Example 5.13. *Sampling for hazardous waste sites.* Many dumps and landfills contain toxic materials. These materials may have been sealed in containers when deposited, but may now be suspected of leaking. But we no longer know where the materials were deposited—containers of hazardous waste may be randomly distributed throughout the landfill, or they may be concentrated in one area, or there may be none at all.

A common practice is to take a systematic sample of grid points and to take soil samples from each to look for evidence of contamination. Choose a point at random in the area, then construct a grid containing that point so that grid points are an equal distance apart. One such grid is shown in Figure 5.7(a). The advantages of taking a systematic sample rather than an SRS are that the systematic sample forces an even coverage of the region and is easier to implement in the field. If you are not worried about periodic patterns in the distribution of toxic materials, and you have little prior knowledge of where the toxic materials might be, a systematic sample is a good design.

With any grid in systematic sampling, you need to worry if the toxic materials are regularly placed so that the grid may miss all of them, as shown in Figure 5.7(b). If this is a concern, you would be better off taking a stratified sample. Lay out the grid, but select a point at random in each square at which to take the soil sample. ■

If periodicity is a concern in a population, one solution is to use **interpenetrating systematic samples** (Mahalanobis, 1946). Instead of taking one systematic sample, take several systematic samples from the population. Then you can use the formulas for cluster samples to estimate variances; each systematic sample acts as one psu. This approach is explored in Exercise 24.

5.6 Model-Based Theory for Cluster Sampling*

In most cluster samples, observations within the same cluster are more similar than observations selected randomly from the population as a whole. A sample of n clusters with m observations observed per cluster gives less information than an SRS of nm observations because the observations within a cluster are dependent. Any model for cluster sampling must include this dependence explicitly.

The one-way ANOVA model with fixed effects provides a theoretical framework for stratified sampling; an analogous model for cluster sampling is the one-way ANOVA model with random effects (Scott and Smith, 1969). Let's look at a simple version of this model:

$$\text{M1: } Y_{ij} = \mu + A_i + \varepsilon_{ij} \tag{5.38}$$

with A_i generated by a distribution with mean 0 and variance σ_A^2, ε_{ij} generated by a distribution with mean 0 and variance σ^2, and all A_is and ε_{ij}s independent.

A random effects model such as M1 allows observations in the same cluster to be positively correlated by specifying a probability distribution for the cluster means. With Model M1,

$$\text{Cov}_{\text{M1}}\left(Y_{ij}, Y_{kl}\right) = \begin{cases} \sigma^2 + \sigma_A^2 & \text{if } i = k \text{ and } j = l \\ \sigma_A^2 & \text{if } i = k \text{ and } j \neq l \\ 0 & \text{if } i \neq k \end{cases} .$$

The model-based intraclass correlation coefficient equals

$$\rho = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2}. \tag{5.39}$$

Note that ρ in (5.39) is always nonnegative, in contrast to the design-based ICC, which can take on negative values.

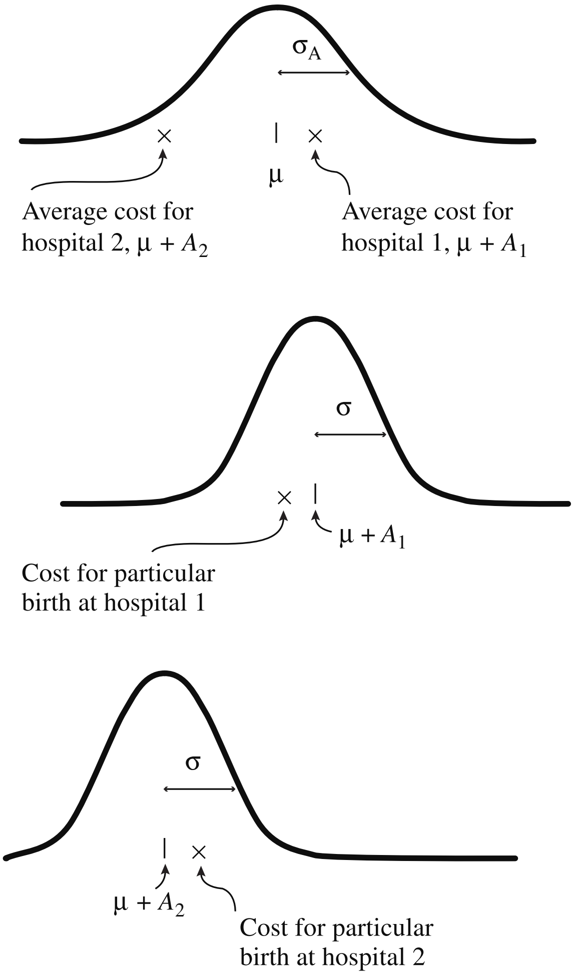
σ_A
μ
Average cost for hospital 2, $\mu + A_2$
Average cost for hospital 1, $\mu + A_1$
σ
$\mu + A_1$
Cost for particular birth at hospital 1
σ
$\mu + A_2$
Cost for particular birth at hospital 2

FIGURE 5.8
Illustration of random effects for hospitals and births.

Figure 5.8 illustrates Model M1, assuming that all random variables are normally distributed, for a two-stage cluster sample taken to estimate the total amount of hospital charges in a country for delivering babies. Hospitals are sampled at the first stage, and birth records from the selected hospitals are sampled at the second stage (twins and triplets count as one record). The average charge per birth varies from hospital to hospital—some hospitals may have higher personnel costs, and others may serve a higher-risk population or have more expensive equipment. That variation is reflected in model M1 by the random variables A_i: $\mu + A_i$ represents the average cost per birth in the ith hospital, and σ_A^2 is the population variance among the hospital means. In addition, costs vary from birth to birth within the hospitals; that variation is incorporated into the model by the term ε_{ij} with variance σ^2.

Costs for births in the same hospital tend to be more similar than costs for births selected randomly across the entire population of hospital births, because the cost for a birth in a given hospital incorporates the hospital-specific characteristics such as personnel costs. This similarity induces a positive correlation among observations in the same cluster.

5.6.1 Estimation Using Models

Now let's find properties of various estimators under Model M1. As in Sections 2.10 and 3.6, we want to predict the unsampled population members in the random variable representing the finite population total, which for cluster sampling is

$$T = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij}.$$

All of the estimators we consider have the form

$$\hat{T} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} b_{ij} Y_{ij}, \tag{5.40}$$

where the b_{ij}s are a set of constants, so let's derive the model-based bias and variance for this general estimator and then substitute the values of b_{ij} for specific estimators later.

Inference in a model-based approach is conditional on the units selected to be in the sample; that is, the inference treats $\mathcal{S}$ and $\mathcal{S}_i$ as fixed, and treats Y_{ij} as a random variable. The bias of the general estimator $\hat{T}$ under model M1 is

$$E_{\text{M1}}[\hat{T} - T] = E_{\text{M1}}\left[\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} b_{ij} Y_{ij} - \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij}\right] = \mu \left(\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} b_{ij} - M_0\right). \tag{5.41}$$

Thus, $\hat{T}$ is model-M1-unbiased when $\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} b_{ij} = M_0$. The model-based (for model M1) variance of $\hat{T} - T$ is (see Exercise 34):

$$V_{\text{M1}}[\hat{T} - T] = \sigma_A^2 \left[\sum_{i \in \mathcal{S}} \left(\sum_{j \in \mathcal{S}_i} b_{ij} - M_i\right)^2 + \sum_{i \notin \mathcal{S}} M_i^2\right] + \sigma^2 \left[\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} (b_{ij}^2 - 2b_{ij}) + M_0\right]. \tag{5.42}$$

Properties of design-based estimators under Model M1. Now let's look at what happens with the design-based estimators we studied in Section 5.3 under Model M1. The random variable corresponding to the design-unbiased estimator $\hat{t}_{\text{unb}}$ in (5.21) is

$$\hat{T}_{\text{unb}} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} \frac{NM_i}{nm_i} Y_{ij};$$

the coefficients b_{ij} are the sampling weights $(NM_i)/(nm_i)$. Using (5.41), the model-based bias of $\hat{T}_{\text{unb}}$ is

$$E_{\text{M1}}[\hat{T}_{\text{unb}} - T] = \mu \left(\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} \frac{NM_i}{nm_i} - M_0\right) = \mu \left(\frac{N}{n} \sum_{i \in \mathcal{S}} M_i - M_0\right).$$

The bias depends on which sample is taken, and the estimator is model-unbiased under (5.38) only when the average of the M_is in the sample equals the average of the M_is in the population, such as will occur when all M_is are the same.

The ratio estimator corresponding to $M_0\hat{\bar{y}}_r$, for $\hat{\bar{y}}_r$ in (5.16), is

$$\hat{T}_r = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}\left(M_0\frac{M_i}{m_i\sum_{k\in\mathcal{S}}M_k}\right)Y_{ij}.$$

Since $\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(M_0M_i)/(m_i\sum_{k\in\mathcal{S}}M_k) = M_0$, $\hat{T}_r$ is model-unbiased under Model M1. If the sample consists of the population units with the largest values of M_i, the ratio estimator compensates by dividing by the sum of the M_i for the sample, which will also be large.

Best linear unbiased predictor under Model M1. If Model M1 really does describe the population and the M_is are unequal, one can find an estimator with smaller variance than both $\hat{T}_{\text{unb}}$ and $\hat{T}_r$.

In a model-based perspective, it is desired to predict the values of Y_{ij} for the unobserved population members. Exercises 35 and 36 show that among all model-unbiased predictors having the form of (5.40), $\hat{T}_{\text{BLUP}}$ in (5.43) has the smallest variance:

$$\begin{aligned}\hat{T}_{\text{BLUP}} &= \sum_{i\in\mathcal{S}}\left(\sum_{j\in\mathcal{S}_i}Y_{ij} + \sum_{j\notin\mathcal{S}_i}\hat{Y}_{ij}\right) + \sum_{i\notin\mathcal{S}}\sum_{j=1}^{M_i}\hat{Y}_{ij}\\ &= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}Y_{ij} + \sum_{i\in\mathcal{S}}\sum_{j\notin\mathcal{S}_i}\left[\rho\alpha_i\frac{1}{m_i}\sum_{l\in\mathcal{S}_i}Y_{il} + (1-\rho\alpha_i)\hat{\mu}\right] + \sum_{i\notin\mathcal{S}}M_i\hat{\mu}, \qquad (5.43)\end{aligned}$$

where $\alpha_i = m_i/[1+\rho(m_i-1)]$ and

$$\hat{\mu} = \left(\sum_{i\in\mathcal{S}}\alpha_i\sum_{j\in\mathcal{S}_i}\frac{Y_{ij}}{m_i}\right)\Bigg/\left(\sum_{i\in\mathcal{S}}\alpha_i\right). \qquad (5.44)$$

Typically, ρ is unknown, and an estimator $\hat{\rho}$ is substituted into (5.43) and (5.44).

The unknown mean μ is estimated by a weighted average of the sample psu means, with α_i proportional to $1/V_{\text{M1}}(\sum_{j\in\mathcal{S}_i}Y_{ij}/m_i)$. Thus, the estimated mean $\hat{\mu}$ relies more heavily on the sample means from psus with larger sample sizes m_i, which have smaller variance under model M1. If psu i is not in the sample, then $\hat{Y}_{ij} = \hat{\mu}$. If psu i is in the sample but ssu j is not, $\hat{Y}_{ij}$ is predicted as a value somewhere between the sample mean of psu i and the overall mean $\hat{\mu}$.

If n/N is small, $\hat{T}_{\text{BLUP}} \approx M_0\hat{\mu}$. Under model M1, after all, every population unit has expected value μ, so if the number of psus in the sample is small relative to the number of psus in the population, it makes sense that the population mean would be approximately estimated by $\hat{\mu}$ and the total by the population size times $\hat{\mu}$.

If $M_i = M$ and $m_i = m$ for all i, then $\hat{\mu} = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}Y_{ij}/(nm)$, $\hat{T}_{\text{unb}} = \hat{T}_r = \hat{T}_{\text{BLUP}} = M_0\hat{\mu}$, and the variance in (5.42) simplifies to

$$V_{\text{M1}}[\hat{T}_{\text{unb}} - T] = M_0^2\left(1-\frac{n}{N}\right)\frac{\sigma_A^2}{n} + M_0^2\left(1-\frac{nm}{NM}\right)\frac{\sigma^2}{mn}. \qquad (5.45)$$

Model assumptions. The assumptions for model-based inference are strong. For Model M1, we assume:

1. The model form is correct, that is, $Y_{ij} = \mu + A_i + \varepsilon_{ij}$ for every unit (i, j) in the population, whether sampled or not. One consequence is that the expected value of the total in psu i, $E\left[\sum_{j=1}^{M_i}Y_{ij}\right]$, is assumed to equal $M_i\mu$.

2. The proposed variance and correlation structures are correct, that is, $V(Y_{ij}) = \sigma_A^2 + \sigma^2$, $\text{Corr}\,(Y_{ij}, Y_{il}) = \rho$ for units $j \neq l$ in the same cluster, and $\text{Corr}\,(Y_{ij}, Y_{kl}) = 0$ for units in different clusters.

In a model-based approach, it does not matter which population units are in the sample because all population units are assumed to follow the model. For that reason, data from a nonprobability sample must be analyzed using models, as discussed in Chapter 15.

But what happens if the model is inappropriate for the population? We can (and should!) perform model diagnostics (see, for example, Demidenko, 2013; Loy et al., 2017) to examine how well the model fits the units in the sample, but unless additional information is available about the units not sampled, we cannot assess whether the model is appropriate for unobserved units. A model-based analyst must assume that the model fits the population units that are not observed. If psus and ssus are selected using simple random sampling, then it is reasonable to assume that a model that fits the observed units also fits the unobserved units since there is nothing "special" about the units that end up in the sample.

With a convenience sample, however, the psus in the sample may differ systematically from those not in the sample—for example, the sampled hospitals in Figure 5.8 might be larger or more urban than nonsampled hospitals—and hence predictions of the unobserved records in unsampled hospitals may be biased. The model-based variance, estimated from the data in the sample, does not capture that bias. Thus, if the model does not fit the unsampled units, the model-based variance can severely underestimate the mean squared error and give the impression that estimates are more accurate than they really are.

That said, Model M1 is a reasonable approximation for many real-life situations in which there is clustering, and it is widely used in practice for nonprobability samples of clusters. Available auxiliary information can be included in the model to improve prediction, as in Section 4.6. Of course, many other models have been proposed that assume units are dependent; see the For Further Reading section for references.

You can use various estimators in conjunction with a model-based approach. Thus, you can estimate the population total by $\hat{T}_{\text{unb}}$, $\hat{T}_r$, $\hat{T}_{\text{BLUP}}$, or another estimator. But the statistical properties of the estimator come from the assumed model. In practice, if estimating multiple quantities or statistics for domains, it may be desirable to use an estimator in which the constants b_{ij} are the same for every y variable, even if the variance is higher than that of $\hat{T}_{\text{BLUP}}$. The constants used in (5.43), which are given explicitly in Exercise 36, can differ for each response variable when the m_i vary from psu to psu. Thus, if you use $\hat{T}_{\text{BLUP}}$ to estimate the total hospital charges for California, the total hospital charges for Oregon, and the total hospital charges for California and Oregon together, the combined California–Oregon total might not equal the sum of the individual totals for the two states. If the constants b_{ij} are the same for every response variable, this problem will not occur.

Example 5.14. Let's fit Model M1 to the schools data from Example 5.7. Looking at Figure 5.3, it seems plausible that the within-school variances σ_i^2 are the same for all schools. Using statistical software for mixed models, we obtain estimated variance components $\hat{\sigma}_A^2 = 33.85$ and $\hat{\sigma}^2 = 102.83$. This results in $\hat{\rho} = 33.85/(33.85 + 102.83) = 0.25$, approximately the same as the value of R_a^2 in Example 5.8. Most software packages for mixed models estimate μ using (5.44). For the schools data, $m_i = 20$ students are sampled from each psu, so $\hat{\mu} = 34.66$ is the average of the individual school means in Table 5.7. The model-based standard error of $\hat{\mu}$ is 1.97. Note that $\hat{\mu}$ is slightly larger than $\hat{\bar{y}}_r = 33.12$ from Example 5.7. The model-based standard error, however, differs from the design-based standard error of 1.76; the model-based standard error estimates variability under M1, while the design-based standard error estimates variability under repeated sampling with the sampling design. ■

5.6.2 Design Using Models

Models are extremely useful for designing a cluster sample. Using a model for design does not mean you have to use a model for analysis of your survey data after it is collected; rather, the model provides a useful way of summarizing information you can use to make the survey design more efficient.

Suppose that Model M1 seems reasonable for your population and that all psu sizes in the population are equal. Then you would like to design the survey to minimize the variance in (5.45), subject to cost constraints. Using the cost function in (5.35), the model-based variance is minimized when

$$m = \sqrt{\frac{c_1\sigma^2}{c_2\sigma_A^2}}.$$

Suppose that the M_is are unequal and that Model M1 holds. We can use the variance in (5.42) to determine the optimal subsampling size m_i for each cluster. This approach was used by Royall (1976) for more general models than considered in this section. For $\hat{T}_r$, $b_{ij} = M_i/(m_i \sum_{k\in\mathcal{S}} M_k)$, and the variance is minimized when m_i is proportional to M_i (see Exercise 39).

5.7 Chapter Summary

Cluster sampling is commonly used in large surveys, but estimates obtained from cluster samples usually have greater variance than if we were able to measure the same number of observation units using an SRS. If it is much less expensive to sample clusters than individual elements, though, cluster sampling can provide more precision per dollar spent.

All of the formulas in this chapter for cluster sampling with equal probabilities are special cases of the general results for two-stage cluster sampling with unequal psu sizes, to be derived in Chapter 6. They can be applied to any two-stage cluster sample in which the psus are selected with equal probability. These formulas were given in (5.21), (5.27), (5.30), and (5.32) and are repeated here:

$$\hat{t}_{\text{unb}} = \frac{N}{n}\sum_{i\in\mathcal{S}} \hat{t}_i = \frac{N}{n}\sum_{i\in\mathcal{S}} M_i \bar{y}_i, \tag{5.46}$$

$$\hat{V}(\hat{t}_{\text{unb}}) = N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\sum_{i\in\mathcal{S}}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}, \tag{5.47}$$

$$\hat{\bar{y}}_r = \frac{\displaystyle\sum_{i\in\mathcal{S}} M_i\bar{y}_i}{\displaystyle\sum_{i\in\mathcal{S}} M_i}, \tag{5.48}$$

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{\overline{M}^2}\left[\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN}\sum_{i\in\mathcal{S}}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}\right], \tag{5.49}$$

with $\overline{M} = \sum_{i\in\mathcal{S}} M_i/n$,

$$s_t^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N}\right)^2$$