Random and Stratified Sampling

- **Task 1**

Consider the *Horvitz-Thompson* estimator of the population total

$$\hat{T}_X = \sum_{i=1}^{n} \frac{1}{\pi_i} X_i,$$

proof that the the variance estimator is

$$\widehat{Var}\left(\hat{T}_X\right) = \sum_{i,j} \left(\frac{X_i X_j}{\pi_{ij}} - \frac{X_i}{\pi_i} \frac{X_j}{\pi_j} \right).$$

- **Task 2**

In addition to the mean, another parameter of interest is the *proportion* (or percentage) of the population with a particular attribute. Results for a proportion can follow directly from those for a mean, since a proportion is just a special case of a mean, by setting $Y_i = 1$ if the i-th element has the attribute and $Y_i = 0$ if not.

1. Determine the variance and confidence intervals formulas in the case of simple random sampling for P, the population proportion with a given attribute.

2. Deduce which is the expression for the sample size required to estimate a proportion p with a bound on error of estimation ϵ.

3. Consider the next problem: The major of a certain village wants to conduct a survey to determine the proportion of people that favors a certain decision about a new factory. Since interviewing $N = 2000$ persons is not possible for her, she need to estimate p with a bound of estimation of magnitude $\epsilon = 0.05$. Show which is the best sample size to do it (wihout estimating the variance). Write a program in `R` .

- **Task 3**

Suppose that a researcher is interested in estimating the *proportion* of households that watches a particular TV program. The population is divided into 2 towns and a residential zone. Show and argue about which are the estimators of the population proportion p_{st}, the estimated variance and the bound on the error of estimation.

Suppose that data contain 155 households in town A, 62 in Town B and 93 in the rural area. A simple random sample is taken from each stratum:

Stratum	**Sample size**	**Number of householders viewing the show**
1	$n_1 = 20$	16
2	$n_2 = 8$	2
3	$n_3 = 12$	6

Write a program in `R` to estimate the proportion of households viewing the show and the bound on the error of estimation.

Suppose that the advertising firm wants to conduct a new survey. The proportions may be estimated from the previous survey sample and the cost of obtaining an observation is € 9 for stratum 1 and 2, and € 16 for stratum 3. The firm wants to estimate the population proportion p with a bound on the error equal to 0.1.

Write a program in `R` to estimate the sample size and the strata sample sizes that give the bound at minimum cost.

- **Task 4**

Resume and comment the techniques of *Poststratification* and *Quota Sampling* with Lohr (2009) text.

[redacted].

4.4 Poststratification

Suppose a sampling frame lists all households in an area, and you would like to estimate the average amount spent on food in a month. One desirable stratification variable might be household size because large households might be expected to have higher food bills than smaller households. From U.S. census data, the distribution of household size in the region is known:

Number of Persons in Household	Percentage of Households
1	25.75
2	31.17
3	17.50
4	15.58
5 or more	10.00

The sampling frame, however, does not include information on household size—it only lists the households. Thus, although you know the population size in each subgroup, you cannot take a stratified sample because you do not know the stratum membership of the units in your sampling frame. You can, however, take an SRS and record the amount spent on food as well as the household size for each household in your sample. If n, the size of the SRS, is large enough, then the sample is likely to resemble a stratified sample with proportional allocation: We would expect about 26% of the sample to be one-person households, about 31% to be two-person households, and so on.

Considering the different household-size groups to be different domains, we can use the methods from Section 4.2 to estimate the average amount spent on groceries for each domain. Take an SRS of size n. Let $n_1, n_2, \ldots, n_H$ be the numbers of units in the various household-size groups (domains) and let $\bar{y}_1, \ldots, \bar{y}_H$ be the sample means for the groups. In this case, since the poststrata are formed *after* the sample is taken, the sample domain sizes $n_1, n_2, \ldots, n_H$ are random quantities. If we selected another SRS from the population, the poststratum sizes in the sample would change. Since the poststratum sizes in the population are known, however, we can use the known values of N_h in the estimation.

To see how poststratification fits in the framework of ratio estimation, define $x_{ih} = 1$ if observation i is in poststratum h and 0 otherwise. Let $u_{ih} = y_i x_{ih}$. Then $t_{xh} = \sum_{i=1}^{N} x_{ih} = N_h$ and

$$t_{uh} = \sum_{i=1}^{N} u_{ih} = \text{population total of variable } y \text{ in poststratum } h.$$

For each poststratum h, we can estimate the total in the poststratum by

$$\hat{t}_{uh} = \sum_{i \in \mathcal{S}} \frac{N}{n} u_{ih}$$

[$\hat{t}_{uh}$ is the domain total estimator in (4.14)]. We can then use ratio estimation to obtain:

$$\hat{t}_{uhr} = \frac{t_{xh}}{\hat{t}_{xh}} \hat{t}_{uh} = \frac{N_h}{\hat{N}_h} \hat{t}_{uh} = N_h \bar{y}_h,$$

where $\bar{y}_h$ is the sample mean of the observations in poststratum h.

The poststratified estimator of the population total is

$$\hat{t}_{y\text{post}} = \sum_{h=1}^{H} \hat{t}_{uhr} = \sum_{h=1}^{H} \frac{N_h}{\hat{N}_h} \hat{t}_{uh} = \sum_{h=1}^{H} N_h \bar{y}_h;$$

ratio estimation is used within each poststratum to estimate the population total in that poststratum.

The poststratified estimator of $\bar{y}_U$ is

$$\bar{y}_{\text{post}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h. \tag{4.21}$$

If N_h/N is known, n_h is reasonably large ($\geq$ 30 or so), and n is large, then we can use the variance for proportional allocation as an approximation to the poststratified variance:

$$\hat{V}(\bar{y}_{\text{post}}) \approx \left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} \frac{N_h}{N} \frac{s_h^2}{n}. \tag{4.22}$$

This approximation is valid only when the expected sample sizes in each poststratum are large, however (see Exercise 37).

Many large surveys use poststratification to improve efficiency of the estimators or to correct for the effects of differential nonresponse in the poststrata (see Chapter 8). We discuss poststratification for general survey designs in Section 11.7.

[redacted]

3.7 Quota Sampling

Many samples that masquerade as stratified random samples are actually quota samples. In **quota sampling**, the population is divided into different subpopulations just as in stratified random sampling, but with one important difference: Probability sampling is not used to choose individuals in the subpopulation for the sample. In extreme versions of quota sampling, choice of units in the sample is entirely at the discretion of the interviewer, so that a sample of convenience is chosen within each subpopulation.

In quota sampling, specified numbers (quotas) of particular types of population units are required in the final sample. For example, to obtain a quota sample with $n = 3000$, you might specify that the sample contain 1000 white males, 1000 white females, 500 men of color, and 500 women of color, but you might give no further instructions about how these quotas are to be filled. Thus, quota sampling is not a form of probability sampling—we do not know the probabilities with which each individual is included in the sample. It is often used when probability sampling is impractical, overly costly, or considered unnecessary, or when the persons designing the sample just do not know any better.

The big drawback of quota sampling is that we do not know if the units chosen for the sample exhibit selection bias. If selection of units is totally up to the interviewer, she or he is likely to choose the most accessible members of the population—for instance, persons who are easily reached by telephone, households without menacing dogs, or areas of the forest close to the road. The most accessible members of a population are likely to differ in a systematic way from less accessible members. Thus, unlike in stratified random sampling, we cannot say that the estimator of the population total from quota sampling is unbiased over repeated sampling—one of our usual criteria of goodness in probability samples. In fact, in quota samples, we cannot measure sampling error over repeated samples and we have no way of estimating the bias from the sample data. Since selection of units is up to the individual interviewer, we cannot expect that repeating the sample will give similar results. Thus, anyone drawing inferences from a quota sample must necessarily take a model-based approach.

EXAMPLE 3.13 The 1945 survey on reading habits taken for the Book Manufacturer's Institute (Link and Hopf, 1946), like many surveys in the 1940s and 1950s, used a quota sample. Some of the classifications used to define the quota classes were area, city size, age, sex, and socioeconomic status; a local supervising psychologist in each city determined the blocks of the city in which interviewers were to interview people from a specified socioeconomic group. The interviewers were then allowed to choose the specific households to be interviewed in the designated city blocks.

The quota procedure followed in the survey did not result in a sample that reflected demographic characteristics of the 1945 U.S. population. The following table compares the educational background of the survey respondents with figures from the 1940 U.S. Census, adjusted to reflect the wartime changes in population.

Distribution by Educational Levels	4,000 People Interviewed (%)	U.S. Census, Urban and Rural Nonfarm (%)
8th grade or less	28	48
1–3 years high school	18	19
4 years high school	25	21
1–3 years college	15	7
4 or more years college	13	5

SOURCE: Link and Hopf (1946).

The oversampling of better-educated persons casts doubt on many of the statistics given in the book. The study concluded that 31% of "active readers" (those who had read at least one book in the past month) had bought the last book they read, and that 25% of all last books read by active readers cost \$1 or less. Who knows whether a stratified random sample would have given the same results? ■

In the 1948 U.S. presidential elections, all of the major polls printed just a few days before the election predicted that Dewey would defeat Truman handily. In fact, of course, Truman won the election. According to Mosteller et al. (1949), one of the problems of those polls was that they all used quota sampling, not a probability-based method—the polling debacle in 1948 spurred many survey organizations in the United States to turn away from quota sampling, at least for a few years. The polls that erred in predicting the winner in the British general election of 1992 all used quota methods in selecting persons to interview in their homes or in the street; the primary quota classes used were sex, age, socio-economic class, and employment status. Although we may never know exactly what went wrong in those polls (see Crewe, 1992, for some other explanations), the use of quota samples may have played a part—if interviewing persons "in the street," it is certainly plausible that persons from a quota class that are accessible differ from persons that are less accessible.

While quota sampling is not as good as probability sampling under ideal conditions, it may give better results than a completely haphazard sample because it at least forces the inclusion of members of the different quota groups. Quota samples have the advantage of being less expensive than probability samples. The quality of the data from quota samples can be improved by allowing the interviewer less discretion in the choice of persons or households to be included in the sample. Many survey organizations use probability sampling along with quotas; they use probability sampling to select small blocks of potential respondents, and then take a quota sample within each block, using variables such as age, sex, and race.

Because we do not know the probabilities with which units were sampled, we must take a model-based approach, and make strong assumptions about the data structure, when analyzing data from a quota sample. The model generally adopted is that of Section 3.6—within each subclass the random variables generating the subpopulation are assumed independent and identically distributed. Such a model implies that any selection of units from the quota class will give a representative sample; if the model holds, then quota sampling will likely give good estimates of the population quantity. If the model does not hold, then the estimates from quota sampling may be badly biased.

EXAMPLE 3.14 Sanzo et al. (1993) used a combination of stratified random sampling and quota sampling for estimating the prevalence of *Coxiella burnetii* infection within the Basque country in northern Spain. *Coxiella burnetii* can cause Q fever, which can lead to complications such as heart and nerve damage. Reviews of Q fever patient records from Basque hospitals showed that about three-fourths of the victims were male, about half were between 16 and 30 years old, and victims were disproportionately likely to be from areas with low population density.

The authors stratified the target population by population density and then randomly selected health care centers from the three strata. In selecting persons for blood testing, however, "a probabilistic approach was rejected as we considered that the refusal rate of blood testing would be high" (p. 1185). Instead, they used quota sampling to balance the sample by age and gender; physicians asked patients who needed laboratory tests whether they would participate in the study, and recruited subjects for the study until the desired sample sizes in the six quota groups were reached for each stratum.

Because a quota sample was taken instead of a probability sample, persons analyzing the data must make strong assumptions about the representativeness of the sample in order to apply the results to the general population of the Basque country. First, the assumption must be made that persons attending a health clinic for laboratory tests (the sampled population of the study) are neither more nor less likely to be infected than persons who would not be visiting the clinic. Second, one must assume that persons who are requested and agree to do the study are similar in terms of the infection to persons in the same quota class having laboratory tests that do not participate in the study. These are strong assumptions: the authors of the article argue that the assumptions are justified, but of course they cannot prove that the assumptions hold unless follow-up investigations are done.

If they had taken a probability sample of persons instead of the quota sample, they would not have had to make these strong assumptions. A probability sample of persons, however, would have been exhorbitantly expensive when compared with the quota sampling scheme used, and a probability sample would also have taken longer to design and implement. With the quota sample, the authors were able to collect information about the public health problem; it is unclear whether the results can be generalized to the entire population, but the data do provide a great deal of quick information on the prevalence of infection that can be used in future investigation of who is likely to be infected, and why. ■

Deville (1991, p. 177) argues that quota samples may be useful for market research, when the organization requesting the survey is aware of the model being used. Persons collecting official statistics about crime, unemployment, or other matters that are used for setting public policy should use probability samples, however.

Quota samples, while easier to collect than a probability sample, suffer from the same disadvantages as other convenience samples. Some survey organizations now use quota sampling to recruit volunteers for online surveys; they accumulate respondents until they have specified sample sizes in the desired demographic classes. In such online surveys, the respondents in each quota class are self-selected—if, as argued by Couper (2000), Internet users who volunteer for such surveys differ from members of the target population in those quota classes, results will be biased.

3.8 Chapter Summary

Stratification uses additional information about a population in the survey design. In the simplest form, stratified random sampling, we take an SRS of size n_h in stratum h, for each of the H strata in the population. To use stratification, we must know the population size N_h for each stratum; we must also know the stratum membership for every unit in the population. The inclusion probability for unit i in stratum h is $\pi_{hi} = n_h/N_h$; consequently, the sampling weight for that unit is $w_{hi} = N_h/n_h$.

To estimate the population total t using a stratified random sample, let $\hat{t}_h$ estimate the population total in stratum h. Then

$$\hat{t}_{\text{str}} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj}$$