have control over the total sample size, *n,* and the number of clusters included in the sample while keeping an *epsem* design. With PPS, each element in the frame will be selected with a PPS of the cluster to which it belongs.

Steps in Sample Design

Step 1. Given a population, *N,* choose the desired sample size *n*.

Step 2. Choose either the number of clusters *a* to include in the sample or the number of cluster elements *b* to include in the sample, where $n = a \times b$.

Step 3. Draw a *RN* from 1 to *N* with replacement.

Step 4. Select the cluster that has the cumulative sum that first exceeds the *RN*.

Step 5. Sample *b* elements in that cluster using SRS, systematic sampling, stratified sampling, or cluster sampling.

Step 6. Repeat steps 3 to 5 *a* times.[12]

Because PPS is an *epsem* design, the mean is estimated as follows:

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

[12] Note that the same cluster can be selected multiple times because the *RN* drawing is with replacement. So, if a cluster is selected *k* times, $k \times b$ elements will be drawn from that cluster.

While the variance is calculated as follows:

$$\text{var}(\bar{y}) = \frac{1-f}{a} s_a^2$$

where:

$$s_a^2 = \frac{1}{a-1} \sum_{j=1}^{a} (\bar{y}_j - \bar{y})^2$$

$$y_j = \frac{\sum_{\beta=1}^{b} y_{j\beta}}{b} \text{ (cluster sample mean).}$$

The probability of inclusion is given by the following:

$$f = \frac{aB_\alpha}{N} \frac{b}{B_\alpha}$$

where:

B_α = cluster size
a = number of cluster included in sample
b = number of cluster elements included in the sample.

Finally the weight is as follows:

$$w = f^{-1} = \frac{NB_\alpha}{aB_\alpha b}.$$

An alternative method[13] consists of selecting clusters with PPS *without replacement* and then sample elements are drawn within selected clusters at random (SRS). This procedure preserves the characteristics of *epsem* design and it is implemented in the following steps.

Steps in Sample Design

Step 1. Given a population, N, choose the desired sample size n.

Step 2. Choose the number of clusters a to be included in the sample and the number of elements b to be included in the sample clusters, where $n = a \times b$.

[13] I owe the inclusion of this method to Dr. Yusuf.

Step 3. Let *i* designate a cluster in the population, with a total of *A* clusters in the population. Furthermore, let *j* indicate a sample element within a cluster with a total of B_i units or elements in the ith cluster. Thus, $\sum_{i}^{A} B_i = N$.

Step 4. List the clusters and cumulated clusters' sizes B_i as follows:

$$B_1 = C_1$$
$$B_1 + B_2 = C_1 + B_2 = C_2$$
$$B_1 + B_2 + B_3 = C_2 + B_3 = C_3\text{, etc. and}$$
$$B_1 + B_2 + B_3 + \ldots . + B_A = C_3 + \ldots . + B_A = C_A$$

Step 5. Compute interval of selection $I = \frac{N}{a}$ where *a* is the number of clusters to be selected.

Step 6. Select an RN with $0 < RN \leq I$.

Step 7. Systematically select *a* clusters with cumulated C_i ($i = 1, 2, \ldots . a$) containing the selection vector elements RN, RN+I, RN+2I, ... RN+ $(a - 1)$I.

Step 8. Select *b* sample elements in each selected cluster using SRS, systematic sampling, stratified sampling, or cluster sampling.

Note that a cluster may be selected more than once only if its size is larger than the interval of selection I. The procedure, however, retains *epsem* characteristics of the sample. A refined standard practice is to select the clusters with sizes larger than the interval of selection I automat ically (that is, with probability 1) and attach weight $w = 1x\frac{b}{B_i}$ to the sample elements selected from such clusters.

If a selected cluster has less than *b* sample elements, say *l,* all elements of the cluster are selected and the remainder $(b - l)$ elements are selected from the adjacent (serially/geographically) cluster. PPS selection (without replacement) is an *epsem* selection with weight *w* for all sample elements given by the following:

$$w = \frac{N}{n}.$$

The estimate of the population mean is as follows:

$$\overline{y} = \frac{\sum_{t=1}^{n} y_t}{n}.$$

The estimate of the variance of the sample mean $\bar{y}$ is given by the following:

$$\mathrm{var}(\bar{y}) = \frac{(1-f)}{a} s_a^2$$

where:

$$s_a^2 = \frac{1}{(a-1)} \sum_{i=1}^{a} (\bar{y}_i - \bar{y})^2$$

$$\bar{y}_i = \frac{\sum_{i=1}^{b} y_{ij}}{b} \text{ (cluster sample mean).}$$

How to Deal with Population Frame Problems

The accuracy of any sampling procedure rests not only on the correct application of the relevant theoretical model, but also, and critically, on the accuracy of the frame. The perfect frame in which each unit is listed exhaustively and uniquely and in which no foreign elements appear is a rare event. Often frames are riddled with problems. Identifying and correcting them remains an important part of sampling.

Frame problems are important because they affect the underlying probability of inclusion of the sample units, thus tainting the original sample design and original weights. Hence, the weights assigned to each element at the design stage must be recalibrated at the estimation stage if frame problems occur.[14] Kish (1965) identifies four categories of problems that can be attributed to faulty frames:

Problem 1. Noncoverage

Some population elements might not be included in the list. This can happen because frames are inadequate or incomplete (Kalton 1983). In business surveys, it is often the case in developing nations that frames are out of date, thus failing to include all elements of the target population.

[14] In addition to frame problems, survey nonresponse has an added impact on the weights; this needs to be taken into account when analyzing the data. The literature has developed a number of different procedures to adjust weights for both *unit nonresponse* and *item nonresponse*. While a discussion on adjustments for unit nonresponse is presented in the next section, methodologies to handle item nonresponse go beyond the scope of these notes.

Solutions at the Design Stage

Solution 1. *Redefine the target population* to exclude the missing elements. This solution is acceptable only when the excluded group is a very small proportion of the target population (Kalton 1983).

Solution 2. *Add supplementary frames* in which the missing elements are included. This solution is preferable to solution 1 although it might generate another problem, duplication. This problem however is less pervasive and can be easily handled (see below) (Kalton 1983).

Solution 3. *Adopt a linking procedure* to attach missing elements to existing elements in a clear, practical, and unique manner. Hence, when the existing element is drawn, all linked elements are also selected. This solution has the same drawbacks as a cluster sampling (Kish 1965).

Solution at the Estimation Stage

Solution 1. *Poststratification* uses stratification weights after the completion of the survey. This method allows the adjustments of weights in a way more respondent to the actual population (table 4.1) Poststratification weight adjustment is particularly useful in the event of an outdated frame list. Hence, for example, if an establishment listed as small at the design stage is (after the survey) discovered to belong to a different size category, it is essential to adjust the weights accordingly with a procedure similar to poststratification.

Table 4.1
Weight Adjustments for Noncoverage

	Design			**Poststratification**		
	Population	**Sample**	**Design Weight**	**Final Population**	**Adjustment Factor**	**Weight Adjusted**
Strata	**N**	n_h	$w=p^{-1}$	N_{ps}	$w_{ps}=N_{ps}/N_h$	$w_{jps}=w*w_{ps}$
A	5,000	250	20	?	1.4	28
B	15,000	500	30	?	1.4	42
Total	20,000			28,000		

Source: Author's calculations.

Problem 2. Duplicates

Sometimes, especially when the frame is constructed as the combination of different frames, some elements may appear more than once. This has an impact on the probability of inclusion and thus needs to be taken into account.

Solutions at the Design Stage

Solution 1. *Adopt a unique identification method.* That is first determine a precise order for each separate listing (the first most important, the second, and so on). Then for each element in the first list eliminate duplicates appearing in the subsequent lists (Kish 1965).

Solution 2. *Adjust the sample of subsequent lists.* Draw an independent sample from each list. Then check and eliminate in the sample of any subsequent lists the elements that appear in the full previous lists (Kish 1965).

Solution at the Estimation Stage

Solution 1. *Use weight adjustment.* Reestimate the weights of the duplicate listing to account for their higher probability of inclusion. If, for instance, two independent samples A and B are drawn from two lists with probability f_a and f_b, all A sample elements that appear also in the B list should have a weight $f = f_a \times f_b$ (Kish 1965).[15]

Problem 3. Blanks or Foreign Elements

Blanks are frame elements without corresponding population elements, while foreign elements are units that belong to the frame but are outside the scope of the survey (Kalton 1983). In business surveys, this might be a company that went out of business but is still listed in the frame, or it could be a company that is operating in an industry outside the research interest.

Solutions at the Design Stage

Solution 1. *Ignore the selected element.* The implication of this is that the total sample at the end of the survey will be lower than the desired size. As a consequence, if it is possible to estimate the proportion γ of blanks and foreign elements in the frame, the sample size must be adjusted as follows:

[15] And, similarly, all B sample elements that appear in the A list.

$$n' = \frac{n}{(1-\gamma)}.$$

One common mistake in this case is to replace the blank or foreign element with the element next to it. This practice should be avoided because it is nonrandom and assigns a higher probability of inclusion to the elements next to the blanks and foreign (Kalton 1983).

Solution 2. *Conduct a two-stage selection.* In the first stage, a screening interview is conducted to determine whether the sampled elements exist and meet the objective of the study. Then a second selection is performed to determine a subsample of eligible elements. This approach is appropriate when only a small fraction of the population is of interest (Kalton 1983).

Problem 4. Clustering

Sometimes a listing of elements might include some clusters. In business surveys, this happens when a frame of establishments also includes firms (groups of establishments).

Solutions at the Design Stage

Solution 1. *Take all elements in the cluster.* This solution has the advantage of being easily applicable. With this approach, each cluster element will have a probability of inclusion equal to the probability of selection of the cluster itself. Hence, no reweighting is necessary. This method presents two disadvantages:

- It generates higher variance the larger is the cluster and the higher is the intraclass correlation.
- It could generate response contamination if units in the same cluster are influenced by the other element's responses. This is particularly true for attitude questions.

Solution 2. *Take only one element in the cluster.* This subsampling procedure eliminates the above disadvantages, but carries its own drawbacks. First, it is harder to implement, and second, it changes the probability of selection of the elements and thus requires reweighting. The first drawback relates to which rule should be followed in the selection of the single cluster element. To avoid selection bias, it is important for a random procedure be adopted. Kalton (1983) argues that it is unrealistic to rely exclusively on the interviewer's ability and willingness to

Table 4.2
Kish's Selection Grid

Number of Establishments in Cluster	If Questionnaire Contains Table							
	A	B	C	D	E	F	G	H
	Then Select the Establishment in Position							
1	1	1	1	1	1	1	1	1
2	1	1	1	1	2	2	2	2
3	1	1	1	2	2	3	3	3
4	1	1	2	2	3	3	4	4
Share of questionnaires containing each table	17%	8%	8%	17%	17%	8%	8%	17%

Source: Kish 1965.

apply a random process (that is, a table of random numbers). He or she might inadvertently misapply this method and select the respondent on the basis of his or her availability instead, without the researcher being able to check the procedure adopted. To avoid this bias, a widely used method is the Kish selection grid (table 4.2). This is an objective and checkable procedure to select one respondent in a cluster while keeping equal probability of selection among the cluster's elements. With this method, all eligible establishments in the cluster are first ordered on the basis of some clear, precise, and objective measure (such as total sales, number of employees, and so on). In each questionnaire a table is printed instructing the interviewer to select the respondent corresponding to a specific position in the ordered list. In this way, the interviewer has no control over the selection process. The randomness is introduced by the fact that different tables are printed on different questionnaires, each assigning a different probability of selection and giving an overall equal probability of selection for all elements (Kalton 1983).[16] So, for instance, if the interviewer has a form with table C printed on it, when he or she encounters a cluster with four eligible respondents, he or she will select the second element in the ordered list (table 4.2).

[16] As an alternative method, instead of printing the grid on each questionnaire, it can be printed on a number of letters given at random to interviewers.

The adoption of this procedure changes the probability of inclusion of the cluster elements. Hence, at the estimation stage, the weights must be recalibrated. The probability of inclusion of element *i* is dependent not only on the probability of selection of the cluster but also on the total number of elements in the cluster. So the new probability is as follows:

$$f = f_c \times f_a = \frac{n}{N} \times \frac{1}{A}$$

where:

f_c = probability of selecting the cluster
f_a = probability of selecting one unit *a* in the cluster of A elements.

The weight of *a*th element is thus:

$$\omega = f^{-1} = \frac{NA}{n}.$$

Impact of Mergers, Acquisitions, and Separations on Sampling Weights

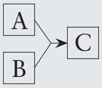
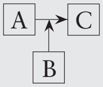
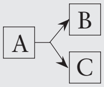
In business surveys, frame problems are particularly frequent because establishments continuously change industry, form, and structure making the maintenance of an up-to-date listing extremely difficult. Hence, the sample designer must be particularly careful in identifying and handling missing, blanks, duplicates, or clusters. In fact, what appears to be a nonexisting establishment, in reality, could be a new establishment (or firm). The fact that establishments merge or split might give the false appearance of blanks and duplicates. If not properly handled, these phenomena might taint the underlying weights assigned at the design stage.

Three general scenarios can happen: *mergers, acquisitions,* and *separations*. In mergers and acquisitions two establishments,[17] A and B, form a new establishment, C. Hence, A and B no longer exist and the new establishment C incorporates the assets and liabilities of both A and B. In these situations, depending on how up to date the frame is,

[17] In this discussion, we assume the establishment to be the unit of analysis as defined in box 4.1.

Table 4.3

Frame Accuracy and Sampling Weights: The Case of Mergers, Acquisitions, and Separations in Establishment Surveys

In the Event of: Phenomenon	The Frame Includes:	The Sample Designer Can:		Unit Weight Adjustments
Merger 	A or B	Select C		$w_a=w_b=w_c=N_h/n_h$
	A and B	Select C		$w_c=N_h/2n_h$
	A and C or B and C	Either treat A=blank (or B=blank[a])		n.a.
		or select C		$w_c=N_h/2n_h$
Acquisition 	A and B and C	Either treat A=blank and B=blank[a]		n.a.
		or select C		$w_c=N_h/3n_h$
	A	Treat A as a cluster:	Select one of 2	$w_i=2N_h/n_h$[c]
			Select all 2	$w_i=N_h/n_h$
Separations 		Either treat A=blank[b]		n.a.
	A and B	or treat A as a cluster	Select one of 2	w_i $2N_h/n_h$ for $i \neq B$ or $2N_h/3n_h$ for $i=B$
			Select all 2	w_i N_h/n_h for $i \neq B$ $N_h/2n_h$ for $i=B$
	A and C	Same as in example above, with C=B		n.a.
	A and B and C	Either treat A=blank[a]		
		or treat A as cluster	Select one of 2	$w_i=2N_h/3n_h$
			Select all 2	$w_i=N_h/2n_h$

Source: Author's creation.

Note: Assuming a stratified simple random sampling of size n_h in a population of N_h with corresponding weights $w=N_h/n_h$.

a. This is preferable.

b. This alternative is less desirable, because it implies C having a zero probability of selection.

c. If the cluster has three elements, then $w_i=3N_b/n_b$.

n.a. = not applicable.

different selection criteria should be adopted (see table 4.3). If the frame contains only one of the two original establishments, A or B, the new establishment C can be included in the sample (linking methodology) without any weight adjustments. On the contrary, if a combination of any of the two original establishments is in the frame—A and

B, A and C, or B and C—then different selection criteria can be adopted, having a different impact on the unit weight. Thus if both A and B are present in the listing, then C (the new establishment) must be selected and the unit weight of C must be modified to $\frac{N_h}{2n_h}$, because C had twice the probability of selection:[18]

$$f_C = f_A + f_B = \frac{n_h}{N_h} + \frac{n_h}{N_h} = \frac{2n_h}{N_h}$$

where f_C is the probability of selection of C.

If A and C or B and C are included in the frame, then the designer has two alternatives. One is to consider A (or B respectively) as blank. This is acceptable because C is present itself in the list and hence has a non-zero probability of selection, while A (or B) no longer exist and can be considered as frame problems (blank). No weight adjustment is necessary in this case. Alternatively, although A or B is randomly selected, the designer can choose to include C (the new establishment) in the sample (linking methodology discussed above). In this case, however, we need to adjust the sampling weight for C because it has twice the probability of selection. Thus, the weight for C becomes $\omega_C = \frac{N_h}{2n_h}$ as discussed above. The decision on the approach to follow must be made *before* the sample is drawn. In fact, in our last example, if C is randomly drawn and either A or B are also in the frame, the weight of C will depend on whether we consider A (or B) as blank, irrespective of whether they are actually drawn. For simplicity of calculation, it is recommended to consider A or B as blanks. In this scenario in fact C will have the same weight as all other elements in the stratum and no weight adjustment is needed.

The last case within mergers and acquisitions is when all three elements, A, B, and C, are in the frame. Once again, the same logic applies. If we adopt (a priori) the policy of considering A and B as blanks, no adjustment is necessary. If C is selected, then it is part of the sample, while if A and/or B are drawn, they are dropped. Alternatively, if A and/or B are selected and we link them to C (hence C is included in the sample although not directly selected), then we need to adjust the ω_C because C

[18] Recall that the weight is the inverse of the probability of selection.

now has three times the probability of selection compared with other elements in its stratum. Therefore $\omega_C = \frac{N_h}{3n}$ because

$$f_C = f_A + f_B + f_C = \frac{n_h}{N_h} + \frac{n_h}{N_h} + \frac{n_h}{N_h} = \frac{3n_h}{N_h}.$$

Separations present more complicated scenarios. The first is when only A (the old establishment) is in the frame. This is the typical case of cluster and, as such, we have two options. We can either include all the members of the cluster (that is, B and C) in our sample or include only one of them. In the first case, no weight adjustment is necessary. In the second, we need to adjust the weight because the probability of selection of each member of the cluster is not equal to the probability of selection of the cluster itself. Hence, the probability of selection of the element included in the sample, either B or C, will be equal to the probability that the cluster is selected multiplied by the probability that the element is drawn given that the cluster is selected:

$$f_i = f_A \times f_{i/A} = \frac{n_h}{N_h} \times \frac{1}{2} = \frac{n_h}{2N_h}$$

where i = B or C.[19]

Similarly, when A and B are present in the frame, we have two options. We can consider A as a blank and disregard it. This is not preferable because it would imply a zero probability of selection for C. Alternatively, we can treat A as a cluster and, again, we have two choices—either to include all elements of the cluster or just one. In both cases, however, we need to adjust the unit weights. If we select all elements of the cluster, the weight of C remains the same as all other elements of the stratum and no adjustment is needed. When we estimate the weight of B, however, we need to consider the fact that B has twice the chances of being selected—once if drawn directly and once if A is drawn. Hence, the weight of B is

[19] Note that throughout this section we assume a cluster of two elements. If more elements are in the cluster, the value of the probability of selection will change accordingly. Hence, if three elements are in the cluster, the probability is $f_i = f_A \times f_{i|A} = \frac{n_h}{N_h} \times \frac{1}{3} = \frac{n_h}{3N_h}$.

the inverse of the sum of the probability of A (the cluster) being selected and the probability of B itself being selected is:

$$f_B = f_A + f_B = \frac{n_h}{N_h} + \frac{n_h}{N_h} = \frac{2n_h}{N_h}.$$

Conversely, if we select only one of the elements of the cluster, then the weight of the chosen element must be adjusted accordingly. If this chosen cluster element is C, then the weight is $\omega_C = \frac{2N_h}{n_h}$ because its probability of selection is as follows:

$$f_C = f_A \times f_{C/A} = \frac{n_h}{N_h} \times \frac{1}{2} = \frac{n_h}{2N_h}.$$

If, however, the element drawn is B, then the unit weight of B is $\omega_B = \frac{2N_h}{3n_h}$ because its probability of selection will depend on both A being drawn and B itself being drawn:

$$f_B = f_A \times f_{B/A} + f_B = \frac{n_h}{N_h} \times \frac{1}{2} + \frac{n_h}{N_h} = \frac{3n_h}{2N_h}.$$

Finally, if all three elements A, B, and C are in the frame, we have two options. We can treat A as a blank or treat A as a cluster. The first option is preferable because it will not taint the weights of the other elements and does not involve any weight adjustments. The second option has two alternatives: we can select just one element of the cluster or all of them. If we chose only one, its weight will be estimated as the inverse of the following:

$$f_i = f_A \times f_{i/A} + f_i = \frac{n_h}{N_h} \times \frac{1}{2} + \frac{n_h}{N_h} = \frac{3n_h}{2N_h}.$$

If we chose all elements of the cluster, then the weight of each of them is simply the inverse of the following:

$$f_i = f_A + f_i = \frac{n_h}{N_h} + \frac{n_h}{N_h} = \frac{2n_h}{N_h}$$

for i = B or C.

Weight Adjustments and Poststratification

While it is not necessary to use weighted results with SRS and other *epsem* methods, when stratified random sampling is adopted, results must be weighted if population parameters need to be estimated (box 4.3). In this case, sampling weights estimated at the design stage (called design weights) must be corrected to compensate for unit nonresponse and frame problems. Both of these phenomena can have a significant impact on the final sampling weights and must to be taken into account at the end of the fieldwork.

Suppose the sampling strategy in a business survey is a stratified SRS with proportional allocation. The strata are determined using sector (garments and textiles), size (small, medium, and large), and location (north and south). Furthermore, as part of the sample design, some strata were collapsed and large establishments were selected with certainty. The sample structure and the estimated weights are summarized in table 4.4. The design weights are estimated as the inverse of the probability of selection:

$$\omega_h = f_h^{-1} = \left(\frac{n_h}{N_h}\right)^{-1} = \frac{N_h}{n_h}.$$

After the data collection is completed, the design weights need to be adjusted before the analysis can commence. The true weights (w) are obtained by multiplying the design weights (w_{DES}) by an adjustment factor for unit nonresponse (w_{RES}) and an adjustment factor for frame problems (w_{FP}):

$$w = w_{DES} * w_{NR} * w_{FP}.$$

The unit nonresponse adjustment factor is calculated by estimating the proportion of the total sample that participated to the survey. Hence,

$$w_{NR} = \frac{n_h}{n_h^r}.$$

where:

n_h^r = number of respondents who participated in the survey and belong to stratum h[20]

n_h = number of sampled elements in stratum h at the design stage

[20] Note that this number includes all respondents to the survey, even those who are sampled in a different stratum than h (because if inaccurate classification) but belong to h.

Box 4.3

Why it is Important to Use Weights with Stratified Sampling

Suppose the car company you work for wants to issue a warranty on transmissions. You are asked to estimate the cost of such a warranty. To do so, you need to estimate at what mileage, on average, cars require transmission repair. Assume you have the list of car owners shown in box table 4.3.1, and your budget allows you to sample nine elements. The true value, unknown to you and that you need to estimate, is 60,120 miles.

Box Table 4.3.1

List of Car Owners

Car ID	Repair Mileage (unknown)	Location	Car ID	Repair Mileage (unknown)	Location
1	85,900	Rural	11	31,500	Urban
2	99,500	Rural	12	48,600	Urban
3	82,100	Rural	13	45,500	Urban
4	70,000	Rural	14	38,500	Urban
5	74,100	Rural	15	49,000	Urban
6	77,000	Rural	16	42,000	Urban
7	68,500	Rural	17	45,500	Urban
8	94,500	Rural	18	35,000	Urban
9	69,700	Rural	19	42,000	Urban
10	65,200	Rural	20	38,500	Urban
		Average	**60,120**		

Source: Author's creation.

The first choice at your disposal is to follow a simple random sampling (SRS) methodology. From the nine samples, you obtain an average value of 56,767 miles.

Nonetheless you have reason to believe from previous discussions with mechanics that transmissions require repair earlier if the car is driven in rural areas than if it is driven in urban locations. Luckily your list includes this information. You decide to sample the nine elements using stratified random sampling. Because you suspect that in rural areas the variance of the parameter you want to estimate (repair mileage) is twice as high as in urban locations, you decide to sample more cars in rural stratum (6) than in the urban stratum (3), as described in box table 4.3.2.

(*continued*)

Box 4.3 (continued)

Box Table 4.3.2

Results Using Simple Random Sampling and Stratified Sampling

Simple Random Sampling		Stratified Random Sampling			
Car ID	Repair Mileage	Stratum	Car ID	Repair Mileage	Weight
3	82,100	Rural	1	85,900	1.67
4	70,000		3	82,100	1.67
6	77,000		4	70,000	1.67
10	65,200		5	74,100	1.67
12	48,600		8	94,500	1.67
13	45,500		10	65,200	1.67
16	42,000	Urban	14	38,500	3.33
19	42,000		15	49,000	3.33
20	38,500		20	38,500	3.33
Average		Simple average		Weighted average	
56,767		66,422		60,317	

Source: Author's creation.

With stratification, you can still estimate the mileage repair value by simple average over all nine observations. The value obtained is 66,422, which is even less accurate than the value obtained using SRS. Hence, while using the simple average with stratification improves the estimate of the parameter within each stratum, it biases its estimate of the whole population. To obtain a more accurate estimate than SRS, we must use the weights. If we do that we obtain a value of 60,317, which is closer to the true population than SRS (box table 4.3.2).

Estimating the adjustment factor for frame problems is more complex. As mentioned earlier, four main categories of frame problems might occur: noncoverage, duplicates, blanks or foreign elements, and clustering. While blanks and foreign elements are dealt with at the design stage, clustering and duplicates have an impact on the weights of the individual sampling unit.[21] Noncoverage, on the contrary, is a source of bias for the

[21] The implications for weight adjustment have already been discussed in the text as well as in table 4.3, so we are not including them again in this example.

Table 4.4

Sample Design: Stratified Sample Random Sampling

Stratum	Stratification Criteria			Stratum Size	Sample Size	Replace-ments	Total Sample	Probability of Selection	Design Weight
	Sector	Size	Location	N^0_h	n^0_h	R_h	$n_h=n^0_h+R_h$	ps	$w=ps^{-1}$
A	Garments	Small	North	1,000	76	24	100	0.10000	10.00
B	Garments	Small	South	2,000	164	36	200	0.10000	10.00
C	Garments	Medium	North	2,000	182	18	200	0.10000	10.00
D	Garments	Medium	South	600	82	13	95	0.15833	6.32
E	Garments	Large	North	200	200	0	200	1.00000	1.00
F	Garments	Large	South	180	180	0	180	1.00000	1.00
G	Textiles	Small & Med.	North	2,200	81	46	127	0.05773	17.32
H	Textiles	Small & Med.	South	1,800	95	25	120	0.06667	15.00
I	Textiles	Large	North	300	300	0	300	1.00000	1.00
K	Textiles	Large	South	220	220	0	220	1.00000	1.00

Source: Author’s creation.

whole stratum. Two main noncoverage problems can take place: inaccurate coverage or incomplete coverage. Inaccurate coverage arises when an establishment is listed in the frame but no longer exists. Incomplete coverage occurs when the frame is not updated and some establishments that should be listed in a specific stratum are instead listed in another stratum, with both strata included in the sample stratification. The adjustment factor for frame problems (w_{FP}) depends on both of these factors, and its estimation is not trivial. The fact that an establishment cannot be located does not necessarily mean that it went out of business. It is possible, for example, that it moved to a different location within the study area, in which cases it should be included in the weight adjustments. The fieldwork can provide extremely useful information for their estimation.

This step (the estimation of the adjustment factor from frame problems) requires the estimation of three parameters: (1) the proportion of units p_{h}^{out} that are present in each stratum *h* but should not be because (a) they do not exist or exist outside the target population, n_{h}^{oos},[22] or (b) they belong to other strata in the sample, $n_h^{in} \Rightarrow i$; (2) the proportion of units that belong to *h* but are found in any of the other strata $i \neq h$ of our target population, $p_i^{in} \Rightarrow h$; and (3) the net rate of growth of each stratum since last update, g_h. After these three values are determined, the adjustment factor for frame problems is estimated as follows:

$$w_{FP} = \frac{N_h^{PS}}{N_h}$$

where:

$$N_h^{PS} = N_h \times \left(1 + g_h - p_h^{out}\right) + p_{i \Rightarrow h}^{in} \sum_{i \neq h} N_i$$

$$p_h^{out} = \frac{n_h^{oos} + n_{h \Rightarrow i}^{in}}{n_h} = \frac{n_h^{oos}}{n_h} + \frac{n_{h \Rightarrow i}^{in}}{n_h} = p_h^{oos} + p_{h \Rightarrow i}^{in}$$

$$p_{i \Rightarrow h}^{in} = \frac{\sum_{i \neq h} n_{i \Rightarrow h}^{in}}{\sum_{i \neq h} n_i}.$$

While g_h must be estimated on the basis of prior knowledge,[23] and the estimation of p_{h}^{out} is pretty straightforward, $p_i^{in} \Rightarrow h$ is complicated

[22] That is, they do not belong to any of the strata in our sample.

[23] Because we are interested in the net growth rate of the stratum, an analysis of the dynamics of the population under study over the past few years might provide useful information.

by the fact that units belonging to one stratum can appear in any other strata surveyed. This requires an accurate recording of what happens during the fieldwork to properly reclassify the units and obtain correct weights.[24]

Example

As an illustration of this methodology, let us assume for simplicity that we have only the first three strata of table 4.4. After the fieldwork is completed, records indicate that (in stratum A) 85 establishments participated, 5 refused, and 10 were inaccurately classified (4 out of scope and 6 belonging to other strata in the sample). Furthermore, because of frame inaccuracy, 60 establishments belong to stratum A, but they have been sampled in strata B and C as shown in table 4.5. Let us also assume a similar pattern for strata B and C, so that at the end of the fieldwork the results are as shown in table 4.6.

After this is done, the final weights can be estimated as described above. Hence, for stratum A, the final weight is estimated as follows:

$$p_A^{out} = \frac{n_A^{out}}{n_A} = \frac{4 + (1 + 5)}{100} = 0.10$$

$$p_{i \Rightarrow A}^{in} = \frac{\sum\limits_{i \neq A} n_{i \Rightarrow A}^{in}}{\sum\limits_{i \neq A} n_i} = \frac{30 + 30}{200 + 200} = 0.15$$

$$N_A^{PS} = N_A \times (1 + g_A - p_A^{out}) + p_{i \Rightarrow A}^{in} \sum_{i \neq A} N_i = 1000 \times (1 + 0.05 - 0.10) + 0.15 \times 2000 = 1250$$

$$w_{FP} = \frac{N_A^{PS}}{N_A} = \frac{1250}{1000} = 1.25$$

$$w_{NR} = \frac{n_b}{n_b^r} = \frac{100}{85 + 30 + 30} = 0.69$$

$$w = w_{DES} \times w_{NR} \times w_{FP} = 10 \times 0.69 \times 1.25 = 8.62$$

All other weights are shown in table 4.7.

[24] To this end, appendix 7 reports the minimum amount of information to collect during the fieldwork to facilitate both weight adjustment and the compilation of figure 2.3 presented at the end of chapter 2.

Table 4.5

Weight Adjustment Components for Stratum A

					Inaccurate	Incomplete		Share of			
Stratum	**Stratum Size**	**Sample Size**	**Respondents**	**Refusals**	**Out of Scope**	**To Other Strata**	**From Other Strata**	**OOS**	$\mathbf{IN}_{h \longrightarrow i}$	$\mathbf{IN}_{i \longrightarrow h}$	**Growth Rate**
	N_h^0	n^0		n^{ref}	n^{oos}	$n^{in}_{h \longrightarrow i}$	$n^{in}_{i \longrightarrow h}$	p_h^{oos}	$p^{in}_{h \longrightarrow i}$	$p^{in}_{i \longrightarrow h}$	g_h
A	1,000	100	85	5	4	to B 1	from B 30	0.04	0.06	0.15	0.05
						to C 5	from C 30				
B	1,000	200				to A 30					
						to C					
C	1,000	200				to A 30					
						to B					

Source: Author's creation.

Table 4.6

Weight Adjustment Components for All Strata

					Inaccurate	Incomplete				Share of			
Stratum	Stratum Size	Sample Size	Respondents	Refusals	Out of Scope	To Other Strata		From Other Strata		OOS	$IN_{h\rightarrow i}$	$IN_{i\rightarrow h}$	Growth Rate
	N_h^0	n^0		n^{ref}	n^{oos}	$n^{in}_{h\rightarrow i}$		$n^{in}_{i\rightarrow h}$		p^{oos}_h	$p^{in}_{h\rightarrow i}$	$p^{in}_{i\rightarrow h}$	g_h
A	1,000	100	85	5	4	to B	1	from B	30	0.04	0.06	0.15	0.05
						to C	5	from C	30				
B	1,000	200	145	5	10	to A	30	from A	1	0.05	0.20	0.003	0.08
						to C	10	from C	0				
C	1,000	200	160	10	0	to A	30	from A	5	0.0	0.15	0.05	0.03
						to B	0	from C	10				
							76	=	76				

Source: Author's creation.

Table 4.7

Estimation of Final Weights

								Response	Frame	
	Stratum	Sample	Share	Share	Growth	Poststratification	Design	Adjustment	Problem	Final
Stratum	Size	Size	of Out	of In	Rate	Population	Weight	Factor	Adj Factor	Weight
	N^0_h	n^0	p^{out}_h	p^{in}_h	g_h	N^{PS}	w_{des}	w_{res}	w_{fp}	w
A	1,000	100	0.10	0.15	0.05	1,250	10	0.69	1.25	8.621
B	1,000	200	0.25	0.003	0.08	840	5	1.37	0.84	5.753
C	1,000	200	0.15	0.05	0.03	1,030	5	1.14	1.03	5.886

Source: Author’s creation.

Sampling in Practice: How to Maximize the Sample Representativeness while Minimizing the Survey Cost through the Use of Poststratification

A common challenge when designing a survey is the desire to analyze many characteristics of the population.[25] Because of budget and time constraints, it is not possible to guarantee a certain level of precision and confidence for all of these dimensions. What this practical example proposes is to address this problem through the careful choice of key strata in drawing the sample and the use of poststratification. These techniques allow for a degree of redesigning of the sample distribution, after the survey is completed, to maintain a predetermined level of precision for different dimensions of analysis within the fixed sample size.

Let us assume that we have obtained or compiled data on the population of manufacturing establishments. The frame list includes the following: (1) sector of activity, (2) location—region, and (3) size—number of employees. The population distribution is presented in table 4.8. Sampling can be designed in the following six steps.

Steps in Sampling Design

Step 1. *Determine the sampling parameters.* The size and composition of the sample will depend on three factors:

- The objective of the study,[26]
- The available budget, and
- The desired level of precision and confidence.

Let's assume that we want to compare characteristics of the Investment Climate environment across locations, as well as estimate the determinants of firm productivity across sectors. Let's further suppose that the available budget allows for a sample size of approximately 850 establishments. Finally, let's assume that we wish to reach a level of statistical significance corresponding to 90 percent confidence and 5 percent precision.[27]

[25] In an Investment Climate Survey, it is often of interest to analyze the business climate across location, size of firms, sector, export orientation, foreign ownership, and so on.

[26] The objective of the study will determine the size of the target population and the characteristic of analysis.

[27] The parameters we wish to estimate are proportions, thus we use the corresponding formula described in the SRS methodology.

Table 4.8
Population Distribution by Sector, Region, and Size

By Sector	
1 Apparel	1,070
2 Basic metals	235
3 Chemical & chemical products	588
4 Electrical machinery	238
5 Electronics	134
6 Food & beverage	2,348
7 Furniture	727
8 Leather products	370
9 Machinery and equipment	400
10 Medical equipment	53
11 Metal products	1,182
12 Motor vehicles	208
13 Nonmetallic mineral products	1,220
14 Other transport equipment	366
15 Paper	599
16 Publishing	438
17 Rubber & plastic products	766
18 Textiles	611
19 Tobacco	24
20 Wood & wood products	904
21 Other (unclassified)	276
Grand Total	12,757

By Region	
Central North	490
Highland	250
Mekong River Delta	1,361
North East	661
North West	78
Red River Delta	3,705
South East	5,466
Southern Central Coastal	746
Grand Total	12,757

By Size	
Small	9,355
Medium	3,086
Large	316
Grand Total	12,757

Source: Author's calculations.

Step 2. *Divide the population in strata and estimate different sampling schemes*. Given that the characteristics of interest are location and sector, we need to stratify the population by each of them separately. Afterward, using table 4.9, we can determine the minimum sample size needed to reach the desired level of statistical significance.

In our example, let's start with a stratification by sector. Because the population includes 21 sectors, we will have 21 strata (table 4.8). Given the desired level significance, we can use table 4.9 to calculate the min-

Table 4.9

Sample Size Requirements for 90 Percent Confidence Interval

		SIZE OF POPULATION												
		50	**100**	**200**	**300**	**500**	**750**	**1,000**	**2,000**	**3,000**	**5,000**	**10,000**	**100,000**	**1,000,000**
		MINIMUM SAMPLE SIZE NEEDED												
DESIRED LEVEL OF PRECISION	**10.0%**	29	41	52	57	61	64	65	67	68	69	69	70	70
	7.5%	36	55	77	88	99	106	110	117	119	121	122	124	124
	7.0%	37	59	83	97	111	120	125	133	136	138	140	142	142
	6.5%	38	62	90	106	124	135	142	152	156	160	162	165	165
	6.0%	40	66	98	118	140	154	162	177	182	186	190	193	194
	5.5%	41	70	107	130	158	176	187	207	214	220	225	230	230
	5.0%	**42**	**74**	**116**	**145**	**179**	**203**	**218**	**245**	**255**	**264**	**271**	**278**	**279**
	4.0%	45	81	137	178	233	276	304	358	380	401	418	434	436
	3.0%	47	89	159	216	304	381	437	558	616	671	719	769	774
	2.0%	49	95	179	256	389	524	635	931	1,102	1,292	1,484	1,713	1,740
	1.0%	50	99	194	288	467	677	875	1,554	2,097	2,912	4,108	6,518	6,924

Source: Author’s calculations.

Note: Assumes highest level of variance within the population.

imum sample size for each stratum. Assuming for the moment that there are no budget constraints, if we aim for a sample representative of all 21 sectors, we would need a sample of 3,436 elements, out of a population of 12,575. Similarly, if our target is a sample representative of all of the regions in the country, we would need a sample of 1,526 units (table 4.10).

Unfortunately, budget constraints rarely allow such a freedom. To meet our budget constraint of approximately 850 units, we can adopt one, or a combination, of the following two strategies:

- Merge strata[28] so that we shrink the overall number of strata, or
- Eliminate strata from our population frame.

Each of these strategies has its disadvantages. While combining strata might compromise the meaningfulness of sectoral comparisons, removing strata lessens the representativeness of the sample at the national level. The best approach is probably a combination of the two. Some strata are merged while others are kept in the stratification. This approach has the advantage of allowing both national as well as sectoral analysis while keeping the sample size at a reasonable level.

The sample designer must weigh the benefits and costs of these approaches. Assuming we decide to follow the second strategy, we must decide which sectors and/or locations to keep in the target population.[29] In reaching this decision, a number of factors must be taken into account, including the following:

- *The importance of these sectors within the objective of the study*. If the purpose of the study is to estimate productivity by focusing on the most important sectors, then the least important sectors within manufacturing should be dropped.
- *The distribution of firms by other relevant dimensions* (for example, location, size, ownership, and so on). To the extent that other dimensions are of analytical interest, the sectors included in the target population should include these dimensions. More specifically, if we wish to estimate the impact of Investment Climate vari-

[28] The sample designer should remember that the definition of strata is dependent on its analytical purpose. Hence, strata can be combined if appropriate to the purpose of study.

[29] In our case, only sector appears to be the binding constraint. After we eliminate some of the sectors in our target population, the total population in terms of location and ownership will also decrease. Hence the corresponding sample sizes will go below or fall near the 800 mark.

Table 4.10

Stratification and Required Sample Size for 90 Percent Confidence and 5 Percent Error

By Sector	Population	Required Sample
Apparel	1,070	221
Basic metals	235	128
Chemical & chemical products	588	189
Electrical machinery	238	128
Electronics	134	91
Food & beverage	2,348	249
Furniture	727	202
Leather products	370	159
Machinery and equipment	400	164
Medical equipment	53	45
Metal products	1,182	226
Motor vehicles	208	119
Nonmetallic mineral products	1,220	227
Other transport equipment	366	158
Paper	599	190
Publishing	438	170
Rubber & plastic products	766	204
Textiles	611	191
Tobacco	24	22
Wood & wood products	904	213
Other (unclassified)	276	139
Total	**12,757**	**3,436**

By Region	Population	Required Sample
Central North	490	178
Highland	250	132
Mekong River Delta	1,361	231
North East	661	196
North West	78	61
Red River Delta	3,705	259
South East	5,466	265
Southern Central Coastal	746	203
Total	**12,757**	**1,526**

Source: Author’s calculations.

ables on firm performance in different locations, then the sectors included in the target population must be present in these locations.

- *The ability to perform international comparisons at the sectoral level.* If in the comparator countries some sectors have already been covered, then the same sectors must be included in the target population.
- *The required sample size and replacements.* Because of nonresponse and frame problems, a number of elements to draw from the population frame must be higher than the actual desired sample size.

Although it would be impractical to show all possible scenarios, let's suppose we decide to adopt employment contribution as selection criterion. As table 4.11 shows, it appears that Mekong River Delta, Red River Delta, South East, and Southern Central Coastal are the most important locations, covering close to 90 percent of employment. Similarly apparel, food and beverages, leather products, nonmetallic mineral products, and textiles are the sectors with the highest concentration of employment (close to 70%).

If we limit our target population to these four regions and five sectors, we can reestimate the minimum required sample size (table 4.12). While the stratification by location is within budget, the minimum required sample size by sector does not meet our budget constraint. Unless we can find additional funds, we need once again to trim the number of strata (or to combine some of them) until we reach a sample size within budget, while we remain satisfied with the sectoral and location coverage.

Step 3. *Reconcile and select the strata sampling strategy to implement in the field.* Suppose we have decided to keep the four most important locations, as well as four out of the five of the sectors previously identified. Our final target population is presented in table 4.13.[30] The next question is which stratification to implement in the field out of the two possible alternatives—sector or location. This choice is important because the stratification criteria implemented in the field is the only one for which we can directly control the level of statistical significance desired.[31]

[30] Although the employment contribution of nonmetallic products is slightly higher than textiles, the decision to keep the latter might be determined by other considerations, such as the ability to use textiles in international comparisons.

[31] The other stratification, which we will reconstruct at the end of the fieldwork (see step 6), will have a level of significance determined indirectly by the number of elements that fall in that stratification.

Table 4.11
Employment Contribution by Sector and Location

By Location		By Sector	
Central North	3%	Apparel	17%
Highland	1%	Basic metals	1%
Mekong River Delta	7%	Chemical & chemical products	3%
North East	5%	Electrical machinery	3%
North West	0%	Electronics	1%
Red River Delta	24%	Food & beverage	15%
South East	52%	Furniture	5%
Southern Central Coastal	7%	Leather products	19%
		Machinery and equipment	2%
		Medical equipment	0%
		Metal products	3%
		Motor vehicles	1%
		Nonmetallic mineral products	8%
		Other transport equipment	3%
		Paper	2%
		Publishing	1%
		Rubber & plastic products	4%
		Textiles	7%
		Tobacco	1%
		Wood & wood products	4%
		Other (unclassified)	2%

Source: Author's calculations.

If we were to implement a stratification by sector the expected sample distribution by location would be as shown in table 4.14. Because we would randomly select elements within each sector, the expected distribution of our sampled elements by location will be approximately proportional to the underlying population distribution. Consequently, because we cannot directly control the number of elements that will fall in each of the location strata, the expected levels of precision by location will slightly differ from the desired levels of 5 percent.

Table 4.12

Target Population (Four Regions and Five Sectors) and Required Sample

Sector	Region Mekong River Delta	Red River Delta	South East	Southern Central Coastal	Total	Required Sample
Apparel	45	280	650	50	1,025	**219**
Food & beverage	817	374	699	152	2,042	**245**
Leather products	13	83	256	7	359	**157**
Nonmetallic mineral	153	315	444	65	977	**217**
Textiles	28	222	270	23	543	**184**
Grand Total	1,056	1,274	2,319	297	4,946	**1,023**
Required Sample	**221**	**229**	**249**	**144**	**842**	**total**

Source: Author's calculations.

Similarly, if we were to implement stratification by location, the expected sample distribution by sector is also shown in table 4.14.

At this point, if we are satisfied with the expected levels of precision, we can proceed to the next step and implement the stratification by sector (or location). If, however, we want to increase the expected level of precision of, say, the stratification by location, we have two options:

- We can oversample some sector strata to increase the number of elements that would fall in the desired locations.

Table 4.13

Final Target Population (Four Regions and Four Sectors) and Required Sample Size

Stratification by Sector	Required Sample	**Stratification by Location**	Required Sample
Apparel	**219**	Mekong River Delta	**195**
Food & beverage	**245**	Red River Delta	**197**
Leather products	**157**	South East	**219**
Textiles	**184**	Southern Central Coastal	**120**

Source: Author's calculations.

Table 4.14

Expected Sample Sizes and Levels of Statistical Significance

	When Sectoral Stratification is Implemented		
	Required Sample[a]	Expected Sample Distribution	Expected Level of Precision[b]
Mekong River Delta	213	123	7.0%
Red River Delta	216	216	5.0%
South East	243	427	3.6%
Southern Central Coastal	127	40	12.0%

	When Location Stratification is Implemented		
	Required Sample[a]	Expected Sample Distribution	Expected Level of Precision[b]
Apparel	219	185	5.60%
Food & beverage	245	451	3.50%
Leather products	157	59	10%
Textiles	184	104	7.40%

Source: Author's calculations.

a. To reach 5 percent precision.

b. Keeping the 90 percent level of confidence.

- We can perform an additional stratification and directly control the number of observations in each new stratum.

In most cases, the first option is hard to implement. Because distributions are often skewed the required oversample could be high.[32] For instance, if we want to obtain more observations in Southern Central Coastal, we could oversample the sector that has the highest concentration in that location (food). This approach, however, will not guarantee the desired location sample size of 120 unless we increase the sample size for food dramatically, which would have obvious implications for our budget.[33]

[32] Recall that with this approach we oversample only the sector strata. Consequently, we have only an indirect control on the desired level of precision for location.

[33] To ensure a sample of 120 in the Southern Central Coastal, we need to increase the sample size of food by more than 700 percent (that is, we need to sample 1,600 elements instead of 245).

The second option is more viable. It consists in performing an additional stratification and then assigning a number of elements within each double-strata to approach a desired level of precision for both sector and location. For example, let's assume we first stratify by sector and then stratify each sector by location as shown in table 4.15. We now have 16 strata. When assigning the sample elements to each sector stratum, we can now directly control the total number of elements in each sector-location stratum. This does not violate sampling protocol as long as the final selection of each individual element within each stratum remains random. Hence, in apparel, instead of selecting randomly 219 elements (which would give us the random distribution described in table 4.15, column 7), we can now directly assign the 219 elements to each of the four locations within apparel. To approach the desired distribution within each location, our goal is to assign more observations to Mekong River Delta and Southern Central Coastal while reducing the sample size in South East.[34] The final sample distribution is shown in the final column of table 4.15.[35]

The reassignment of sample units in the second stratification is not always easy. Note that, in our case, we oversampled the second sector to get as close as possible to the desired level of precision at the location level while meeting all the other constraints. Our final sample increased to 850 elements distributed in 16 strata (final column of table 4.15). The expected level of precision for location increases now to 6.1 percent in Southern Central Coastal (down from 12 percent), 5.7 percent in Mekong River Delta, and 4 percent in South East.

A similar approach can be followed if we have other dimensions (for example, size) for which we wish to ensure a level of precision ex ante. Suppose we are concerned that large firms might be underrepresented in the final sample because of their skewed distribution. If this is the case, we can envisage two situations:

- We are satisfied with the expected level of precision of location, but we are concerned about the expected precision by size alone. Then

[34] Note that in Red River Delta, because the number of observations in the sector stratification is exactly equal to the expected number of observations by location, we leave the sample size as it is.

[35] Note that although we have some flexibility in reassigning observations between strata, we still have a number of constraints to meet, including the availability of replacements. In textiles, for example, even if we wish to reduce the sample size for the South-East, we cannot increase any other location because we would run out of replacements.

Table 4.15

Stratification by Sector and Location

First Level of Stratification				Second Level of Stratification			
Strata No.	Sector	Population	Sample	Location	Population	Expected Sample Size in Case of Random Selection within the Whole Sector	Direct Imputation of Sample Size
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	Apparel	1,025	219	Mekong River Delta	45	10	30
2				Red River Delta	280	60	60
3				South East	650	139	103
4				Southern Central Coastal	50	11	26
5	Food & beverage	2,042	245	Mekong River Delta	817	98	118
6				Red River Delta	374	45	45
7				South East	699	84	68
8				Southern Central Coastal	152	18	59
9	Leather products	359	157	Mekong River Delta	13	6	8
10				Red River Delta	83	36	36
11				South East	256	112	109
12				Southern Central Coastal	7	3	4
13	Textiles	543	184	Mekong River Delta	28	10	17
14				Red River Delta	222	75	75
15				South East	270	92	78
16				Southern Central Coastal	23	8	14
			806			807	850

Source: Author's calculations.

Note: Values might not add up exactly because of rounding.

we can use size as the dimension for the second stratification, exactly as shown above.

- We are concerned about the expected level of precision of both location and size. In this case, we should first estimate the expected level of precision by sector corresponding to a double stratification location-size. If this is satisfactory, then we proceed with the location-size stratification as shown above, disregarding sector. If we cannot exclude sector from the stratification, then we need to add a third level of stratification and follow the same methodology as presented above.

Step 4. *Implement the sampling strategy*. After the sample strategy has been finalized, the actual number of elements to be drawn from the underlying population will have to be adjusted to take into account two main factors:

- The accuracy of the population frame, and
- The expected nonresponse rate.

Often the available frame lists are old and inaccurate. Furthermore, not all selected respondents will participate in the survey. For these reasons an estimated inaccuracy rate and refusal rate must be incorporated in the calculation of the final number of elements to be drawn in each stratum. Assuming an average inaccuracy rate of 5 percent and a refusal rate of 50 percent (equally distributed across strata), then the total number of elements to be drawn must be adjusted accordingly (see table 4.16).[36]

After the total number of elements has been determined, the elements must be drawn from the population frame *randomly and in one draw* (see box 4.4). The sample and replacements must be selected at the same time to ensure that each element within each stratum has the same probability of selection. During the fieldwork, the order in which the interviews are conducted is important. First, all the elements in the sample must be interviewed. Then each of the respondents that does not exist (frame problem) or refuses to participate (nonresponse) has to be substituted in the exact order in which they appear in the drawing.

[36] Note that in the first strata, because the estimated number of sample + replacements is slightly higher than the population, all the elements of the strata will be included in the sample. It is nevertheless important to sample them, because the order of interviewing is critical.

Table 4.16

Sample, Replacements, and Total Elements to Draw

	First Level of Stratification	Second Level of Stratification					
Strata No.	Sector	Location	Population	Population Adjusted[a]	Sample	Replacements	Total No. of Elements to Draw
1	Apparel	Mekong River Delta	45	43	30	13	43
2		Red River Delta	280	279	60	30	90
3		South East	650	649	103	52	155
4		Southern Central Coastal	50	49	26	13	39
5	Food & beverage	Mekong River Delta	817	816	118	59	177
6		Red River Delta	374	373	45	23	68
7		South East	699	698	68	34	102
8		Southern Central Coastal	152	151	59	30	89
9	Leather products	Mekong River Delta	13	12	8	4	12
10		Red River Delta	83	82	36	18	54
11		South East	256	255	109	55	164
12		Southern Central Coastal	7	6	4	2	6
13	Textiles	Mekong River Delta	28	27	17	9	26
14		Red River Delta	222	221	75	38	113
15		South East	270	269	78	39	117
16		Southern Central Coastal	23	22	14	7	21
					850	423	1,273

Source: Author's calculations.

[a] Note the adjustment is only for frame problems (inaccuracy of listing).

Box 4.4

Using SAS to Draw Samples

Modern computing technology has made it easy to perform the actual drawing of sample elements. As for other software programs, SAS has simple commands to randomly select sample elements for a variety of sample designs: simple random sampling (SRS), stratified sampling, systematic sampling, and probability proportional to size (PPS).

(1) **Simple random sampling.** Assume the population frame is stored in the file "frame" and we wish to draw an SRS of $n = 12$ elements. The commands needed in SAS are as follows:

```
proc surveyselect data=frame method=srs n=12 out=sampleSRS;
run;
```

(2) **Stratified sampling.** Suppose we have designed a stratification and the file "frame" contains a strata variable (industry). We can then perform the following:

(a) *Equal allocation* with $n = 4$ in each stratum

```
proc surveyselect data=frame method=srs n=4 out=sampleESTSRS;
strata industry;
run;
```

(b) *Proportionate allocation* with a common rate of 20 percent in each stratum

```
proc surveyselect data=frame method=srs rate=.20
seed=1953 out=samplePSTR;
strata industry;
run;
```

(c) *Disproportionate allocation* with a $n_1 = 3$, $n_2 = 5$, and $n_3 = 4$:

```
proc surveyselect data=frame method=srs n=(3,5,4) out=sampleDSTSRS;
strata industry;
run;
```

(3) **Systematic sampling.** Suppose we wish to draw a systematic sample with a sampling rate of one-quarter. In SAS, the commands to use are as follows:

```
proc surveyselect data=frame method=sys samrate=0.25 out=samplePPS;
run;
```

(4) **Probability proportional to size.** Suppose you wish to obtain a sample with PPS allocation, with $n = 9$ and 'labor' being the size variable. The SAS commands are as follows:

```
proc surveyselect data=frame method=pps n=9 out=samplePPS;
size labor;
run;
```

Hence if, say, strata 1 include 30 sample elements and 13 replacements, if respondent number 3 refuses to participate, he or she will have to be substituted with element number 31 and so on. It is now obvious how important it is to accurately estimate the inaccuracy and nonresponse rate before the drawing of the elements. After the list of drawn elements is exhausted, we cannot draw additional elements from the original population frame without changing the probability of selection of the elements, making the calculation of weights extremely difficult.

Step 5. *Estimate the weights.* After the sample size has been determined, the weights can be estimated (design weights), as shown in table 4.17.[37] At the end of the fieldwork these design weights must be adjusted as described above to obtain the final weights.

Step 6. *Perform the poststratification.* In our example, the final weights refer to the double stratification, sector-location. With respect to these two dimensions, we can make statistically significant inferences at any time without any further adjustment. However, if we wish to make statistically significant inferences with respect to dimensions not expressly included in the sample design, we need to "poststratify" the sample distribution and estimate the corresponding sample weights.

Poststratification means stratifying after the fieldwork has been completed. To poststratify we need to know the population of the dimension of interest. Let's assume that we want to poststratify our population by ownership status—foreign direct investment (FDI) and private. For each stratum, we need to identify the population and the sample corresponding to the new dimension and then estimate the weights in the usual way. Table 4.18 shows the distribution of strata (32) with the relevant information on population, sample, and weights. These new weights can now be used to make statistically significant inferences on ownership status of our target population.

[37] The weights are the inverse of the probability of selection (w=N/n, where N=population and n=sample within each stratum).

Table 4.17

Final Sample and Weights

Strata No.	Sector	Location	Original Population	Final Population[a]	Sample	Design Weight	Final Sample	Final Weights
1	Apparel	Mekong River Delta	45	45	30	1.500	30	1.500
2		Red River Delta	280	280	60	4.667	60	4.667
3		South East	650	650	103	6.311	103	6.311
4		Southern Central Coastal	50	50	26	1.923	26	1.923
5	Food & beverage	Mekong River Delta	817	817	118	6.924	118	6.924
6		Red River Delta	374	374	45	8.311	45	8.311
7		South East	699	699	68	10.279	68	10.279
8		Southern Central Coastal	152	152	59	2.576	59	2.576
9	Leather products	Mekong River Delta	13	13	8	1.625	8	1.625
10		Red River Delta	83	83	36	2.306	36	2.306
11		South East	256	256	109	2.349	109	2.349
12		Southern Central Coastal	7	7	4	1.750	4	1.750
13	Textiles	Mekong River Delta	28	28	17	1.647	17	1.647
14		Red River Delta	222	222	75	2.960	75	2.960
15		South East	270	270	78	3.462	78	3.462
16		Southern Central Coastal	23	23	14	1.643	14	1.643

Source: Author's calculations.

a. This adjustment must include both frame problems because of inaccuracy and expected growth rate of stratum.

Table 4.18

Poststratification by Ownership

		Sample Design				Poststratification			
Strata No.	Sector	Location	Final Population	Final Sample	Final Weight	Post Strata	Population	Sample	Weights
1	Apparel	Mekong River	45	30	1.500	Private	35	25	1.400
2		Delta				FDI	10	5	2.000
3		Red River	280	60	4.667	Private	220	50	4.400
4		Delta				FDI	60	10	6.000
5		South East	650	103	6.311	Private	510	80	6.375
6						FDI	140	23	6.087
7		Southern	50	26	1.923	Private	35	25	1.400
8		Central Coastal				FDI	15	1	15.000
9	Food & beverage	Mekong River	817	118	6.924	Private	615	84	7.321
10		Delta				FDI	202	34	5.941
11		Red River	374	45	8.311	Private	194	39	4.974
12		Delta				FDI	180	6	30.000
13		South East	699	68	10.279	Private	618	59	10.475
14						FDI	81	9	9.000
15		Southern	152	59	2.576	Private	102	25	4.080
16		Central Coastal				FDI	50	34	1.471
17	Leather products	Mekong River	13	8	1.625	Private	13	8	1.625
18		Delta				FDI	0	0	—
19		Red River	83	36	2.306	Private	75	29	2.586
20		Delta				FDI	8	7	1.143
21		South East	256	109	2.349	Private	220	57	3.860
22						FDI	36	52	0.692
23		Southern	7	4	1.750	Private	7	4	1.750
24		Central Coastal				FDI	0	0	—

(*continued*)

Table 4.18 (continued)

Poststratification by Ownership

		Sample Design				Poststratification			
Strata No.	Sector	Location	Final Population	Final Sample	Final Weight	Post Strata	Population	Sample	Weights
25	Textiles	Mekong River Delta	28	17	1.647	Private	28	17	1.647
26						FDI	0	0	—
27		Red River Delta	222	75	2.960	Private	188	61	3.082
28						FDI	34	14	2.429
29		South East	270	78	3.462	Private	220	66	3.333
30						FDI	50	12	4.167
31		Southern Central Coastal	23	14	1.643	Private	23	14	1.643
32						FDI	0	0	—

Source: Author's calculations.

Note: FDI = Foreign direct investment. — = Not applicable.