question. Of the remaining 134 voters, 112 opposed the incinerator, so the council estimates the proportion by

$$\hat{p} = 112/134 = .83582$$

with

$$\hat{V}(\hat{p}) = .83582(1 - .83582)/134 = 0.00102.$$

Are these estimates valid? Why, or why not?

2 Senturia et al. (1994) describe a survey taken to study how many children have access to guns in their households. Questionnaires were distributed to all parents who attended selected clinics in the Chicago area during a one-week period for well or sick child visits.

a Suppose that the quantity of interest is percentage of the households with guns. Describe why this is a cluster sample. What is the psu? The ssu? Is it a one-stage or two-stage cluster sample? How would you estimate the percentage of households with guns, and the standard error of your estimate?

b What is the sampling population for this study? Do you think this sampling procedure results in a representative sample of households with children? Why, or why not?

3 Kleppel et al. (2004) report on a study of wetlands in upstate New York. Four wetlands were selected for the study: Two of the wetlands drain watersheds from small towns and the other two drain suburban watersheds. Quantities such as pH were measured at two to four randomly selected sites within each of the four wetlands.

a Describe why this is a cluster sample. What are the psus? The ssus? How would you estimate the average pH in the suburban wetlands?

b The authors used Student's two-sample *t* test to compare the average pH from the sites in the suburban wetlands with the average pH from the sites in the small town wetlands, treating all sites as independent. Is this analysis appropriate? Why, or why not?

4 Survey evidence is often introduced in court cases involving trademark violation and employment discrimination. There has been controversy, however, about whether nonprobability samples are acceptable as evidence in litigation. Jacoby and Handlin (1991) selected 26 from a list of 1285 scholarly journals in the social and behavioral sciences. They examined all articles published during 1988 for the selected journals, and recorded (1) the number of articles in the journal that described empirical research from a survey (they excluded articles in which the authors analyzed survey data which had been collected by someone else) and (2) the total number of articles for each journal which used probability sampling, nonprobability sampling, or for which the sampling method could not be determined. The data are in file journal.dat.

a Explain why this is a cluster sample.

b Estimate the proportion of articles in the 1285 journals that use nonprobability sampling, and give the standard error of your estimate.

c The authors conclude that, because "an overwhelming proportion of … recognized scholarly and practitioner experts rely on non-probability sampling designs," courts "should have no problem admitting otherwise well-conducted non-probability surveys and according them due weight" (p. 175). Comment on this statement.

B. Working with Survey Data

11 An accounting firm is interested in estimating the error rate in a compliance audit it is conducting. The population contains 828 claims, and the firm audits an SRS of 85 of those claims. In each of the 85 sampled claims, 215 fields are checked for errors. One claim has errors in 4 of the 215 fields, 1 claim has 3 errors, 4 claims have 2 errors, 22 claims have 1 error, and the remaining 57 claims have no errors. (Data courtesy of Fritz Scheuren.)

a Treating the claims as psus and the observations for each field as ssus, estimate the error rate, defined to be the average number of errors per field, along with the standard error for your estimate.

b Estimate (with standard error) the total number of errors in the 828 claims.

c Suppose that instead of taking a cluster sample, the firm had taken an SRS of 85× 215 = 18,275 fields from the 178,020 fields in the population. If the estimated error rate from the SRS had been the same as in (a), what would the estimated variance $\hat{V}(\hat{p}_{\text{SRS}})$ be? How does this compare with the estimated variance from (a)?

12 Use the data in coots.dat to estimate th

Show that the above definition is equivalent to (5.8). HINT: First show that

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}^{M}(y_{ij}-\bar{y}_U)(y_{ik}-\bar{y}_U)+\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_U)^2 = M(\text{SSB}).$$

23 For the quantities in the population ANOVA table (Table 5.1), show that

$$\text{MSW} = \frac{NM-1}{NM}S^2(1-\text{ICC})$$

and

$$\text{MSB} = \frac{NM-1}{M(N-1)}S^2[1+(M-1)\text{ICC}].$$

24 Suppose in a two-stage cluster sample that all population cluster sizes are equal ($M_i = M$ for all i), and that all sample sizes for the clusters are equal ($m_i = m$ for all i).

a Show (5.30).

b Show that MSW = $S^2(1-R_a^2)$ and that

$$\text{MSB} = S^2\left[\frac{N(M-1)R_a^2}{N-1}+1\right].$$

c Using (a) and (b), express $V(\hat{\bar{y}})$ as a function of n, m, N, M, and R_a^2.

d Show that if S^2 and the sample and population sizes are fixed, and if $(m-1)/m > n/N$, then $V(\hat{\bar{y}})$ is an increasing function of R_a^2.

25 Suppose in a two-stage cluster sample that all population cluster sizes are equal ($M_i = M$ for all i), and that all sample sizes for the clusters are equal ($m_i = m$ for all i).

a Show that $\hat{t}_{\text{unb}} = \hat{t}_r$, and, hence, that $\hat{\bar{y}}_{\text{unb}} = \hat{\bar{y}}_r$.

b Fill in the formulas for the sums of squares in the ANOVA table below, for the sample data.

Source	df	Sum of Squares	Mean Square
Between psus	$n-1$		msb
Within psus	$n(m-1)$		msw
Total	$nm-1$		msto

c Show that $E[\text{msw}]$ = MSW and

$$E[\text{msb}] = \frac{m}{M}\text{MSB}+\left(1-\frac{m}{M}\right)\text{MSW},$$

where MSB and MSW are the between and within mean squares, respectively, from the *population* ANOVA table given in Table 5.1.

d Show that

$$\widehat{\text{MSB}} = \frac{M}{m}\text{msb}-\left(\frac{M}{m}-1\right)\text{msw}$$

is an unbiased estimator of MSB.

e Show, using (5.24) or (5.28), that

$$\hat{V}(\hat{\bar{y}}_{\text{unb}}) = \left(1 - \frac{n}{N}\right)\frac{\text{msb}}{nm} + \frac{1}{N}\left(1 - \frac{m}{M}\right)\frac{\text{msw}}{m}.$$

26 For the situation in Exercise 25, let msto represent the mean square total from the sample ANOVA table.

a Write msto as a function of msb and msw, and use the results of Exercise 25(c) to find $E[\text{msto}]$.

b Show that $E[\text{msto}] \approx S^2$ if n and N are large.

c Show that

$$\hat{S}^2 = \frac{M(N-1)}{m(NM-1)}\text{msb} + \frac{(m-1)NM + M - m}{m(NM-1)}\text{msw}$$

is an unbiased estimator of S^2.

27 (Requires calculus.) Show that if $M_i = M$ and $m_i = m$ for all i, and if the cost function is $C = c_1 n + c_2 nm$, then

$$m_{\text{opt}} = \sqrt{\frac{c_1 M(\text{MSW})}{c_2(\text{MSB} - \text{MSW})}} = \sqrt{\frac{c_1 M(N-1)(1-R_a^2)}{c_2(NM-1)R_a^2}}$$

minimizes the variance for fixed total cost C. HINT: Show the result with MSW and MSB first, then use Exercise 24(b).

28 (Requires trigonometry.) In Example 5.12, a systematic sampling scheme was proposed for detecting hazardous wastes in landfills. How far apart should sampling points be placed? Suppose that if there is leakage, it will spread to a circular region with radius R. Let D be the distance between adjacent sampling points in the same row or column.

a Calculate the probability with which a contaminant will be detected. HINT: Consider three cases, with $R < D$, $D \le R \le \sqrt{2}D$, and $R > \sqrt{2}D$.

b Propose a sampling design that gives a higher probability that a contaminant will be detected than the square grid, but does not increase the number of sampling points.

29 (Requires knowledge of random effects models.) Under Model M1 in (5.34), a one-way random effects model, the intraclass correlation coefficient ρ may be estimated by

$$\hat{\rho} = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}^2},$$

where $\hat{\sigma}_A^2$ and $\hat{\sigma}^2$ estimate the variance components σ_A^2 and σ^2. The methods of moments estimators for one-stage cluster sampling when all clusters are of the same size are $\hat{\sigma}^2 = \text{msw}$ and $\hat{\sigma}_A^2 = (\text{msb} - \text{msw})/M$, where msw and msb are the within and between mean squares from the sample ANOVA table.

a What is $\hat{\rho}$ in Example 5.4? How does it compare with $\widehat{\text{ICC}}$?

b Calculate $\hat{\rho}$ for Populations A and B in Example 5.3. Why do these differ from the ICC?

38 *IPUMS exercises.*

a Generate a frequency table of the number of persons within each psu.

b Suppose that it costs $50 per interview to collect data using an SRS. If a cluster sample is taken, it costs $100 per psu chosen, plus $20 for each interview taken. Select an SRS of 10 psus. In each of the selected psus, take a subsample of persons with sample size proportional to the population size within that psu. Your total cost for the sample should be about the same as for the SRS you took in Chapter 2.

c Using the sample you selected, estimate the population mean of *inctot* and give the standard error of your estimate. Also estimate the population total of *inctot* and give its standard error. How do these estimates compare with those from the SRS you took in Chapter 2?

5.4.2 Choosing Subsampling Sizes

The goal in designing a survey is generally to get the most information possible for the least cost and inconvenience. In this section, we concentrate on designing a two-stage cluster survey when all psus have the same number, M, of ssus; designing cluster samples will be treated more generally in Chapters 6 and 7. One approach for equal-sized clusters, discussed in Cochran (1977), is to minimize the variance in (5.21) for a fixed cost. If $M_i = M$ and $m_i = m$ for all psus, then $V(\hat{\bar{y}}_{\text{unb}})$ may be rewritten (see Exercise 24) as:

$$V(\hat{\bar{y}}_{\text{unb}}) = \left(1 - \frac{n}{N}\right)\frac{\text{MSB}}{nM} + \left(1 - \frac{m}{M}\right)\frac{\text{MSW}}{nm} \tag{5.30}$$

where MSB and MSW are the between and within mean squares, respectively, in Table 5.1, the population ANOVA table.

If MSW $= 0$ and hence $R_a^2 = 1$, for R_a^2 defined in (5.11), then each element within a psu equals the psu mean. In that case you may as well take $m = 1$; examining more than one element per psu just costs extra time and money without increasing precision. For other values of R_a^2, the optimal allocation depends on the relative costs of sampling psus and ssus.

Consider the simple cost function

$$\text{total cost} = C = c_1 n + c_2 nm, \tag{5.31}$$

where c_1 is the cost per psu (not including the cost of measuring ssus) and c_2 is the cost of measuring each ssu. One can easily determine, using calculus, that the values

$$n_{\text{opt}} = \frac{C}{c_1 + c_2 m_{\text{opt}}}$$

and

$$m_{\text{opt}} = \sqrt{\frac{c_1 M(N-1)(1-R_a^2)}{c_2(NM-1)R_a^2}} \tag{5.32}$$

minimize the variance for fixed total cost C under this cost function (see Exercise 27); often, though, a number of different values will work about equally well, and graphing the projected variance of the estimator will give more information than merely computing one fixed solution. A graphical approach also allows you to perform what-if analyses on the designs: What if the costs or the cost function are slightly different? Or the value of R_a^2 is changed slightly? You can also explore different cost functions with this approach.

In (5.32), the value R_a^2 is from the population ANOVA table. In practice, we can estimate it from pilot survey data by $\hat{R}_a^2 = \widehat{\text{MSW}}/\hat{S}^2$. In large populations, the ratio $M(N-1)/(NM-1)$ will be close to 1, so we can use $\hat{m}_{\text{opt}} = \sqrt{c_1(1-\hat{R}_a^2)/(c_2\hat{R}_a^2)}$.

EXAMPLE 5.10 Would subsampling have been more efficient for Example 5.2 than the one-stage cluster sample that was used? We do not know the population quantities, but have population mean. Again, the y's of Chapter and the x's are the psu sizes M_i. As in (5.15),

$$\hat{\bar{y}}_r = \frac{\displaystyle\sum_{i\in\mathcal{S}} \hat{t}_i}{\displaystyle\sum_{i\in\mathcal{S}} M_i} = \frac{\displaystyle\sum_{i\in\mathcal{S}} M_i\bar{y}_i}{\displaystyle\sum_{i\in\mathcal{S}} M_i}. \tag{5.26}$$

Using the sampling weights in (5.19) with $w_{ij} = (NM_i)/(nm_i)$, we can rewrite $\hat{\bar{y}}_r$ as

$$\hat{\bar{y}}_r = \frac{\hat{t}_{\text{unb}}}{\hat{M}_0} = \frac{\displaystyle\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}y_{ij}}{\displaystyle\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}}. \tag{5.27}$$

The weights are different, but the form of the estimator is the same as in (5.16). The variance estimator is again based on the approximation in (4.10):

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{\bar{M}^2}\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN\bar{M}^2}\sum_{i\in\mathcal{S}} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}, \tag{5.28}$$

where s_i^2 is defined in (5.23),

$$s_r^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}} (M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2, \tag{5.29}$$

and $\bar{M}$ is the average psu size. As with $\hat{t}_{\text{unb}}$, the second term in (5.28) is usually negligible compared with the first term, and most survey software packages calculate the variance using only the first term.

The data in the file

on the units included in the sample. This is more easily done in the general setting of unequal probability sampling; to avoid proving the same result twice, we shall prove the general result in Section 6.6.[1]

To estimate $V(\hat{t}_{\text{unb}})$, let

$$s_t^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N}\right)^2 \tag{5.22}$$

be the sample variance among the estimated psu totals and let

$$s_i^2 = \frac{1}{m_i-1}\sum_{j\in\mathcal{S}_i}(y_{ij}-\bar{y}_i)^2 \tag{5.23}$$

be the sample variance of the ssus sampled in psu i. As will be shown in Section 6.6, an unbiased estimator of the variance in (5.21) is given by

$$\hat{V}(\hat{t}_{\text{unb}}) = N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\sum_{i\in\mathcal{S}}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}. \tag{5.24}$$

The standard error, $\text{SE}(\hat{t}_{\text{unb}})$, is of course the square root of (5.24).

Remark. In many situations when N is large, the contribution of the second term in (5.24) to the variance estimator is negligible compared with that of the first term. We show in Section 6.6 that

$$E[s_t^2] = S_t^2 + \frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{S_i^2}{m_i}.$$

We expect the sample variance of the estimated psu totals $\hat{t}_i$ to be larger than the sample variance of the true psu totals t_i because $\hat{t}_i$ will be different if we take a different subsample in psu i. Thus, if N is large, the first term in (5.24) is approximately unbiased for the theoretical variance in (5.21). To simplify calculations, most software packages for analyzing survey data (including SAS software) estimate the variance using only the first term of (5.24), often omitting the finite population correction (fpc), $(1 - n/N)$. The estimator

$$\hat{V}_{\text{WR}}(\hat{t}_{\text{unb}}) = N^2\frac{s_t^2}{n} \tag{5.25}$$

estimates the with-replacement variance for a cluster sample, as will be seen in Section 6.3. If the first-stage sampling fraction n/N is small, there is little difference between the variance from a with-replacement sample and that from a without-replacement sample. Alternatively, a replication method of variance estimation from Chapter 9 can be used.

If we know the total number of elements in the population, M_0, we can estimate the population mean by $\hat{\bar{y}}_{\text{unb}} = \hat{t}_{\text{unb}}/M_0$ with standard error $\text{SE}(\hat{\bar{y}}_{\text{unb}}) = \text{SE}(\hat{t}_{\text{unb}})/M_0$.

[1] Working with the additional level of abstraction will allow us to see the structure of the variance more clearly, without floundering in the notation of the special case of equal probabilities discussed in this chapter. If you prefer to see the proof before you use the variance results, read Section 6.6 now.