# Chapter 1

# Introduction

**1.1** Target population: Unclear, but presumed to be readers of *Parade* magazine.

Sampling frame: Persons who know about the telephone survey.

Sampling unit = observation unit: One call. (Although it would also be correct to consider the sampling unit to be a person. The survey is so badly done that it is difficult to tell what the units are.)

As noted in Section 1.3, samples that consist only of volunteers are suspect. This is especially true of surveys in which respondents must pay to participate, as here—persons willing to pay 75 cents a call are likely to have strong opinions about the legalization of marijuana, and it is impossible to say whether pro- or anti-legalization adherents are more likely to call. This survey is utterly worthless for measuring public opinion because of its call-in format. Other potential biases, such as requiring a touch-tone telephone, or the sensitive subject matter or the ambiguity of the wording (what does "as legal as alcoholic beverages" mean?) probably make little difference because the call-in structure destroys all credibility for the survey by itself.

**1.2** Target population: All mutual funds.

Sampling frame: Mutual funds listed in newspaper.

Sampling unit = observation unit: One listing.

As funds are listed alphabetically by company, there is no reason to believe there will be any selection bias from the sampling frame. There may be undercoverage, however, if smaller or new funds are not listed in the newspaper.

**1.3** Target population: Not specified, but a target population of interest would be persons who have read the book.

Sampling frame: Persons who visit the website

Sampling unit = observation unit: One review.

The reviews are contributed by volunteers. They cannot be taken as representative of readers' opinions. Indeed, there have been instances where authors of competing books have written negative reviews of a book, although amazon.com tries to curb such practices.

**1.4** Target population: Persons eligible for jury duty in Maricopa County.

Sampling frame: County residents who are registered voters or licensed drivers over 18.

Sampling unit = observation unit: One resident.

Selection bias occurs largely because of undercoverage and nonresponse. Eligible jurors may not appear in the sampling frame because they are not registered to vote and they do not possess an Arizona driver's license. Addresses on either list may not be up to date. In addition, jurors fail to appear or are excused; this is nonresponse.

A similar question for class discussion is whether there was selection bias in selecting which young men in the U.Swere to be drafted and sent to Vietnam.

**1.5** Target population: All homeless persons in study area.

Sampling frame: Clinics participating in the Health Care for the Homeless project.

Sampling unit: Unclear. Depending on assumptions made about the survey design, one could say either a clinic or a homeless person is the sampling unit.

Observation unit: Person.

Selection bias may be a serious problem for this survey. Even though the demographics for HCH patients are claimed to match those of the homeless population (but do we *know* they match?) and the clinics are readily accessible, the patients differ in two critical ways from non-patients: (1) they needed medical treatment, and (2) they went to a clinic to get medical treatment. One does not know the likely direction of selection bias, but there is no reason to believe that the same percentages of patients and non-patients are mentally ill.

**1.6** Target population: Female readers of *Prevention* magazine.

Sampling frame: Women who see the survey in a copy of the magazine.

Sampling unit = observation unit: One woman.

This is a mail-in survey of volunteers, and we cannot trust any statistics from it.

**1.7** Target population: All cows in region.

Sampling frame: List of all farms in region.

Sampling unit: One farm.

Observation unit: One cow.

There is no reason to anticipate selection bias in this survey. The design is a single-

stage cluster sample, discussed in Chapter 5.

**1.8** Target population: Licensed boarding homes for the elderly in Washington state.

Sampling frame: List of 184 licensed homes.

Sampling unit = observation unit: One home.

Nonresponse is the obvious problem here, with only 43 of 184 administrators or food service managers responding. It may be that the respondents are the larger homes, or that their menus have better nutrition. The problem with nonresponse, though, is that we can only conjecture the direction of the nonresponse bias.

**1.13** Target population: All attendees of the 2005 JSM.

Sampling population: E-mail addresses provided by the attendees of the 2005 JSM.

Sampling unit: One e-mail address.

It is stated that the small sample of conference registrants was selected randomly. This is good, since the ASA can control the quality better and follow up on non-respondents. It also means, since the sample is selected, that persons with strong opinions cannot flood the survey. But nonresponse is a potential problem—response is not mandatory and it might be feared that only attendees with strong opinions or a strong sense of loyalty to the ASA will respond to the survey.

**1.14** Target population: All professors of education

Sampling population: List of education professors

Sampling unit: One professor

Information about how the sample was selected was not given in the publication, but let's assume it was a random sample. Obviously, nonresponse is a huge problem with this survey. Of the 5324 professors selected to be in the sample, only 900 were interviewed. Professors who travel during summer could of course not be contacted; also, summer is the worst time of year to try to interview professors for a survey.

**1.15** Target population: All adults

Sampling population: Friends and relatives of American Cancer Society volunteers

Sampling unit: One person

Here's what I wrote about the survey elsewhere:

"Although the sample contained Americans of diverse ages and backgrounds, and the sample may have provided valuable information for exploring factors associated with development of cancer, its validity for investigating the relationship between amount of sleep and mortality is questionable. The questions about amount of sleep and insomnia were not the focus of the original study, and the survey was not designed to obtain accurate responses to those questions. The design did not allow

researchers to assess whether the sample was representative of the target population of all Americans. Because of the shortcomings in the survey design, it is impossible to know whether the conclusions in Kripke et al. (2002) about sleep and mortality are valid or not." (pp. 97–98)

Lohr, S. (2008). "Coverage and sampling," chapter 6 of *International Handbook of Survey Methodology*, ed. E. deLeeuw, J. Hox, D. Dillman. New York: Erlbaum, 97–112.

**1.25** Students will have many different opinions on this issue. Of historical interest is this excerpt of a letter written by James Madison to Thomas Jefferson on February 14, 1790:

> A Bill for taking a census has passed the House of Representatives, and is with the Senate. It contained a schedule for ascertaining the component classes of the Society, a kind of information extremely requisite to the Legislator, and much wanted for the science of Political Economy. A repetition of it every ten years would hereafter afford a most curious and instructive assemblage of facts. It was thrown out by the Senate as a waste of trouble and supplying materials for idle people to make a book. Judge by this little experiment of the reception likely to be given to so great an idea as that explained in your letter of September.

# Chapter 2

# Simple Probability Samples

**2.1** (a) $\bar{y}_U = \dfrac{98 + 102 + 154 + 133 + 190 + 175}{6} = 142$

(b) For each plan, we first find the sampling distribution of $\bar{y}$.

**Plan 1:**

| Sample number | $P(\mathcal{S})$ | $\bar{y}_{\mathcal{S}}$ |
|---|---|---|
| 1 | 1/8 | 147.33 |
| 2 | 1/8 | 142.33 |
| 3 | 1/8 | 140.33 |
| 4 | 1/8 | 135.33 |
| 5 | 1/8 | 148.67 |
| 6 | 1/8 | 143.67 |
| 7 | 1/8 | 141.67 |
| 8 | 1/8 | 136.67 |

(i) $E[\bar{y}] = \dfrac{1}{8}(147.33) + \dfrac{1}{8}(142.33) + \cdots + \dfrac{1}{8}(136.67) = 142.$

(ii) $V[\bar{y}] = \dfrac{1}{8}(147.33 - 142)^2 + \dfrac{1}{8}(142.33 - 142)^2 + \cdots + \dfrac{1}{8}(136.67 - 142)^2 = 18.94.$

(iii) Bias $[\bar{y}] = E[\bar{y}] - \bar{y}_U = 142 - 142 = 0.$

(iv) Since Bias $[\bar{y}] = 0$, MSE $[\bar{y}] = V[\bar{y}] = 18.94$

**Plan 2:**

| Sample number | $P(\mathcal{S})$ | $\bar{y}_{\mathcal{S}}$ |
|---|---|---|
| 1 | 1/4 | 135.33 |
| 2 | 1/2 | 143.67 |
| 3 | 1/4 | 147.33 |

(i) $E[\bar{y}] = \dfrac{1}{4}(135.33) + \dfrac{1}{2}(143.67) + \dfrac{1}{4}(147.33) = 142.5.$

(ii)

$$\begin{aligned} V[\bar{y}] &= \frac{1}{4}(135.33 - 142.5)^2 + \frac{1}{2}(143.67 - 142.5)^2 + \frac{1}{4}(147.33 - 142.5)^2 \\ &= 12.84 + 0.68 + 5.84 \\ &= 19.36. \end{aligned}$$

(iii) Bias $[\bar{y}] = E[\bar{y}] - \bar{y}_U = 142.5 - 142 = 0.5$.

(iv) MSE $[\bar{y}] = V[\bar{y}] + (\text{Bias}\,[\bar{y}])^2 = 19.61$.

(c) Clearly, Plan 1 is better. It has smaller variance and is unbiased as well.

**2.2** (a) Unit 1 appears in samples 1 and 3, so $\pi_1 = P(\mathcal{S}_1) + P(\mathcal{S}_3) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$.

Similarly,

$$\begin{aligned} \pi_2 &= \frac{1}{4} + \frac{3}{8} = \frac{5}{8} \\ \pi_3 &= \frac{1}{8} + \frac{1}{4} = \frac{3}{8} \\ \pi_4 &= \frac{1}{8} + \frac{3}{8} + \frac{1}{8} = \frac{5}{8} \\ \pi_5 &= \frac{1}{8} + \frac{1}{8} = \frac{1}{4} \\ \pi_6 &= \frac{1}{8} + \frac{1}{8} + \frac{3}{8} = \frac{5}{8} \\ \pi_7 &= \frac{1}{4} + \frac{1}{8} = \frac{3}{8} \\ \pi_8 &= \frac{1}{4} + \frac{1}{8} + \frac{3}{8} + \frac{1}{8} = \frac{7}{8}. \end{aligned}$$

Note that $\sum_{i=1}^{8} \pi_i = 4 = n$.

(b)

| Sample, $\mathcal{S}$ | $P(\mathcal{S})$ | $\hat{t}$ |
|---|---|---|
| $\{1,3,5,6\}$ | 1/8 | 38 |
| $\{2,3,7,8\}$ | 1/4 | 42 |
| $\{1,4,6,8\}$ | 1/8 | 40 |
| $\{2,4,6,8\}$ | 3/8 | 42 |
| $\{4,5,7,8\}$ | 1/8 | 52 |

Thus the sampling distribution of $\hat{t}$ is:

| $k$ | $P(\hat{t} = k)$ |
|---|---|
| 38 | 1/8 |
| 40 | 1/8 |
| 42 | 5/8 |
| 52 | 1/8 |

**2.3** No, because thick books have a higher inclusion probability than thin books.

**2.4** (a) A total of $\binom{8}{3} = 56$ samples are possible, each with probability of selection $\frac{1}{56}$. The R function *samplist* below will (inefficiently!) generate each of the 56 samples. To find the sampling distribution of $\bar{y}$, I used the commands

```
samplist <- function(popn,sampsize){
    popvals <- 1:length(popn)
    temp <- comblist(popvals,sampsize)
    matrix(popn[t(temp)],nrow=nrow(temp),byrow=T)
}

comblist <- function(popvals, sampsize)
{
popsize <- length(popvals)
if(sampsize > popsize)
stop("sample size cannot exceed population size")
nvals <- popsize - sampsize + 1
nrows <- prod((popsize - sampsize + 1):popsize)/prod(1:sampsize)
ncols <- sampsize
yy <- matrix(nrow = nrows, ncol = ncols)
if(sampsize == 1) {yy <- popvals}
else {
nvals <- popsize - sampsize + 1
nrows <- prod(nvals:popsize)/prod(1:sampsize)
ncols <- sampsize
yy <- matrix(nrow = nrows, ncol = ncols)
rep1 <- rep(1, nvals)
if(nvals > 1) {
for(i in 2:nvals)
rep1[i] <- (rep1[i - 1] * (sampsize + i - 2))/(i - 1)
}
rep1 <- rev(rep1)
yy[, 1] <- rep(popvals[1:nvals], rep1)
for(i in 1:nvals) {
yy[yy[, 1] == popvals[i], 2:ncols] <- Recall(
popvals[(i + 1):popsize], sampsize - 1)
}
}
yy
}
temp1 <-samplist(c(1,2,4,4,7,7,7,8),3)
temp2 <-apply(temp1, 1, mean)
table(temp 2)
```

The following, then, is the sampling distribution of $\bar{y}$.

| $k$ | $P(\bar{y} = k)$ |
|---|---|
| $2\frac{1}{3}$ | 2/56 |
| 3 | 1/56 |
| $3\frac{1}{3}$ | 4/56 |
| $3\frac{2}{3}$ | 1/56 |
| 4 | 6/56 |
| $4\frac{1}{3}$ | 8/56 |
| $4\frac{2}{3}$ | 2/56 |
| 5 | 6/56 |
| $5\frac{1}{3}$ | 7/56 |
| $5\frac{2}{3}$ | 3/56 |
| 6 | 6/56 |
| $6\frac{1}{3}$ | 6/56 |
| 7 | 1/56 |
| $7\frac{1}{3}$ | 3/56 |

Using the sampling distribution,

$$E[\bar{y}] = \frac{2}{56}\left(2\frac{1}{3}\right) + \cdots + \frac{3}{56}\left(7\frac{1}{3}\right) = 5.$$

The variance of $\bar{y}$ for an SRS without replacement of size 3 is

$$V[\bar{y}] = \frac{2}{56}\left(2\frac{1}{3} - 5\right)^2 + \cdots + \frac{3}{56}\left(7\frac{1}{3} - 5\right)^2 = 1.429.$$

Of course, this variance could have been more easily calculated using the formula in (2.7):

$$V[\bar{y}] = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} = \left(1 - \frac{3}{8}\right)\frac{6.8571429}{3} = 1.429.$$

(b) A total of $8^3 = 512$ samples are possible when sampling with replacement. Fortunately, we need not list all of these to find the sampling distribution of $\bar{y}$. Let $X_i$ be the value of the $i$th unit drawn. Since sampling is done with replacement, $X_1, X_2$, and $X_3$ are independent; $X_i$ $(i = 1, 2, 3)$ has distribution

| $k$ | $P(X_i = k)$ |
|---|---|
| 1 | 1/8 |
| 2 | 1/8 |
| 4 | 2/8 |
| 7 | 3/8 |
| 8 | 1/8 |

Using the independence, then, we have the following probability distribution for $\bar{X}$, which serves as the sampling distribution of $\bar{y}$.

| $k$ | $P(\bar{y}=k)$ | $k$ | $P(\bar{y}=k)$ |
|---|---|---|---|
| 1 | 1/512 | $4\frac{2}{3}$ | 12/512 |
| $1\frac{1}{3}$ | 3/512 | 5 | 63/512 |
| $1\frac{2}{3}$ | 3/512 | $5\frac{1}{3}$ | 57/512 |
| 2 | 7/512 | $5\frac{2}{3}$ | 21/512 |
| $2\frac{1}{3}$ | 12/512 | 6 | 57/512 |
| $2\frac{2}{3}$ | 6/512 | $6\frac{1}{3}$ | 36/512 |
| 3 | 21/512 | $6\frac{2}{3}$ | 6/512 |
| $3\frac{1}{3}$ | 33/512 | 7 | 27/512 |
| $3\frac{2}{3}$ | 15/512 | $7\frac{1}{3}$ | 27/512 |
| 4 | 47/512 | $7\frac{2}{3}$ | 9/512 |
| $4\frac{1}{3}$ | 48/512 | 8 | 1/512 |

The with-replacement variance of $\bar{y}$ is

$$V_{\text{wr}}[\bar{y}] = \frac{1}{512}(1-5)^2 + \cdots + \frac{1}{512}(8-5)^2 = 2.$$

Or, using the formula with population variance (see Exercise 2.28),

$$V_{\text{wr}}[\bar{y}] = \frac{1}{n}\sum_{i=1}^{N}\frac{(y_i - \bar{y}_U)^2}{N} = \frac{6}{3} = 2.$$

**2.5** (a) The sampling weight is $100/30 = 3.3333$.

(b) $\hat{t} = \sum_{i\in\mathcal{S}} w_i y_i = 823.33$.

(c) $\hat{V}(\hat{t}) = N^2\left(1-\frac{n}{N}\right)\frac{s_y^2}{n} = 100^2\left(1-\frac{30}{100}\right)\frac{15.9781609}{30} = 3728.238$, so

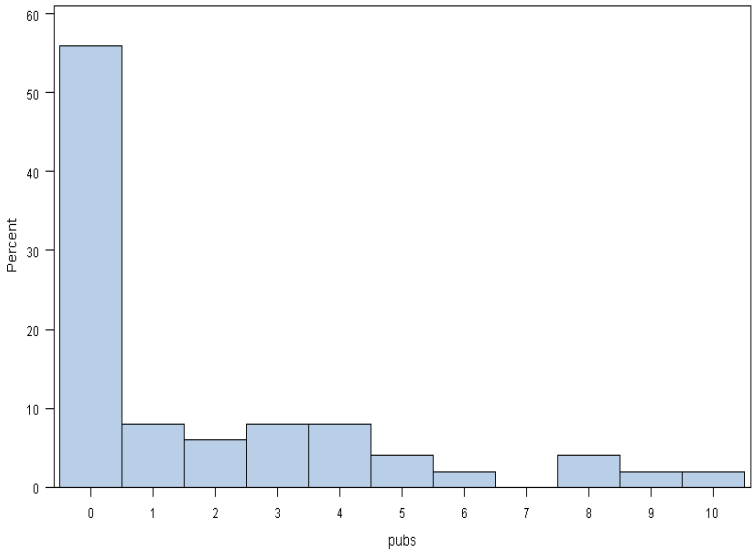
$$\text{SE}\,(\hat{t}) = \sqrt{3728.238} = 61.0593$$

and a 95% CI for $t$ is

$$823.33 \pm (2.045230)(61.0593) = 823.33 \pm 124.8803 = [698.45, 948.21].$$

The fpc is $(1 - 30/100) = .7$, so it reduces the width of the CI.

**2.6** (a)




The data are quite skewed because 28 faculty have no publications.

(b) $\bar{y} = 1.78$; $s = 2.682$;

$$\text{SE}\,[\bar{y}] = \frac{2.682}{\sqrt{50}}\sqrt{1 - \frac{50}{807}} = 0.367.$$

(c) No; a sample of size 50 is probably not large enough for $\bar{y}$ to be normally distributed, because of the skewness of the original data.

The sample skewness of the data is (from SAS) 1.593. This can be calculated by hand, finding

$$\frac{1}{n}\sum_{i\in\mathcal{S}}(y_i - \bar{y})^3 = 28.9247040$$

so that the skewness is $28.9247040/(2.682^3) = 1.499314$. Note this estimate differs from SAS PROC UNIVARIATE since SAS adjusts for df using the formula skewness $= \dfrac{n}{(n-1)(n-2)}\displaystyle\sum_{i\in\mathcal{S}}(y_i - \bar{y})^3/s^3$. Whichever estimate is used, however, formula (2.23) says we need a minimum of

$$28 + 25(1.5)^2 = 84$$

observations to use the central limit theorem.

(d) $\hat{p} = 28/50 = 0.56$.

$$\text{SE}\,(\hat{p}) = \sqrt{\frac{(0.56)(0.44)}{49}\left(1 - \frac{50}{807}\right)} = 0.0687.$$

A 95% confidence interval is

$$0.56 \pm 1.96(0.0687) = [0.425, 0.695].$$

**2.07** (a) A 95% confidence interval for the proportion of entries from the South is

$$\frac{175}{1000} \pm 1.96\sqrt{\frac{\frac{175}{1000}\left(1-\frac{175}{1000}\right)}{1000}} = [.151, .199].$$

(b) As 0.309 is not in the confidence interval, there is evidence that the percentages differ.

**2.08** Answers will vary.

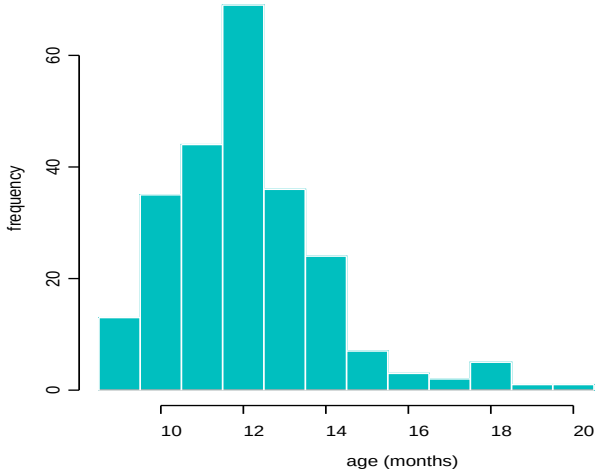
**2.09** If $n_0 \leq N$, then

$$\begin{aligned} z_{\alpha/2}\sqrt{1-\frac{n}{N}}\frac{S}{\sqrt{n}} &= z_{\alpha/2}\sqrt{1-\frac{n_0}{N(1+\frac{n_0}{N})}}\frac{S}{\sqrt{n_0}}\sqrt{1+\frac{n_0}{N}} \\ &= z_{\alpha/2}\sqrt{1+\frac{n_0}{N}-\frac{n_0}{N}}\frac{S}{\sqrt{n_0}} \\ &= z_{\alpha/2}\frac{S}{\frac{z_{\alpha/2}S}{e}} \\ &= e \end{aligned}$$

**2.10** Design 3 gives the most precision because its sample size is largest, even though it is a small fraction of the population. Here are the variances of $\bar{y}$ for the three samples:

| Sample Number | $V(\bar{y})$ |
|---|---|
| 1 | $(1-400/4000)S^2/400 = 0.00225S^2$ |
| 2 | $(1-30/300)S^2/30 = 0.03S^2$ |
| 3 | $(1-3000/300{,}000{,}000)S^2/3000 = 0.00033333S^2$ |

**2.11** (a)




The histogram appears skewed with tail on the right. With a mildly skewed distribution, though, a sample of size 240 is large enough that the sample mean should be normally distributed.

(b) $\bar{y} = 12.07917$; $s^2 = 3.705003$; $\text{SE}\,[\bar{y}] = \sqrt{s^2/n} = 0.12425$.

(Since we do not know the population size, we ignore the fpc, at the risk of a slightly-too-large standard error.)

A 95% confidence interval is

$$12.08 \pm 1.96(0.12425) = [11.84, 12.32].$$

(c) $n = \dfrac{(1.96)^2(3.705)}{(0.5)^2} = 57.$

**2.12** (a) Using (2.17) and choosing the maximum possible value of $(0.5)^2$ for $S^2$,

$$n_0 = \frac{(1.96)^2 S^2}{e^2} = \frac{(1.96)^2(0.5)^2}{(0.1)^2} = 96.04.$$

Then

$$n = \frac{n_0}{1 + n_0/N} = \frac{96.04}{1 + 96.04/580} = 82.4.$$

(b) Since sampling is with replacement, no fpc is used. An approximate 95% confidence interval for the proportion of children not overdue for vaccination is

$$\frac{27}{120} \pm 1.96\sqrt{\frac{\frac{27}{120}\left(1 - \frac{27}{120}\right)}{120}} = [0.15, 0.30]$$

**2.13** (a) We have $\hat{p} = .2$ and

$$\hat{V}(\hat{p}) = \left(1 - \frac{745}{2700}\right) \frac{(.2)(.8)}{744} = 0.0001557149,$$

so an approximate 95% CI is

$$0.2 \pm 1.96\sqrt{0.0001557149} = [.176, .224].$$

(b) The above analysis is valid only if the respondents are a random sample of the selected sample. If respondents differ from the nonrespondents—for example, if the nonrespondents are more likely to have been bullied—then the entire CI may be biased.

**2.14** Here is SAS output:

```
The SURVEYMEANS Procedure

            Data Summary

Number of Observations           150
Sum of Weights                   864

   Class Level Information

Class
Variable      Levels    Values

sex                2    f m


                          Statistics

                                 Std Error
Variable   Level        Mean      of Mean       95% CL for Mean
-----------------------------------------------------------------
sex        f        0.306667     0.034353    0.23878522 0.37454811
           m        0.693333     0.034353    0.62545189 0.76121478

                          Statistics

Variable   Level         Sum      Std Dev        95% CL for Sum
-----------------------------------------------------------------
sex        f      264.960000    29.680756    206.310434 323.609566
           m      599.040000    29.680756    540.390434 657.689566
-----------------------------------------------------------------
```

**2.15** (a) $\bar{y}$ =301,953.7, $s^2$ =118,907,450,529.

$$\text{CI} : 301953.7 \pm 1.96\sqrt{\frac{s^2}{300}\left(1 - \frac{300}{3078}\right)}, \text{ or } [264883, 339025]$$

(b) $\bar{y} = 599.06$, $s^2 = 161795.4$

CI : $[556, 642]$

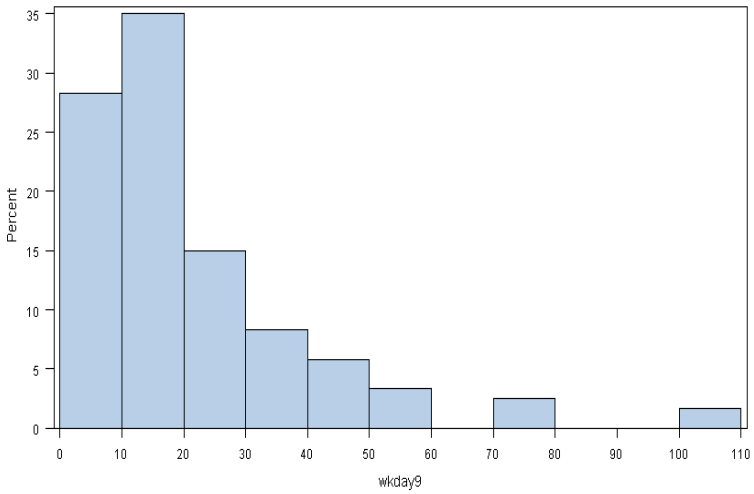
(c) $\bar{y} = 56.593$, $s^2 = 5292.73$

CI : $[48.8, 64.4]$

(d) $\bar{y} = 46.823$, $s^2 = 4398.199$

CI : $[39.7, 54.0]$

**2.16** (a) The data appear skewed with tail on right.



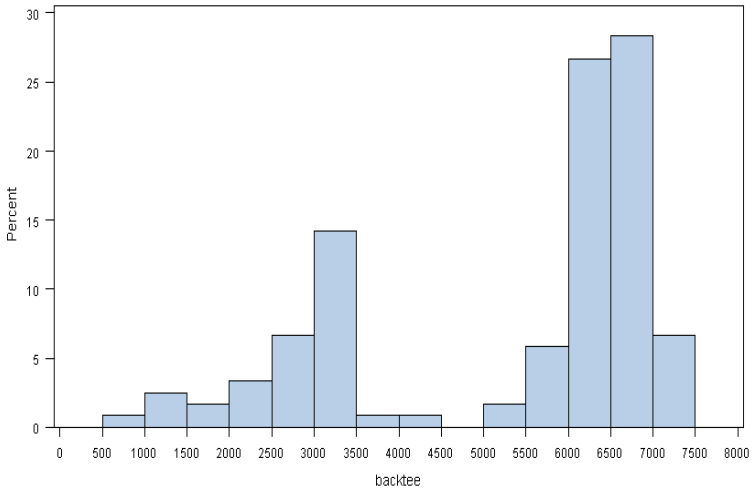

(b) $\bar{y} = 5309.8$, $s^2 = 3{,}274{,}784$, SE $[\bar{y}] = 164.5$

Here is SAS code for problems 2.16 and 2.17:

```
filename golfsrs 'C:\golfsrs.csv';
options ls=78 nodate nocenter;

data golfsrs;
   infile golfsrs delimiter="," dsd firstobs=2;
     /* The dsd option allows SAS to read the missing values between
        successive delimiters */
   sampwt = 14938/120;
```

```
   input  RN  state $ holes  type $ yearblt  wkday18  wkday9  wkend18
      wkend9  backtee  rating  par  cart18  cart9  caddy $ pro  $ ;

/* Make sure the data were read in correctly */
proc print data=golfsrs;
run;

proc univariate data= golfsrs;
   var wkday9 backtee;
   histogram wkday9 /endpoints = 0 to 110 by 10;
   histogram backtee /endpoints = 0 to 8000 by 500;
run;

proc surveymeans data=golfsrs total = 14938;
   weight sampwt;
   var wkday9 backtee;
run;
```

**2.17** (a) The data appear skewed with tail on left.




(b) $\bar{y} = 5309.8$, $s^2 = 3{,}274{,}784$, $\text{SE}[\bar{y}] = 164.5$

**2.18** $\hat{p} = 85/120 = 0.708$

$$95\%\text{CI: } 85/120 \pm 1.96\sqrt{\frac{85/120\,(1-85/120)}{119}\left(1-\frac{120}{14938}\right)} = .708 \pm .081,$$

or [0.627, 0.790].

**2.19** Assume the maximum value for the variance, with $p = 0.5$. Then use $n_0 = 1.96^2(0.5)^2/(.04)^2$, $n = n_0/(1 + n_0/N)$.

| City | $n_0$ | $n$ |
|---|---|---|
| Buckeye | 600.25 | 535 |
| Gilbert | 600.25 | 595 |
| Gila Bend | 600.25 | 446 |
| Phoenix | 600.25 | 600 |
| Tempe | 600.25 | 598 |

The finite population correction only makes a difference for Buckeye and Gila Bend.

**2.20** Sixty of the 70 samples yield confidence intervals, using this procedure, that include the true value $t = 40$. The exact confidence level is $60/70 = 0.857$.

**2.21** (a) A number of different arguments can be made that this method results in a simple random sample. Here is one proof, which assumes that the random number table indeed consists of independent random numbers. In the context of the problem, $M = 999$, $N = 742$, and $n = 30$. Of course, many students will give a more heuristic argument.

Let $U_1, U_2, U_3, \ldots$, be independent random variables, each with a discrete uniform distribution on $\{0, 1, 2, \ldots, M\}$. Now define

$$T_1 = \min\{i : U_i \in [1, N]\}$$

and

$$T_k = \min\{i > T_{k-1} : U_i \in [1, N], U_i \notin \{U_{T_1}, \ldots, U_{T_{k-1}}\}\}$$

for $k = 2, \ldots, n$. Then for $\{x_1, \ldots, x_n\}$ a set of $n$ distinct elements in $\{1, \ldots, N\}$,

$$P(\mathcal{S} = \{x_1, \ldots, x_n\}) = P(\{U_{T_1}, \ldots, U_{T_n}\} = \{x_1, \ldots, x_n\})$$

$$\begin{aligned} P\{U_{T_1} = x_1, \ldots, U_{T_n} = x_n\} &= E[P\{U_{T_1} = x_1, \ldots, U_{T_n} = x_n \mid T_1, T_2, \ldots, T_n\}] \\ &= \left(\frac{1}{N}\right)\left(\frac{1}{N-1}\right)\left(\frac{1}{N-2}\right)\cdots\left(\frac{1}{N-n+1}\right) \\ &= \frac{(N-n)!}{N!}. \end{aligned}$$

Conditional on the stopping times $T_1, \ldots, T_n$, $U_{T_1}$ is discrete uniform on $\{1, \ldots, N\}$; $(U_{T_2} \mid T_1, \ldots, T_N, U_{T_1})$ is discrete uniform on $\{1, \ldots, N\} - \{U_{T_1}\}$, and so on. Since $x_1, \ldots, x_n$ are arbitrary,

$$P(\mathcal{S} = \{x_1, \ldots, x_n\}) = \frac{n!(N-n)!}{N!} = \frac{1}{\binom{N}{n}},$$

so the procedure results in a simple random sample.

(b) This procedure does not result in a simple random sample. Units starting with 5, 6, or 7 are more likely to be in the sample than units starting with 0 or 1. To see

this, let's look at a simpler case: selecting one number between 1 and 74 using this procedure.

Let $U_1, U_2, \ldots$ be independent random variables, each with a discrete uniform distribution on $\{0, \ldots, 9\}$. Then the first random number considered in the sequence is $10U_1 + U_2$; if that number is not between 1 and 74, then $10U_2 + U_3$ is considered, etc. Let

$$T = \min\{i : 10U_i + U_{i+1} \in [1, 74]\}.$$

Then for $x = 10x_1 + x_2$, $x \in [1, 74]$,

$$\begin{aligned} P(\mathcal{S} = \{x\}) &= P(10U_T + U_{T+1} = x) \\ &= P(U_T = x_1, U_{T+1} = x_2). \end{aligned}$$

For part (a), the stopping times were irrelevant for the distribution of $U_{T_1}, \ldots, U_{T_n}$; here, though, the stopping time makes a difference. One way to have $T = 2$ is if $10U_1 + U_2 = 75$. In that case, you have rejected the first number solely because the second digit is too large, but that second digit becomes the first digit of the random number selected. To see this formally, note that

$$\begin{aligned} P(\mathcal{S} = \{x\}) = P(&10U_1 + U_2 = x \text{ or } \{10U_1 + U_2 \notin [1, 74] \text{ and } 10U_2 + U_3 = x\} \\ &\text{or} \quad \{10U_1 + U_2 \notin [1, 74] \quad \text{and} \quad 10U_2 + U_3 \notin [1, 74] \\ &\text{and} \quad 10U_3 + U_4 = x\} \quad \text{or} \ldots) \\ = P(&U_1 = x_1, U_2 = x_2) \\ &+ \sum_{t=2}^{\infty} P\Bigg( \bigcap_{i=1}^{t-1} \{U_i > 7 \quad \text{or} \quad [U_i = 7 \text{ and } U_{i+1} > 4]\} \\ &\qquad \text{and } U_t = x_1 \text{ and } U_{t+1} = x_2 \Bigg). \end{aligned}$$

Every term in the series is larger if $x_1 > 4$ than if $x_1 \leq 4$.

(c) This method almost works, but not quite. For the first draw, the probability that 131 (or any number in $\{1, \ldots, 149, 170\}$ is selected is 6/1000; the probability that 154 (or any number in $\{150, \ldots, 169\}$) is selected is 5/1000.

(d) This clearly does not produce an SRS, because no odd numbers can be included.

(e) If class sizes are unequal, this procedure does not result in an SRS: students in smaller classes are more likely to be selected for the sample than are students in larger classes.

Consider the probability that student $j$ in class $i$ is chosen on the first draw.

$$\begin{aligned} P\{\text{select student } j \text{ in class } i\} &= P\{\text{select class } i\} P\{\text{select student } j \mid \text{class } i\} \\ &= \frac{1}{20} \, \frac{1}{\text{number of students in class } i}. \end{aligned}$$

(f) Let's look at the probability student $j$ in class $i$ is chosen for first unit in the sample. Let $U_1, U_2, \ldots$ be independent discrete uniform $\{1, \ldots, 20\}$ and let $V_1, V_2, \ldots$

be independent discrete uniform $\{1,\ldots,40\}$. Let $M_i$ denote the number of students in class $i$, with $K=\sum_{i=1}^{20} M_i$. Then, because all random variables are independent,

$$\begin{aligned}
&P(\text{student } j \text{ in class } i \text{ selected})\\
&= P(U_1=i,V_2=j)+P(U_2=i,V_2=j)P\Big(\bigcup_{k=1}^{20}\{U_1=k,V_1>M_k\}\Big)\\
&\quad +\cdots+P\Big\{U_{l+1}=i,V_{l+1}=j\Big\}\prod_{q=1}^{l}P\Big(\bigcup_{k=1}^{20}\{U_q=k,V_q>M_k\}\Big)\\
&\quad +\cdots\\
&= \frac{1}{20}\,\frac{1}{40}\sum_{l=0}^{\infty}\Big[\prod_{q=1}^{l}P\Big(\bigcup_{k=1}^{20}\{U_q=k,V_q>M_k\}\Big)\Big]\\
&= \frac{1}{800}\sum_{l=0}^{\infty}\Big[\sum_{k=1}^{20}\frac{1}{20}\frac{40-M_k}{40}\Big]^l\\
&= \frac{1}{800}\sum_{l=0}^{\infty}\Big[1-\frac{K}{800}\Big]^l\\
&= \frac{1}{800}\frac{1}{1-(1-K/800)}=\frac{1}{K}.
\end{aligned}$$

Thus, before duplicates are eliminated, a student has probability $1/K$ of being selected on any given draw. The argument in part (a) may then be used to show that when duplicates are discarded, the resulting sample is an SRS.

**2.22** (a) From (2.13),

$$\mathrm{CV}(\bar{y})=\frac{\sqrt{V(\bar{y})}}{E(\bar{y})}=\sqrt{1-\frac{n}{N}}\frac{S}{\sqrt{n}\bar{y}_U}.$$

Substituting $\hat{p}$ for $\bar{y}$, and $\frac{N}{N-1}p(1-p)$ for $S^2$, we have

$$\mathrm{CV}(\hat{p})==\sqrt{\Big(1-\frac{n}{N}\Big)\frac{Np(1-p)}{(N-1)np^2}}=\sqrt{\frac{N-n}{N-1}\frac{1-p}{np}}.$$

The CV for a sample of size 1 is $\sqrt{(1-p)/p}$. The sample size in (2.26) will be $z^2_{\alpha/2}\mathrm{CV}^2/r^2$.

(b) I used Excel to calculate these values.

| $p$ | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 | 0.3 | 0.5 |
|---|---|---|---|---|---|---|---|
| Fixed | 4.3 | 21.2 | 42.3 | 202.8 | 384.2 | 896.4 | 1067.1 |
| Relative | 4264176 | 849420 | 422576 | 81100 | 38416 | 9959.7 | 4268.4 |

| $p$ | 0.7 | 0.9 | 0.95 | 0.99 | 0.995 | 0.999 |
|---|---|---|---|---|---|---|
| Fixed | 896.4 | 384.2 | 202.8 | 42.3 | 21.2 | 4.3 |
| Relative | 1829.3 | 474.3 | 224.7 | 43.1 | 21.4 | 4.3 |

**2.23**

$$\begin{aligned} P(\text{no missing data}) &= \frac{\binom{3059}{300}\binom{19}{0}}{\binom{3078}{300}} \\ &= \frac{(2778)(2777)\ldots(2760)}{(3078)(3077)\ldots(3060)} \\ &= 0.1416421. \end{aligned}$$

**2.24**

$$g(n) = L(n) + C(n) = k\left(1 - \frac{n}{N}\right)\frac{S^2}{n} + c_0 + c_1 n.$$

$$\frac{dg}{dn} = -\frac{kS^2}{n^2} + c_1$$

Setting the derivative equal to 0 and solving for $n$ gives

$$n = \sqrt{\frac{kS^2}{c_1}}.$$

The sample size, in the decision theoretic approach, should be larger if the cost of a bad estimate, $k$, or the variance, $S^2$, is larger; the sample size is smaller if the cost of sampling is larger.

**2.25** (a) Skewed, with tail on right.

(b) $\bar{y} = 20.15$, $s^2 = 321.357$, $\text{SE}[\bar{y}] = 1.63$

**2.26** In a systematic sample, the population is partitioned into $k$ clusters, each of size $n$. One of these clusters is selected with probability $1/k$, so $\pi_i = 1/k$ for each $i$. But many of the samples that could be selected in an SRS cannot be selected in a systematic sample. For example,

$$P(Z_1 = 1, \ldots, Z_n = 1) = 0 :$$

since every $k$th unit is selected, the sample cannot consist of the first $n$ units in the population.

**2.27** (a)

$$\begin{aligned} P \text{ (you are in sample)} &= \frac{\binom{99{,}999{,}999}{999}\binom{1}{1}}{\binom{100{,}000{,}000}{1000}} \\ &= \frac{99{,}999{,}999!}{999!\ 99{,}999{,}000!}\frac{1000!\ 99{,}999{,}000!}{100{,}000{,}000!} \\ &= \frac{1000}{100{,}000{,}000} = \frac{1}{100{,}000}. \end{aligned}$$

(b)

$$P \text{ (you are not in any of the 2000 samples)} = \left(1 - \frac{1}{100{,}000}\right)^{2000} = 0.9802$$

(c) $P$ (you are not in any of $x$ samples) $= (1 - 1/100{,}000)^x$. Solving for $x$ in $(1 - 1/100{,}000)^x = 0.5$ gives $x \log(.99999) = \log(0.5)$, or $x = 69314.4$. Almost 70,000 samples need to be taken! This problem provides an answer to the common question, "Why haven't I been sampled in a poll?"

**2.28** (a) We can think of drawing a simple random sample with replacement as performing an experiment $n$ independent times; on each trial, outcome $i$ (for $i \in \{1, \ldots, N\}$) occurs with probability $p_i = 1/N$. This describes a multinomial experiment.

We may then use properties of the multinomial distribution to answer parts (b) and (c):

$$E[Q_i] = np_i = \frac{n}{N},$$
$$V[Q_i] = np_i(1 - p_i) = \frac{n}{N}\left(1 - \frac{1}{N}\right),$$

and

$$\text{Cov}\,[Q_i, Q_j] = -np_ip_j = -\frac{n}{N}\frac{1}{N} \quad \text{for} \quad i \neq j.$$

(b)

$$E[\hat{t}] = \frac{N}{n} E\left[\sum_{i=1}^{N} Q_i y_i\right] = \frac{N}{n} \sum_{i=1}^{N} \frac{n}{N} y_i = t.$$

(c)

$$\begin{aligned}
V[\hat{t}] &= \left(\frac{N}{n}\right)^2 V\left[\sum_{i=1}^{N} Q_i y_i\right] \\
&= \left(\frac{N}{n}\right)^2 \sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \ \text{Cov}\,[Q_{ij} \ \ Q_j] \\
&= \left(\frac{N}{n}\right)^2 \left\{ \sum_{i=1}^{N} y_i^2 np_i(1 - p_i) + \sum_{i=1}^{N}\sum_{j \neq i} y_i y_j (-np_ip_j) \right\} \\
&= \left(\frac{N}{n}\right)^2 \left\{ \frac{n}{N}\left(1 - \frac{1}{N}\right) \sum_{i=1}^{N} y_i^2 - \frac{n}{N}\frac{1}{N} \sum_{i=1}^{N}\sum_{j \neq i} y_i y_j \right\} \\
&= \frac{N}{n} \left\{ \sum_{i=1}^{N} y_i^2 - N\bar{y}_U^2 \right\} \\
&= \frac{N^2}{n} \frac{\sum_{i=1}^{N} (y_i - \bar{y}_U)^2}{N}.
\end{aligned}$$

**2.29** We use induction. Clearly, $\mathcal{S}_0$ is an SRS of size $n$ from a population of size $n$.

Now suppose $\mathcal{S}_{k-1}$ is an SRS of size $n$ from $U_{k-1} = \{1, 2, \ldots, n+k-1\}$, where $k \geq 1$. We wish to show that $\mathcal{S}_k$ is an SRS of size $n$ from $U_k = \{1, 2, \ldots, n+k\}$. Since $\mathcal{S}_{k-1}$ is an SRS, we know that

$$P(\mathcal{S}_{k-1}) = \frac{1}{\left( \begin{array}{c} n+k-1 \\ n \end{array} \right)} = \frac{n!(k-1)!}{(n+k-1)!}.$$

Now let $U_k \sim \text{Uniform}(0,1)$, let $V_k$ be discrete uniform $(1, \ldots, n)$, and suppose $U_k$ and $V_k$ are independent. Let $\mathcal{A}$ be a subset of size $n$ from $U_k$. If $\mathcal{A}$ does not contain unit $n+k$, then $\mathcal{A}$ can be achieved as a sample at step $k-1$ and

$$\begin{aligned} P(\mathcal{S}_k = \mathcal{A}) &= P\left(\mathcal{S}_{k-1} \text{ and } U_k > \frac{n}{n+k}\right) \\ &= P(\mathcal{S}_{k-1}) \frac{k}{n+k} \\ &= \frac{n!k!}{(n+k)!}. \end{aligned}$$

If $\mathcal{A}$ does contain unit $n+k$, then the sample at step $k-1$ must contain $\mathcal{A}_{k-1} = \mathcal{A} - \{n+k\}$ plus one other unit among the $k$ units not in $\mathcal{A}_{k-1}$.

$$\begin{aligned} P(\mathcal{S}_k = \mathcal{A}) &= \sum_{j \in U_{k-1} \cap \mathcal{A}_{k-1}^C} P\left(\mathcal{S}_{k-1} = \mathcal{A}_{k-1} \cup \{j\} \text{ and } U_k \leq \frac{n}{n+k} \text{ and } V_k = j\right) \\ &= k \frac{n!(k-1)!}{(n+k-1)!} \frac{n}{n+k} \frac{1}{n} \\ &= \frac{n!k!}{(n+k)!}. \end{aligned}$$

**2.30** I always use this activity in my classes. Students generally get estimates of the total area that are biased upwards for the purposive sample. They think, when looking at the picture, that they don't have enough of the big rectangles and so tend to oversample them. This is also a good activity for reviewing confidence intervals and other concepts from an introductory statistics class.

# Chapter 3

# Stratified Sampling

**3.2** (a) For each stratum, we calculate $\hat{t}_h = 4\bar{y}_h$

**Stratum 1**

| Sample, $\mathcal{S}_1$ | $P(\mathcal{S}_1)$ | $\{y_i, i \in \mathcal{S}_1\}$ | $\hat{t}_{1\mathcal{S}_1}$ |
|---|---|---|---|
| $\{1,2\}$ | 1/6 | $1,2$ | 6 |
| $\{1,3\}$ | 1/6 | $1,4$ | 10 |
| $\{1,8\}$ | 1/6 | $1,8$ | 18 |
| $\{2,3\}$ | 1/6 | $2,4$ | 12 |
| $\{2,8\}$ | 1/6 | $2,8$ | 20 |
| $\{3,8\}$ | 1/6 | $4,8$ | 24 |

**Stratum 2**

| Sample, $\mathcal{S}_2$ | $P(\mathcal{S}_2)$ | $\{y_i, i \in \mathcal{S}_2\}$ | $\hat{t}_{2\mathcal{S}_2}$ |
|---|---|---|---|
| $\{4,5\}$ | 1/6 | $4,7$ | 22 |
| $\{4,6\}$ | 1/6 | $4,7$ | 22 |
| $\{4,7\}$ | 1/6 | $4,7$ | 22 |
| $\{5,6\}$ | 1/6 | $7,7$ | 28 |
| $\{5,7\}$ | 1/6 | $7,7$ | 28 |
| $\{6,7\}$ | 1/6 | $7,7$ | 28 |

(b) From Stratum 1, we have the following probability distribution for $\hat{t}_1$:

| $j$ | $P(\hat{t}_1 = j)$ |
|---|---|
| 6 | 1/6 |
| 10 | 1/6 |
| 12 | 1/6 |
| 18 | 1/6 |
| 20 | 1/6 |
| 24 | 1/6 |

The sampling distribution for $\hat{t}_2$ is:

| $k$ | $P(\hat{t}_2 = k)$ |
|---|---|
| 22 | 1/2 |
| 28 | 1/2 |

Because we sample *independently* in Strata 1 and 2,

$$P(\hat{t}_1 = j \text{ and } \hat{t}_2 = k) = P(\hat{t}_1 = j)P(\hat{t}_2 = k)$$

for all possible values of $j$ and $k$. Thus,

| $j$ | $k$ | $j+k$ | $P(\hat{t}_1 = j \text{ and } \hat{t}_2 = k)$ |
|---|---|---|---|
| 6 | 22 | 28 | 1/12 |
| 6 | 28 | 34 | 1/12 |
| 10 | 22 | 32 | 1/12 |
| 10 | 28 | 38 | 1/12 |
| 12 | 22 | 34 | 1/12 |
| 12 | 28 | 40 | 1/12 |
| 18 | 22 | 40 | 1/12 |
| 18 | 28 | 46 | 1/12 |
| 20 | 22 | 42 | 1/12 |
| 20 | 28 | 48 | 1/12 |
| 24 | 22 | 46 | 1/12 |
| 24 | 28 | 52 | 1/12 |

So the sampling distribution of $\hat{t}_{\text{str}}$ is

| $k$ | $P(\hat{t}_{\text{str}} = k)$ |
|---|---|
| 28 | 1/12 |
| 32 | 1/12 |
| 34 | 2/12 |
| 38 | 1/12 |
| 40 | 2/12 |
| 42 | 1/12 |
| 46 | 2/12 |
| 48 | 1/12 |
| 52 | 1/12 |

(c)

$$\begin{aligned} E[\hat{t}_{\text{str}}] &= \sum_k kP(\hat{t}_{\text{str}} = k) = 40 \\ V[\hat{t}_{\text{str}}] &= \sum_k (k-40)^2 P(\hat{t}_{\text{str}} = k) = 47\tfrac{1}{3}. \end{aligned}$$

**3.3** (a) $\bar{y}_U = 71.83333$, $S^2 = 86.16667$.

(b) $\begin{pmatrix} 6 \\ 4 \end{pmatrix} = 15$.

(c) The 15 samples are:

| Units in sample | | | | $y$ values in sample | | | | Sample Mean |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 66 | 59 | 70 | 83 | 69.50 |
| 1 | 2 | 3 | 5 | 66 | 59 | 70 | 82 | 69.25 |
| 1 | 2 | 3 | 6 | 66 | 59 | 70 | 71 | 66.50 |
| 1 | 2 | 4 | 5 | 66 | 59 | 83 | 82 | 72.50 |
| 1 | 2 | 4 | 6 | 66 | 59 | 83 | 71 | 69.75 |
| 1 | 2 | 5 | 6 | 66 | 59 | 82 | 71 | 69.50 |
| 1 | 3 | 4 | 5 | 66 | 70 | 83 | 82 | 75.25 |
| 1 | 3 | 4 | 6 | 66 | 70 | 83 | 71 | 72.50 |
| 1 | 3 | 5 | 6 | 66 | 70 | 82 | 71 | 72.25 |
| 1 | 4 | 5 | 6 | 66 | 83 | 82 | 71 | 75.50 |
| 2 | 3 | 4 | 5 | 59 | 70 | 83 | 82 | 73.50 |
| 2 | 3 | 4 | 6 | 59 | 70 | 83 | 71 | 70.75 |
| 2 | 3 | 5 | 6 | 59 | 70 | 82 | 71 | 70.50 |
| 2 | 4 | 5 | 6 | 59 | 83 | 82 | 71 | 73.75 |
| 3 | 4 | 5 | 6 | 70 | 83 | 82 | 71 | 76.50 |

Using (2.9),

$$V(\bar{y}) = \left(1 - \frac{4}{6}\right) \frac{86.16667}{4} = 7.180556.$$

(d) $\binom{3}{2}\binom{3}{2} = 9.$

(e) You cannot have any of the samples from (c) which contain 3 units from one of the strata. This eliminates the first 3 samples, which contain {1, 2, 3} and the three samples containing students {4, 5, 6}. The stratified samples are

| Units in $\mathcal{S}_1$ | | Units in $\mathcal{S}_2$ | | $y$ values in sample | | | | $\bar{y}_{\text{str}}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 4 | 5 | 66 | 59 | 83 | 82 | 72.50 |
| 1 | 2 | 4 | 6 | 66 | 59 | 83 | 71 | 69.75 |
| 1 | 2 | 5 | 6 | 66 | 59 | 82 | 71 | 69.50 |
| 1 | 3 | 4 | 5 | 66 | 70 | 83 | 82 | 75.25 |
| 1 | 3 | 4 | 6 | 66 | 70 | 83 | 71 | 72.50 |
| 1 | 3 | 5 | 6 | 66 | 70 | 82 | 71 | 72.25 |
| 2 | 3 | 4 | 5 | 59 | 70 | 83 | 82 | 73.50 |
| 2 | 3 | 4 | 6 | 59 | 70 | 83 | 71 | 70.75 |
| 2 | 3 | 5 | 6 | 59 | 70 | 82 | 71 | 70.50 |

$$V(\bar{y}_{\text{str}}) = \left(1 - \frac{2}{3}\right)\left(\frac{3}{6}\right)^2 \frac{31}{2} + \left(1 - \frac{2}{3}\right)\left(\frac{3}{6}\right)^2 \frac{44.33333}{2} = 3.14.$$

The variance is smaller because the extreme samples from (c) are excluded by the stratified design. The variances $S_1^2 = 31$ and $S_2^2 = 44.33$ are much smaller than the population variance $S^2$.

**3.4** Here is SAS code for creating the data set:

```
data acls;
   input stratum $ popsize returns percfem;
   females = round(returns*percfem/100);
   males = returns - females;
   sampwt = popsize/returns;
   datalines;
Literature 9100 636 38
Classics   1950 451 27
Philosophy 5500 481 18
History    10850 611 19
Linguistics 2100 493 36
PoliSci    5500 575 13
Sociology  9000 588 26
;

proc print data=acls;
run;

data aclslist;
   set acls;
   do i = 1 to females;
     femind = 1;
 output;
   end;
   do i = 1 to males;
     femind = 0;
 output;
   end;

/* Check whether we created the data set correctly*/
proc freq data=aclslist;
   tables stratum * femind;
run;

proc surveymeans data=aclslist mean clm sum clsum;
   stratum stratum;
   weight sampwt;
   var femind;
run;
```

We obtain $\hat{t} = 10858$ with SE 313. These values differ from those in Example 4.4 because of rounding.

**3.5** (a) The sampled population consists of members of the organizations who would respond to the survey.

(b)

$$\begin{aligned}\hat{p}_{\text{str}} &= \sum_{h=1}^{7} \frac{N_h}{N}\hat{p}_h \\ &= \left(\frac{9{,}100}{44{,}000}\right)(0.37) + \left(\frac{1{,}950}{44{,}000}\right)(0.23) + \cdots + \left(\frac{9{,}000}{44{,}000}\right)(0.41) \\ &= 0.334.\end{aligned}$$

$$\begin{aligned}\text{SE}\left[\hat{p}_{\text{str}}\right] &= \sqrt{\sum_{h=1}^{7}\left(1-\frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h-1}} \\ &= \sqrt{1.46\times10^{-5} + 5.94\times10^{-7} + \cdots + 1.61\times10^{-5}} \\ &= 0.0079.\end{aligned}$$

**3.6** (a) We use Neyman allocation (= optimal allocation when costs in the strata are equal), with $n_h \propto N_h S_h$.

We take $R_h$ to be the relative standard deviation in stratum $h$, and let $n_h = 900(N_h R_h)/125000$.

| Stratum | $N_h$ | $R_h$ | $N_h R_h$ | $n_h$ |
|---|---|---|---|---|
| Houses | 35,000 | 2 | 70,000 | 504 |
| Apartments | 45,000 | 1 | 45,000 | 324 |
| Condos | 10,000 | 1 | 10,000 | 72 |
| Sum | 90,000 | | 125,000 | 900 |

(b) Let's suppose we take a sample of 900 observations. (Any other sample size will give the same answer.)

With proportional allocation, we sample 350 houses, 450 apartments, and 100 condominiums. If the assumptions about the variances hold,

$$\begin{aligned}V_{\text{str}}[\hat{p}_{\text{str}}] &= \left(\frac{350}{900}\right)^2\frac{(.45)(.55)}{350} + \left(\frac{450}{900}\right)^2\frac{(.25)(.75)}{450} + \left(\frac{100}{900}\right)^2\frac{(.03)(.97)}{100} \\ &= .000215.\end{aligned}$$

If these proportions hold in the population, then

$$p = \frac{35}{90}(.45) + \frac{45}{90}(.25) + \frac{10}{90}(.03) = 0.3033$$

and, with an SRS of size 900,

$$V_{\text{srs}}[\hat{p}_{\text{srs}}] = \frac{(0.3033)(1-.3033)}{900} = .000235.$$

The gain in efficiency is given by

$$\frac{V_{\text{str}}[\hat{p}_{\text{str}}]}{V_{\text{srs}}[\hat{p}_{\text{srs}}]} = \frac{.000215}{.000235} = 0.9144.$$

For any sample size $n$, using the same argument as above, we have

$$V_{\text{str}}[\hat{p}_{\text{str}}] = \frac{.193233}{n} \quad \text{and}$$

$$V_{\text{srs}}[\hat{p}_{\text{srs}}] = \frac{.211322}{n}.$$

We only need $0.9144n$ observations, taken in a stratified sample with proportional allocation, to achieve the same variance as in an SRS with $n$ observations.

Note: The ratio $V_{\text{str}}[\hat{p}_{\text{str}}]/V_{\text{srs}}[\hat{p}_{\text{srs}}]$ is the *design effect*, to be discussed further in Section 7.5.

**3.7** (a) Here are summary statistics for each stratum:

| | Stratum | | | |
|---|---|---|---|---|
| | Biological | Physical | Social | Humanities |
| average | 3.142857 | 2.105263 | 1.230769 | 0.4545455 |
| variance | 6.809524 | 8.210526 | 4.358974 | 0.8727273 |

Since we took a simple random sample in each stratum, we use

$$(102)(3.142857) = 320.5714$$

to estimate the total number of publications in the biological sciences, with estimated variance

$$(102)^2\left(1 - \frac{7}{102}\right)\frac{6.809524}{7} = 9426.327.$$

The following table gives estimates of the total number of publications and estimated variance of the total for each of the four strata:

| Stratum | Estimated total number of publications | Estimated variance of total |
|---|---|---|
| Biological Sciences | 320.571 | 9426.33 |
| Physical Sciences | 652.632 | 38982.71 |
| Social Sciences | 267.077 | 14843.31 |
| Humanities | 80.909 | 2358.43 |
| Total | 1321.189 | 65610.78 |

We estimate the total number of refereed publications for the college by adding the totals for each of the strata; as sampling was done independently in each stratum, the variance of the college total is the sum of the variances of the population stratum totals. Thus we estimate the total number of refereed papers as 1321.2, with standard error $\sqrt{65610.78} = 256.15$.

(b) From Exercise 2.6, using an SRS of size 50, the estimated total was $\hat{t}_{\text{srs}} = 1436.46$, with standard error 296.2. Here, stratified sampling ensures that each division of the college is represented in the sample, and it produces an estimate with a smaller standard error than an SRS with the same number of observations. The sample variance in Exercise 2.8 was $s^2 = 7.19$. Only Physical Sciences had a

sample variance larger than 7.19; the sample variance in Humanities was only 0.87. Observations within many strata tend to be more homogeneous than observations in the population as a whole, and the reduction in variance in the individual strata often leads to a reduced variance for the population estimate.

(c)

| | $N_h$ | $n_h$ | $\hat{p}_h$ | $\frac{N_h}{N}\hat{p}_h$ | $\left(1-\frac{n_h}{N_h}\right)\frac{N_h^2}{N^2}\frac{\hat{p}_h(1-\hat{p}_h)}{n_h-1}$ |
|---|---|---|---|---|---|
| Biological Sciences | 102 | 7 | $\frac{1}{7}$ | .018 | .0003 |
| Physical Sciences | 310 | 19 | $\frac{10}{19}$ | .202 | .0019 |
| Social Sciences | 217 | 13 | $\frac{9}{13}$ | .186 | .0012 |
| Humanities | 178 | 11 | $\frac{8}{11}$ | .160 | .0009 |
| Total | 807 | 50 | | .567 | .0043 |

$$\begin{aligned} \hat{p}_{\text{str}} &= 0.567 \\ \text{SE}[\hat{p}_{\text{str}}] &= \sqrt{0.0043} = 0.066. \end{aligned}$$

**3.8** (a) Because the budget for interviews is \$15,000, a total of 15,000/30 = 500 in-person interviews can be taken. The variances in the phone and nonphone strata are assumed similar, so proportional allocation is optimal: 450 phone households and 50 nonphone households would be selected for interview.

(b) The variances in the two strata are assumed equal, so optimal allocation gives

$$n_h \propto N_h/\sqrt{c_h}.$$

| Stratum | $c_h$ | $N_h/N$ | $N_h/(N\sqrt{c_h})$ |
|---|---|---|---|
| Phone | 10 | 0.9 | 0.284605 |
| Nonphone | 40 | 0.1 | 0.015811 |
| Total | | 1.0 | 0.300416 |

The calculations in the table imply that

$$n_{\text{phone}} = \frac{0.284605}{0.300416}n;$$

the cost constraints imply that

$$10n_{\text{phone}} + 40n_{\text{non}} = 10n_{\text{phone}} + 40(n - n_{\text{phone}}) = 15{,}000.$$

Solving, we have

$$\begin{aligned} n_{\text{phone}} &= 1227 \\ n_{\text{non}} &= 68 \\ n &= 1295. \end{aligned}$$

Because of the reduced costs of telephone interviewing, more households can be selected in each stratum.

**3.9** (a) Summary statistics for acres87:

| Region | $N_h$ | $n_h$ | $\bar{y}_h$ | $s_h^2$ | $(N_h/N)\bar{y}_h$ | $\frac{N_h - n_h}{N_h}\frac{N_h^2}{N^2}\frac{s_h^2}{n_h}$ |
|---|---|---|---|---|---|---|
| NC | 1054 | 103 | 308188.3 | 2.943E+10 | 105532.98 | 30225148 |
| NE | 220 | 21 | 109009.6 | 1.005E+10 | 7791.46 | 2211633 |
| S | 1382 | 135 | 212687.2 | 5.698E+10 | 95495.05 | 76782239 |
| W | 422 | 41 | 654458.7 | 3.775E+11 | 89727.61 | 156241957 |
| Total | 3078 | 300 | | | 298547.10 | 265460977 |

For acres87, $\bar{y}_{str} = \sum_h (N_h/N)\bar{y}_h = 298547.1$ and

$$\text{SE}(\bar{y}_{str}) = \sqrt{\sum_h \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}} = 16293.$$

Of course, $\bar{y}_{str}$ could also be calculated using the column of weights in the data set, as:

$$\bar{y}_{\text{str}} = \frac{\sum_{i\in\mathcal{S}} w_i y_i}{\sum_{i\in\mathcal{S}} w_i} = \frac{918927923}{3078} = 298547.1$$

(b) Summary statistics for farms92:

| Region | $N_h$ | $n_h$ | $\bar{y}_h$ | $s_h^2$ | $(N_h/N)\bar{y}_h$ | $\frac{N_h - n_h}{N_h}\frac{N_h^2}{N^2}\frac{s_h^2}{n_h}$ |
|---|---|---|---|---|---|---|
| NC | 1054 | 103 | 750.68 | 128226.50 | 257.06 | 131.71 |
| NE | 220 | 21 | 528.10 | 128645.90 | 37.75 | 28.31 |
| S | 1382 | 135 | 578.59 | 222972.8 | 259.78 | 300.44 |
| W | 422 | 41 | 602.34 | 311508.4 | 82.58 | 128.94 |
| Total | 3078 | 300 | | | 637.16 | 589.40 |

For farms92, $\bar{y}_{str} = \sum_h (N_h/N)\bar{y}_h = 637.16$ and

$$\text{SE}(\bar{y}_{str}) = \sqrt{\sum_h \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}} = 24.28.$$

(c) Summary statistics for largef92:

| Region | $N_h$ | $n_h$ | $\bar{y}_h$ | $s_h^2$ | $(N_h/N)\bar{y}_h$ | $\frac{N_h - n_h}{N_h}\frac{N_h^2}{N^2}\frac{s_h^2}{n_h}$ |
|---|---|---|---|---|---|---|
| NC | 1054 | 103 | 70.91 | 4523.34 | 24.28 | 4.65 |
| NE | 220 | 21 | 8.19 | 90.16 | 0.59 | 0.02 |
| S | 1382 | 135 | 38.84 | 2450.47 | 17.44 | 3.30 |
| W | 422 | 41 | 104.98 | 11328.97 | 14.39 | 4.69 |
| Total | 3078 | 300 | | | 56.70 | 12.66 |

For largef92, $\bar{y}_{str} = \sum_h (N_h/N)\bar{y}_h = 56.70$ and

$$\text{SE}(\bar{y}_{str}) = \sqrt{\sum_h \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}} = 3.56.$$

(d) Summary statistics for smallf92:

| Region | $N_h$ | $n_h$ | $\bar{y}_h$ | $s_h^2$ | $(N_h/N)\bar{y}_h$ | $\frac{N_h - n_h}{N_h}\frac{N_h^2}{N^2}\frac{s_h^2}{n_h}$ |
|---|---|---|---|---|---|---|
| NC | 1054 | 103 | 44.26 | 1286.43 | 15.16 | 1.32 |
| NE | 220 | 21 | 47.24 | 2364.79 | 3.38 | 0.52 |
| S | 1382 | 135 | 47.39 | 6205.45 | 21.28 | 8.36 |
| W | 422 | 41 | 124.39 | 100640.94 | 17.05 | 41.66 |
| Total | 3078 | 300 | | | 56.86 | 51.86 |

For smallf92, $\bar{y}_{str} = \sum_h (N_h/N)\bar{y}_h = 56.86$ and

$$\text{SE}(\bar{y}_{str}) = \sqrt{\sum_h \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}} = 7.20.$$

Here is SAS code for finding these estimates:

```
data strattot;
   input region $  _total_;
   cards;
NE 220
NC 1054
S  1382
W  422
;

proc surveymeans data=agstrat total = strattot mean sum clm clsum df;
   stratum region ;
   var acres87 farms92 largef92 smallf92  ;
   weight strwt;
run;
```

**3.10** For this problem, note that $N_h$, the total number of dredge tows needed to cover the stratum, must be calculated. We use $N_h = 25.6 \times \text{Area}_h$.

(a) Calculate $\hat{t}_h = N_h\bar{y}_h$

| Stratum | $N_h$ | $n_h$ | $\bar{y}_h$ | $s_h^2$ | $\hat{t}_h$ | $N_h^2(1-n_h/N_h)\frac{s_h^2}{n_h}$ |
|---|---|---|---|---|---|---|
| 1 | 5704 | 4 | 0.44 | 0.068 | 2510 | 552718 |
| 2 | 1270 | 6 | 1.17 | 0.042 | 1486 | 11237 |
| 3 | 1286 | 3 | 3.92 | 2.146 | 5041 | 1180256 |
| 4 | 5064 | 5 | 1.80 | 0.794 | 9115 | 4068262 |
| Sum | 13324 | 18 | | | 18152 | 5812472 |

Thus $\hat{t}_{\text{str}} = 18152$ and

$$\text{SE}\,[\hat{t}_{\text{str}}] = \sqrt{5812472} = 2411.$$

(b)

| Stratum | $N_h$ | $n_h$ | $\bar{y}_h$ | $s_h^2$ | $\hat{t}_h$ | $N_h^2(1-n_h/N_h)\frac{s_h^2}{n_h}$ |
|---|---|---|---|---|---|---|
| 1 | 8260 | 8 | 0.63 | 0.083 | 5204 | 707176 |
| 4 | 5064 | 5 | 0.40 | 0.046 | 2026 | 235693 |
| Sum | 13324 | 13 | | | 7229 | 942869 |

Here, $\hat{t}_{\text{str}} = 7229$ and

$$\text{SE}\,[\hat{t}_{\text{str}}] = \sqrt{942869} = 971.$$

**3.11** Note that the paper is somewhat ambiguous on how the data were collected. The abstract says random stratified sampling was used, while on p. 224 the authors say: 'a sampling grid covering 20% of the total area was made ... by picking 40 numbers between one and 200 with the random number generator." It's possible that poststratification was really used, but for exercise purposes, let's treat it as a stratified random sample. Also note that the original data were not available, data were generated that were consistent with summary statistics in the paper.

(a) Summary statistics are in the following table:

| Zone | $N_h$ | $n_h$ | $\bar{y}_h$ | $s_h^2$ |
|---|---|---|---|---|
| 1 | 68 | 17 | 1.765 | 3.316 |
| 2 | 84 | 12 | 4.417 | 11.538 |
| 3 | 48 | 11 | 10.545 | 46.073 |
| Total | 200 | 40 | | |

Using (3.1),

$$\begin{aligned}\hat{t}_{\text{str}} &= \sum_h N_h \bar{y}_h \\ &= 68(1.76) + 84(4.42) + 48(10.55) \\ &= 997.\end{aligned}$$

From (3.3),

$$\begin{aligned}
\hat{V}(\hat{t}_{\text{str}}) &= \left(1-\frac{N_h}{N}\right)\sum_{h=1}^{H} {N_h}^2 \frac{s_h^2}{n_h} \\
&= \left(1-\frac{17}{68}\right) 68^2 \frac{3.316}{17} + \left(1-\frac{12}{84}\right) 84^2 \frac{11.538}{12} + \left(1-\frac{11}{48}\right) 48^2 \frac{46.073}{11} \\
&= 676.5 + 5815.1 + 7438.7 \\
&= 13930.2,
\end{aligned}$$

so

$$\text{SE}(\hat{t}_{y\text{str}}) = \sqrt{13930.2} = 118.$$

SAS code to calculate these quantities is given below:

```
data seals;
   infile seals delimiter="," firstobs=2;
   input zone holes;
   if zone = 1 then sampwt = 68/17;
   if zone = 2 then sampwt = 84/12;
   if zone = 3 then sampwt = 48/11;
run;

data strattot;
   input zone  _total_;
   datalines;
1 68
2 84
3 48
;

proc surveymeans data=seals total=strattot mean clm sum clsum;
   strata zone;
   weight sampwt;
   var holes;
run;

The SURVEYMEANS Procedure

            Data Summary

Number of Strata                   3
Number of Observations            40
Sum of Weights                   200
```

```
                          Statistics

                                    Std Error
Variable            Mean          of Mean       95% CL for Mean

holes           4.985909         0.590132    3.79018761 6.18163058

                          Statistics

Variable            Sum          Std Dev        95% CL for Sum

holes        997.181818       118.026447    758.037521 1236.32612
```

(b) (i) If the goal is estimating the total number of breathing holes, we should use optimal allocation. Using the values of $s_h^2$ from this survey as estimates of $S_h^2$, we have:

| Zone | $N_h$ | $s_h^2$ | $N_h s_h$ |
|---|---|---|---|
| 1 | 68 | 3.316 | 123.83 |
| 2 | 84 | 11.538 | 285.33 |
| 3 | 48 | 46.073 | 325.81 |
| Total | 200 | | 734.97 |

Then $n_1 = (123.83/734.97)n = 0.17n$; $n_2 = 0.39n$; $n_3 = 0.44n$. The high variance in zone 3 leads to a larger sample size in that zone.

(ii) If the goal is to compare the density of the breathing holes in the three zones, we would like to have equal precision for $\bar{y}_h$ in the three strata. Ignoring the fpc, that means we would like

$$\frac{S_1^2}{n_1} = \frac{S_2^2}{n_2} = \frac{S_3^2}{n_3},$$

which implies that $n_h$ should be proportional to $S_h^2$ to achieve equal variances.

Using the sample variances $s_h^2$ instead of the unknown population variances $S_h^2$, this leads to

$$\begin{aligned} n_1 &= \frac{s_1^2}{s_1^2 + s_2^2 + s_3^2} n = 0.05n \\ n_2 &= 0.19n \\ n_3 &= 0.76n. \end{aligned}$$

**3.12** We use $n_h = 300 N_h s_h / \sum_k N_k s_k$

| Region | $N_h$ | $N_h s_h$ | $n_h$ |
|---|---|---|---|
| Northeast | 220 | 19,238,963 | 7 |
| North Central | 1,054 | 181,392,707 | 69 |
| South | 1,382 | 319,918,785 | 122 |
| West | 422 | 265,620,742 | 101 |
| Total | 3,078 | 786,171,197 | 300 |

**3.13** Answers will vary since students select different samples.

**3.14**

| Method | $\hat{p}_{\text{str}}$ | SE$[\hat{p}_{\text{str}}]$ |
|---|---|---|
| Role play | 0.96 | 0.011 |
| Problem solving | 0.82 | 0.217 |
| Simulations | 0.45 | 0.028 |
| Empathy building | 0.45 | 0.028 |
| Gestalt exercises | 0.11 | 0.017 |

Note that the standard errors calculated for role play and for gestalt exercises are unreliable because the formula relies on a normal approximation to $\hat{p}_h$: here, the sample sizes in the strata are small and $\hat{p}_h$'s are close to 0 or 1, so the accuracy of the normal approximation is questionable.

**3.15** (a) An advantage is that using the same number of stores in each stratum gives the best precision for comparing strata if the within-stratum variances are the same. In addition, people may perceive that allocation as fair. A disadvantage is that estimates may lose precision relative to optimal allocation if some strata have higher variances than others.

(b)

$$\bar{y}_{\text{str}} = \sum_{h=1}^{3} \frac{N_h}{N} \bar{y}_h = 3.9386.$$

$$\begin{aligned} \hat{V}(\bar{y}_{\text{str}}) &= \sum_{h=1}^{3} \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h} \\ &= 8.85288 \times 10^{-5} \end{aligned}$$

Thus, a 95% CI is

$$3.9386 \pm 1.96\sqrt{8.85288 \times 10^{-5}} = [3.92, 3.96]$$

This is a very small CI. Remember, though, that it reflects only the sampling error. In this case, the author was unable to reach some of the stores and in addition some of the market basket items were missing, so there was nonsampling error as well.

**3.16** (a)

| Stratum | $N_h$ | $n_h$ | $\bar{y}_h$ | $s_h^2$ | $\hat{t}_h$ | $\hat{V}(\hat{t}_h)$ |
|---|---|---|---|---|---|---|
| 1 | 89 | 19 | 1.74 | 5.43 | 154.6 | 1779.5 |
| 2 | 61 | 20 | 1.75 | 6.83 | 106.8 | 854.0 |
| 3 | 40 | 22 | 13.27 | 58.78 | 530.9 | 1923.7 |
| 4 | 47 | 21 | 4.10 | 15.59 | 192.5 | 907.2 |
| Total | 237 | 82 | | | 984.7 | 5464.3 |

The estimated total number of otter holts is

$$\hat{t}_{\text{str}} = 985$$

with

$$\mathrm{SE}\left[\hat{t}_{\mathrm{str}}\right] = \sqrt{5464} = 73.9.$$

Here is SAS code and output for estimating these quantities:

```
data exer0316;
   infile otters delimiter=',' firstobs=2;
   input section habitat holts;
   if habitat = 1 then sampwt = 89/19;
   if habitat = 2 then sampwt = 61/20;
   if habitat = 3 then sampwt = 40/22;
   if habitat = 4 then sampwt = 47/21;
;

data strattot;
   input habitat  _total_;
   datalines;
1 89
2 61
3 40
4 47
;
proc surveymeans data=exer0316 total = strattot mean clm sum clsum;
   stratum habitat;
   weight sampwt;
   var holts;
run;

The SURVEYMEANS Procedure

            Data Summary

Number of Strata                   4
Number of Observations            82
Sum of Weights                   237


                           Statistics

                               Std Error
Variable            Mean         of Mean       95% CL for Mean

holts           4.154912        0.311903    3.53396136 4.77586336

                           Statistics
```

```
Variable            Sum        Std Dev       95% CL for Sum

holts        984.714229      73.920990   837.548842 1131.87962
```

**3.17** (a) We form a new variable, weight= 1/samprate. Then the number of divorces in the divorce registration area is

$$\sum_{h=1}^{H}(\text{weight})_h(\text{numrecs})_h = 571{,}185.$$

Note that this is the population value, not an estimate, because samprate= $n_h/N_h$ and $(\text{numrecs})_h = n_h$. Thus

$$\sum_{h=1}^{H}(\text{weight})_h(\text{numrecs})_h = \sum_{h=1}^{H}\frac{N_h}{n_h}n_h = N.$$

(b) They wanted a specified precision within each state (= stratum). You can see that, except for a few states in which a census is taken, the number of records sampled is between 2400 and 6200. That gives roughly the same precision for estimates within each of those states. If the same sampling rate were used in each state, states with large population would have many more records sampled than states with small population.

(c) (i) For each stratum,

$$\bar{y}_h = \hat{p}_h = \frac{\text{hsblt20} + \text{hsb20-24}}{\text{numrecs}}.$$

The following spreadsheet shows calculations done to obtain

$$\hat{t}_{\text{str}} = \sum_h N_h\hat{p}_h$$

and

$$\hat{V}(\hat{t}_{\text{str}}) = \sum_h N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1} = \sum_h \text{varcont}_h.$$

| state | rate | $n_h$ | $N_h$ | husb $\leq 24$ | $\hat{p}_h$ | $N_h\hat{p}_h$ | varcont |
|---|---|---|---|---|---|---|---|
| AL | 0.1 | 2460 | 24600 | 295 | 0.11992 | 2950 | 23376 |
| AK | 1 | 3396 | 3396 | 371 | 0.10925 | 371 | 0 |
| CT | 0.5 | 6003 | 12006 | 333 | 0.05547 | 666 | 629 |
| DE | 1 | 2938 | 2938 | 238 | 0.08101 | 238 | 0 |
| DC | 1 | 2525 | 2525 | 90 | 0.03564 | 90 | 0 |
| GA | 0.1 | 3404 | 34040 | 440 | 0.12926 | 4400 | 34491 |
| HI | 1 | 4415 | 4415 | 394 | 0.08924 | 394 | 0 |
| ID | 0.5 | 2949 | 5898 | 380 | 0.12886 | 760 | 662 |
| IL | 1 | 46986 | 46986 | 4349 | 0.09256 | 4349 | 0 |
| IA | 0.5 | 5259 | 10518 | 541 | 0.10287 | 1082 | 971 |
| KS | 0.5 | 6170 | 12340 | 768 | 0.12447 | 1536 | 1345 |
| KY | 0.2 | 3879 | 19395 | 567 | 0.14617 | 2835 | 9685 |
| MD | 0.2 | 3104 | 15520 | 156 | 0.05026 | 780 | 2964 |
| MA | 0.2 | 3367 | 16835 | 163 | 0.04841 | 815 | 3103 |
| MI | 0.1 | 3996 | 39960 | 270 | 0.06757 | 2700 | 22664 |
| MO | 1 | 24984 | 24984 | 2876 | 0.11511 | 2876 | 0 |
| MT | 1 | 4125 | 4125 | 432 | 0.10473 | 432 | 0 |
| NE | 1 | 6236 | 6236 | 620 | 0.09942 | 620 | 0 |
| NH | 1 | 4947 | 4947 | 458 | 0.09258 | 458 | 0 |
| NY | 1 | 67993 | 67993 | 3809 | 0.05602 | 3809 | 0 |
| OH | 0.05 | 2465 | 49300 | 102 | 0.04138 | 2040 | 37171 |
| OR | 0.2 | 3124 | 15620 | 233 | 0.07458 | 1165 | 4314 |
| PA | 0.1 | 3883 | 38830 | 248 | 0.06387 | 2480 | 20900 |
| RI | 1 | 3684 | 3684 | 246 | 0.06678 | 246 | 0 |
| SC | 1 | 13835 | 13835 | 1429 | 0.10329 | 1429 | 0 |
| SD | 1 | 2699 | 2699 | 93 | 0.03446 | 93 | 0 |
| TN | 0.1 | 3042 | 30420 | 426 | 0.14004 | 4260 | 32982 |
| UT | 0.5 | 4489 | 8978 | 591 | 0.13166 | 1182 | 1027 |
| VT | 1 | 2426 | 2426 | 162 | 0.06678 | 162 | 0 |
| VA | 1 | 25608 | 25608 | 2075 | 0.08103 | 2075 | 0 |
| WI | 0.2 | 3384 | 16920 | 280 | 0.08274 | 1400 | 5138 |
| WY | 1 | 3208 | 3208 | 346 | 0.10786 | 346 | 0 |
| Total | | 280983 | 571185 | | | 49039 | 201422 |

Thus, for estimating the total number of divorces granted to men aged 24 or less,

$$\hat{t}_{\text{str}} = 49039$$

and

$$\text{SE}(\hat{t}_{\text{str}}) = \sqrt{201{,}422} = 449.$$

A 95% confidence interval is

$$49039 \pm (1.96)(449) = [48159, 49919]$$

(ii) Similarly, for the women,

$$\hat{t}_{\text{str}} = 4600 + 664 + 1330 + \cdots + 658 = 86619$$

and

$$\text{SE}(\hat{t}_{\text{str}}) = \sqrt{33672 + 0 + 1183 + \cdots 8327 + 0} = 564.$$

A 95% CI is [85513, 87725].

(d) For estimating the proportions,

$$\hat{P}_{\text{str}} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{P}_h$$

and

$$\hat{V}(\hat{p}_{\text{str}}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}$$

(i) For the men:

$$\hat{p}_{\text{str}} = \frac{24600}{571185}\left(\frac{390}{2460}\right) + \frac{3396}{571185}\left(\frac{647}{3396}\right) + \cdots \frac{3208}{571185}\left(\frac{560}{3208}\right) = 0.1928.$$

$$\begin{aligned} \text{SE}(\hat{p}_{\text{str}}) &= \sqrt{(9.06 \times 10^{-8}) + 0 + (6.54 \times 10^{-9}) + \cdots + (3.37 \times 10^{-8}) + 0} \\ &= 1.068 \times 10^{-3}. \end{aligned}$$

A 95% confidence interval for the proportion of men aged 40–49 at the time of the decree is

$$0.1928 \pm 1.96(1.068 \times 10^{-3}) = [0.191, 0.195].$$

(ii) For the women:

$$\hat{p}_{\text{str}} = \frac{24600}{571185}\left(\frac{296}{2460}\right) + \frac{3396}{571185}\left(\frac{495}{3396}\right) + \cdots + \frac{3208}{571185}\left(\frac{437}{3208}\right) = 0.1566$$

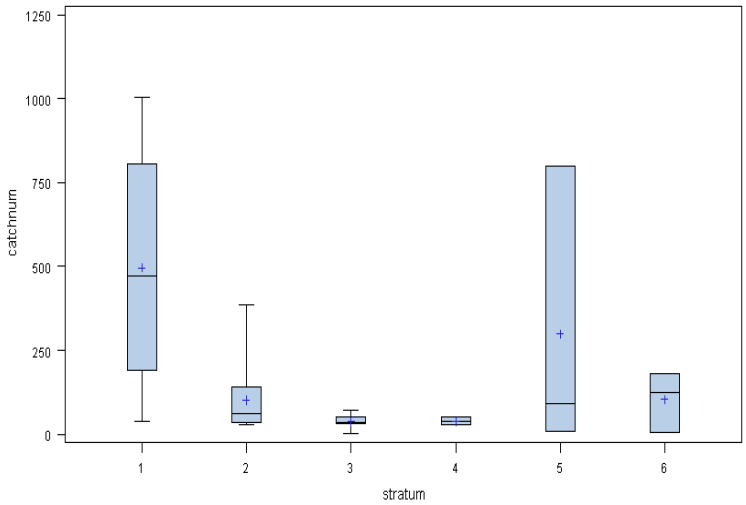
and

$$\begin{aligned} \text{SE}(\hat{p}_{\text{str}}) &= \sqrt{(7.19 \times 10^{-8}) + 0 + (5.80 \times 10^{-9}) + \cdots + (2.83 \times 10^{-8}) + 0} \\ &= 9.75 \times 10^{-4}. \end{aligned}$$

A 95% confidence interval is

$$0.1566 \pm 1.96(9.75 \times 10^{-4}) = [0.155, 0.158].$$

**3.18** (a)




(b) Let $w_h$ be the relative sampling weight for stratum $h$. Then $N_h \propto n_h w_h$. For each response, we may calculate

$$\bar{y}_{\text{str}} = \sum_h n_h w_h \bar{y}_h \Big/ \sum_h n_h w_h;$$

equivalently, we may define a new column

$$\text{weight}_{hi} = \begin{cases} 1 & \text{if } h = 1 \text{ or} 2 \\ 2 & \text{if } h \in \{3,4,5,6\} \end{cases}$$

and calculate

$$\bar{y}_{\text{str}} = \sum_h \sum_i \text{weight}_{hi} y_{hi} \Big/ \sum_h \sum_i \text{weight}_{hi}.$$

Summary statistics for 1974 are:

| | | | | Number of fish | | Weight of fish | |
|---|---|---|---|---|---|---|---|
| Stratum | $n_h$ | $w_h$ | $N_h/N$ | $\bar{y}_h$ | $s_h^2$ | $\bar{y}_h$ | $s_h^2$ |
| 1 | 13 | 1 | 0.213 | 496.2 | 108528.8 | 60.4 | 522.5 |
| 2 | 12 | 1 | 0.197 | 101.9 | 10185.0 | 32.1 | 299.5 |
| 3 | 9 | 2 | 0.295 | 38.2 | 504.4 | 15.0 | 130.9 |
| 4 | 3 | 2 | 0.098 | 39.0 | 147.0 | 6.9 | 9.1 |
| 5 | 3 | 2 | 0.098 | 299.3 | 189681.3 | 39.2 | 3598.4 |
| 6 | 3 | 2 | 0.098 | 103.7 | 8142.3 | 8.2 | 76.4 |

We have, using (3.5) and ignoring the fpc when calculating the standard error,

| Response | $\hat{\bar{y}}_{str}$ | SE($\hat{\bar{y}}_{str}$) |
|---|---|---|
| Number of fish | 180.5 | 32.5 |
| Weight of fish | 29.0 | 4.0 |

SAS code for calculating these estimates is given below.

```
data nybight;
   infile nybight delimiter=',' firstobs=2;
   input year  stratum catchnum catchwt numspp depth temp ;
   select (stratum);
      when (1,2) relwt=1;
  when (3,4,5,6) relwt=2;
   end;
   if year = 1974;
proc surveymeans data=nybight mean clm ;
   stratum stratum;
   var catchnum catchwt;
   weight relwt;
run;
```

(c) The procedure is the same as that in part (b). Summary statistics for 1975 are:

| | | | Number of fish | | Weight of fish | |
|---|---|---|---|---|---|---|
| Stratum | $n_h$ | $w_h$ | $\bar{y}_h$ | $s_h^2$ | $\bar{y}_h$ | $s_h^2$ |
| 1 | 14 | 1 | 486.9 | 94132.0 | 127.0 | 3948.0 |
| 2 | 16 | 1 | 262.7 | 42234.8 | 109.5 | 7189.8 |
| 3 | 15 | 2 | 119.6 | 9592.0 | 33.5 | 867.8 |
| 4 | 13 | 2 | 238.3 | 12647.2 | 84.1 | 1583.6 |
| 5 | 3 | 2 | 119.7 | 789.3 | 20.6 | 18.4 |
| 6 | 3 | 2 | 70.7 | 3194.3 | 12.0 | 255.0 |

| Response | $\hat{\bar{y}}_{str}$ | SE($\hat{\bar{y}}_{str}$) |
|---|---|---|
| Number of fish | 223.9 | 18.5 |
| Weight of fish | 70.6 | 5.7 |

**3.19** (a)

| Stratum | Respondents to survey | Respondents to *breakaga*, $n_h$ |
|---|---|---|
| 1 | 288 | 232 |
| 2 | 533 | 514 |
| 3 | 91 | 86 |
| 4 | 73 | 67 |
| Total | 985 | 899 |

(b) In the table,

$$\text{varcont}_h = \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}.$$

| Stratum | $N_h$ | $n_h$ | $\hat{p}_h$ | $(\frac{N_h}{N})\hat{p}_h$ | varcont |
|---|---|---|---|---|---|
| 1 | 1374 | 232 | .720 | .269 | $1.01 \times 10^{-4}$ |
| 2 | 1960 | 514 | .893 | .475 | $3.90 \times 10^{-5}$ |
| 3 | 252 | 86 | .872 | .060 | $4.05 \times 10^{-6}$ |
| 4 | 95 | 67 | .866 | .022 | $3.46 \times 10^{-7}$ |
| Total | 3681 | 899 | | .826 | $1.44 \times 10^{-4}$ |

Thus,

$$\begin{aligned} \hat{p}_{\text{str}} &= 0.826 \\ \text{SE}\,(\hat{p}_{\text{str}}) &= \sqrt{1.44 \times 10^{-4}} = 0.012. \end{aligned}$$

(c) The weights are as follows:

| Stratum | weight |
|---|---|
| 1 | 5.922 |
| 2 | 3.813 |
| 3 | 2.930 |
| 4 | 1.418 |

The answer again is $\hat{p}_{str} = 0.826$.

(e)

| Stratum | Employee Type | Survey Response Rate (%) | breakaga Response Rate (%) |
|---|---|---|---|
| 1 | Faculty | 58 | 46 |
| 2 | Classified staff | 82 | 79 |
| 3 | Administrative staff | 93 | 88 |
| 4 | Academic professional | 77 | 71 |

The faculty have the lowest response rate (somehow, this did not surprise me). Stratification assumes that the nonrespondents in a stratum are similar to respondents in that stratum.

**3.20** (b) This is done by dividing popsize/sampsize for each county. The first few weights for strata are:

| countynum | countyname | weight |
|---|---|---|
| 1 | Aitkin | 1350.0 |
| 2 | Anoka | 1261.4 |
| 3 | Beltrami | 2750.0 |

(c)

| response | mean | SE | 95% CI |
|---|---|---|---|
| radon | 4.898551 | 0.154362 | [4.59560775, 5.20149511] |
| lograd | 1.301306 | 0.028777 | [1.24482928, 1.35778371] |

(d) The total number of homes with excessive radon is estimated as 722781, with

standard error 28107 and 95% CI [667620, 777942].

**3.22** (a) $n_h = 2000(N_h S_h / \sum_i N_i S_i)$

| Stratum | $N_h/N$ | $S_h$ | $S_h N_h/N$ | $n_h$ |
|---|---|---|---|---|
| 1 | 0.4 | $\sqrt{(.10)(.90)} = .3000$ | .1200 | 1079 |
| 2 | 0.6 | $\sqrt{(.03)(.97)} = .1706$ | .1024 | 921 |
| Total | 1.0 | | .2224 | 2000 |

(b) For both proportional and optimal allocation,

$$S_1^2 = (.10)(.90) = .09 \quad \text{and} \quad S_2^2 = (.03)(.97) = .0291.$$

Under proportional allocation $n_1 = 800$ and $n_2 = 1200$.

$$V_{\text{prop}}(\hat{p}_{\text{str}}) = (.4)^2 \frac{.09}{800} + (.6)^2 \frac{.0291}{1200} = 2.67 \times 10^{-5}.$$

For optimal allocation,

$$V_{\text{opt}}(\hat{p}_{\text{str}}) = (.4)^2 \frac{.09}{1079} + (.6)^2 \frac{.0291}{921} = 2.47 \times 10^{-5}.$$

For an SRS, $p = (.4)(.10) + (.6)(.03) = .058$

$$V_{\text{srs}}(\hat{p}_{\text{srs}}) = \frac{.058(1 - .058)}{2000} = 2.73 \times 10^{-5}.$$

**3.23**

(a) We take an SRS of $n/H$ observations from each of the $N/H$ strata, so there are a total of

$$\left( \begin{array}{c} N/H \\ n/H \end{array} \right)^H = \left[ \frac{(N/H)!}{(n/H)!(N/H - n/H)!} \right]^H$$

possible stratified samples.

(b) By Stirling's formula,

$$\begin{aligned} \left( \begin{array}{c} N \\ n \end{array} \right) &= \frac{N!}{n!(N-n)!} \\ &\approx \frac{\sqrt{2\pi N} \left(\frac{N}{e}\right)^N}{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n \sqrt{2\pi(N-n)} \left(\frac{N-n}{e}\right)^{N-n}} \\ &= \sqrt{\frac{N}{2\pi n(N-n)}} \, \frac{N^N}{n^n (N-n)^{N-n}} \end{aligned}$$

We use the same argument, substituting $N/H$ for $N$ and $n/H$ for $n$ in the equation above, to obtain:

$$\left( \begin{array}{c} N/H \\ n/H \end{array} \right) \approx \sqrt{\frac{NH}{2\pi n(N-n)}} \, \frac{N^{N/H}}{n^{n/H}(N-n)^{(N-n)/H}}.$$

Consequently,

$$\frac{\left(\begin{array}{c} N/H \\ n/H \end{array}\right)^H}{\left(\begin{array}{c} N \\ n \end{array}\right)} \approx \frac{\left[\sqrt{\frac{NH}{2\pi n(N-n)}}\ \frac{N^{N/H}}{n^{n/H}(N-n)^{(N-n)/H}}\right]^H}{\sqrt{\frac{N}{2\pi n(N-n)}}\ \frac{N^N}{n^n(N-n)^{N-n}}}$$
$$= \left[\frac{N}{2\pi n(N-n)}\right]^{(H-1)/2} H^{H/2}.$$

**3.24** We wish to minimize

$$V[\hat{t}_{\text{str}}] = \sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)N_h^2\frac{S_h^2}{n_h}$$

subject to the constraint

$$C = c_0 + \sum_{h=1}^{H} c_h n_h.$$

Lagrange multipliers are often used for such problems. (See, for example, Thomas, G.B. and Finney, R. L. (1982). *Calculus and Analytic Geometry, Fifth edition.* Reading, MA: Addison-Wesley, p. 617.)

Define

$$f(n_1,\ldots,n_H,\lambda) = \sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)N_h^2\frac{S_h^2}{n_h} - \lambda\left(C - c_0 - \sum_{h=1}^{H} c_h n_h\right).$$

Then

$$\frac{\partial f}{\partial n_k} = -N_k^2\frac{S_k^2}{n_k^2} + \lambda c_k$$

$k = 1,\ldots,H$, and

$$\frac{\partial f}{\partial \lambda} = c_0 + \sum_{h=1}^{H} c_h n_h - C.$$

Setting the partial derivatives equal to 0 and solving gives

$$n_k = \frac{N_k S_k}{\sqrt{\lambda c_k}}$$

for $k = 1,\ldots,H$, and

$$\sum_{h=1}^{H} c_h n_h = C - c_0,$$

which implies that

$$\sqrt{\lambda} = \frac{\sum_{h=1}^{H}\sqrt{c_h}N_h S_h}{C - c_0}$$

and hence that

$$n_k = \frac{N_k S_k}{\sqrt{c_k}} \frac{C - c_0}{\sum_{h=1}^{H} \sqrt{c_h} N_h S_h}.$$

Note that we also have $n_k \propto N_k S_k / \sqrt{c_k}$ if we want to minimize the cost for a fixed variance $N$. Then, let

$$g(n_1, \ldots, n_H, \lambda) = c_0 + \sum_{h=1}^{H} c_h n_h - \lambda \left[ V - \sum_{h=1}^{H} \left( 1 - \frac{n_h}{N_h} \right) N_h^2 \frac{S_h^2}{n_h} \right].$$

Then

$$\frac{\partial g}{\partial n_k} = c_k - \lambda N_k^2 S_k^2 / n_k^2$$

and

$$\frac{\partial g}{\partial \lambda} = V - \sum_{h=1}^{H} \left( 1 - \frac{n_h}{N_h} \right) N_h^2 \frac{S_h^2}{n_h}.$$

Setting the partial derivatives equal to 0 and solving gives

$$n_k = \frac{\sqrt{\lambda} N_k S_k}{\sqrt{c_k}}.$$

**3.25** (a) We substitute $n_{h,\text{Neyman}}$ for $n_h$ in (3.4):

$$\begin{aligned}
V_{\text{Neyman}}(\hat{t}_{\text{str}}) &= \sum_{h=1}^{H} \left( 1 - \frac{n_h}{N_h} \right) N_h^2 \frac{S_h^2}{n_h} \\
&= \sum_{h=1}^{H} \left( 1 - \frac{N_h S_h n}{N_h \sum_{l=1}^{H} N_l S_l} \right) N_h^2 \frac{S_h^2 \sum_{l=1}^{H} N_l S_l}{n_h N_h S_h n} \\
&= \sum_{h=1}^{H} \left( 1 - \frac{S_h n}{\sum_{l=1}^{H} N_l S_l} \right) \frac{N_h S_h \sum_{l=1}^{H} N_l S_l}{n} \\
&= \frac{1}{n} \left( \sum_{h=1}^{H} N_h S_h \right)^2 - \sum_{h=1}^{H} N_h S_h^2.
\end{aligned}$$

(b)

$$\begin{aligned} V_{\text{prop}}(\hat{t}_{\text{str}}) - V_{\text{Neyman}}(\hat{t}_{\text{str}}) &= \frac{N}{n}\sum_{h=1}^{H} N_h S_h^2 - \sum_{h=1}^{H} N_h S_h^2 \\ &\quad -\frac{1}{n}\left(\sum_{h=1}^{H} N_h S_h\right)^2 + \sum_{h=1}^{H} N_h S_h^2 \\ &= \frac{N}{n}\sum_{h=1}^{H} N_h S_h^2 - \frac{1}{n}\left(\sum_{h=1}^{H} N_h S_h\right)^2 \\ &= \frac{N^2}{n}\left[\sum_{h=1}^{H}\frac{N_h}{N} S_h^2 - \left(\sum_{h=1}^{H}\frac{N_h}{N} S_h\right)^2\right] \\ &= \frac{N^2}{n}\sum_{h=1}^{H}\frac{N_h}{N}\left[S_h^2 - S_h\sum_{l=1}^{H}\frac{N_l}{N} S_l\right] \end{aligned}$$

But

$$\begin{aligned} \sum_{h=1}^{H}\frac{N_h}{N}\left[S_h - \sum_{l=1}^{H}\frac{N_l}{N} S_l\right]^2 &= \sum_{h=1}^{H}\frac{N_h}{N}\left[S_h^2 - 2S_h\sum_{l=1}^{H}\frac{N_l}{N} S_l + \left(\sum_{l=1}^{H}\frac{N_l}{N} S_l\right)^2\right] \\ &= \sum_{h=1}^{H}\frac{N_h}{N} S_h^2 - \left(\sum_{l=1}^{H}\frac{N_l}{N} S_l\right)^2, \end{aligned}$$

proving the result.

(c) When $H = 2$, the difference from (b) is

$$\begin{aligned} &\frac{N^2}{n}\sum_{h=1}^{2}\frac{N_h}{N}\left(S_h - \sum_{l=1}^{2}\frac{N_l}{N} S_l\right)^2 \\ &= \frac{N^2}{n}\left[\frac{N_1}{N}\left(S_1 - \frac{N_1}{N} S_1 - \frac{N_2}{N} S_2\right)^2 + \frac{N_2}{N}\left(S_2 - \frac{N_1}{N} S_1 - \frac{N_2}{N} S_2\right)^2\right] \\ &= \frac{N^2}{n}\left[\frac{N_1}{N}\left(\frac{N_2}{N} S_1 - \frac{N_2}{N} S_2\right)^2 + \frac{N_2}{N}\left(\frac{N_1}{N} S_2 - \frac{N_1}{N} S_1\right)^2\right] \\ &= \frac{N^2}{n}\frac{N_1}{N}\frac{N_2}{N}\left[\frac{N_1}{N} + \frac{N_2}{N}\right](S_1 - S_2)^2 \\ &= \frac{N_1 N_2}{n}(S_1 - S_2)^2. \end{aligned}$$

**3.34**

(a) In the data step, define the variable one to have the value 1 for every observation. Then $\sum_{i\in\mathcal{S}} w_i 1 = N$. Here, $\sum_{i\in\mathcal{S}} w_i 1 = 85174776$. The standard error is zero

because this is a stratified sample. The weights are $N_h/n_h$ so the sum of the weights in stratum $h$ is $N_h$ exactly. There is no sampling variability.

Here is the code used to obtain these values:

```
proc surveymeans data=vius mean clm sum clsum;
   weight tabtrucks;
   stratum stratum;
   var  one miles_annl mpg;
```

(b) The estimated total number of truck miles driven is 1.115 $\times 10^{12}$; the standard error is 6492344384 and a 95% CI is [1.102$\times 10^{12}$, 1.127$\times 10^{12}$].

(c) Because these are stratification variables, we can calculate estimates for each truck type by summing $w_{hj}y_{hj}$ separately for each $h$. We obtain:

```
proc sort data=vius;
   by trucktype;
proc surveymeans data=vius sum clsum;
   by trucktype;
   weight tabtrucks;
   stratum stratum;
   var miles_annl;
   ods output Statistics=Mystat;
proc print data=Mystat;
run;
```

| Obs | VarName | VarLabel | Sum |
|---|---|---|---|
| 1 | MILES_ANNL | Number of Miles Driven During 2002 | 428294502082 |
| 2 | MILES_ANNL | Number of Miles Driven During 2002 | 541099850893 |
| 3 | MILES_ANNL | Number of Miles Driven During 2002 | 41279084490 |
| 4 | MILES_ANNL | Number of Miles Driven During 2002 | 31752656137 |
| 5 | MILES_ANNL | Number of Miles Driven During 2002 | 72301789843 |

| Obs | LowerCLSum | StdDev | UpperCLSum |
|---|---|---|---|
| 1 | 4.19064E11 | 4708839922 | 4.37525E11 |
| 2 | 5.32459E11 | 4408042207 | 5.4974E11 |
| 3 | 4.05032E10 | 395841910 | 4.2055E10 |
| 4 | 3.107E10 | 348294378 | 3.24353E10 |
| 5 | 7.12861E10 | 518195242 | 7.33175E10 |

(d) The estimated average mpg is 16.515427 with standard error 0.039676; a 95% CI is [16.4377, 16.5932]. These CIs are very small because the sample size is so large.

# Chapter 4

# Ratio and Regression Estimation

**4.2**

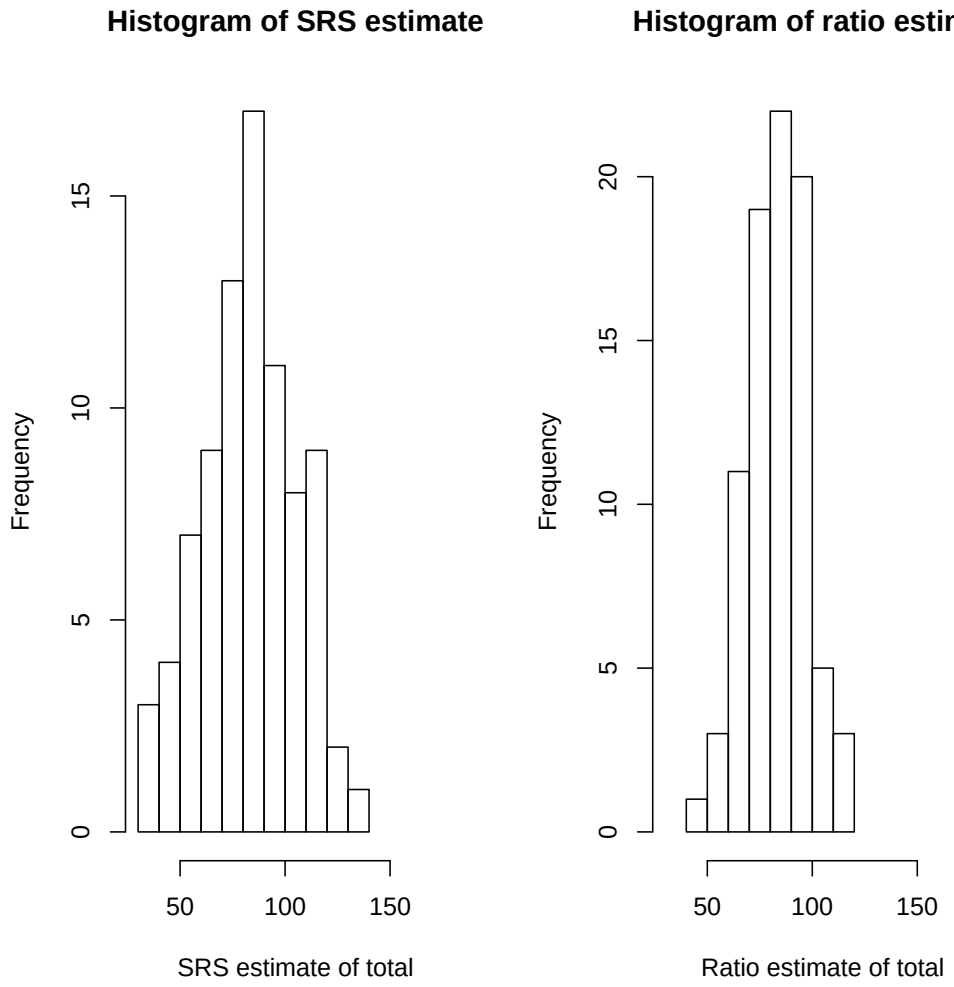
(a) We have $t_x = 69$. $t_y = 83$, $S_x = 4.092676$, $S_y = 5.333333$, $R = 0.8112815$, and $B = 1.202899$.

(b)

| Sample Number | Sample $\mathcal{S}$ | $\bar{x}_{\mathcal{S}}$ | $\bar{y}_{\mathcal{S}}$ | $\hat{B}$ | $\hat{t}_{\text{SRS}}$ | $\hat{t}_{yr}$ |
|---|---|---|---|---|---|---|
| 1 | $\{1,2,3\}$ | 10.333 | 10.000 | 0.968 | 90.000 | 66.774 |
| 2 | $\{1,2,4\}$ | 10.667 | 11.333 | 1.063 | 102.000 | 73.313 |
| 3 | $\{1,2,5\}$ | 8.000 | 8.333 | 1.042 | 75.000 | 71.875 |
| 4 | $\{1,2,6\}$ | 7.667 | 6.000 | 0.783 | 54.000 | 54.000 |
| 5 | $\{1,2,7\}$ | 10.333 | 11.000 | 1.065 | 99.000 | 73.452 |
| 6 | $\{1,2,8\}$ | 7.667 | 8.000 | 1.043 | 72.000 | 72.000 |
| 7 | $\{1,2,9\}$ | 8.333 | 7.000 | 0.840 | 63.000 | 57.960 |
| 8 | $\{1,3,4\}$ | 12.000 | 13.333 | 1.111 | 120.000 | 76.667 |
| 9 | $\{1,3,5\}$ | 9.333 | 10.333 | 1.107 | 93.000 | 76.393 |
| 10 | $\{1,3,6\}$ | 9.000 | 8.000 | 0.889 | 72.000 | 61.333 |
| 11 | $\{1,3,7\}$ | 11.667 | 13.000 | 1.114 | 117.000 | 76.886 |
| 12 | $\{1,3,8\}$ | 9.000 | 10.000 | 1.111 | 90.000 | 76.667 |
| 13 | $\{1,3,9\}$ | 9.667 | 9.000 | 0.931 | 81.000 | 64.241 |
| 14 | $\{1,4,5\}$ | 9.667 | 11.667 | 1.207 | 105.000 | 83.276 |
| 15 | $\{1,4,6\}$ | 9.333 | 9.333 | 1.000 | 84.000 | 69.000 |
| 16 | $\{1,4,7\}$ | 12.000 | 14.333 | 1.194 | 129.000 | 82.417 |
| 17 | $\{1,4,8\}$ | 9.333 | 11.333 | 1.214 | 102.000 | 83.786 |
| 18 | $\{1,4,9\}$ | 10.000 | 10.333 | 1.033 | 93.000 | 71.300 |
| 19 | $\{1,5,6\}$ | 6.667 | 6.333 | 0.950 | 57.000 | 65.550 |
| 20 | $\{1,5,7\}$ | 9.333 | 11.333 | 1.214 | 102.000 | 83.786 |
| 21 | $\{1,5,8\}$ | 6.667 | 8.333 | 1.250 | 75.000 | 86.250 |
| 22 | $\{1,5,9\}$ | 7.333 | 7.333 | 1.000 | 66.000 | 69.000 |
| 23 | $\{1,6,7\}$ | 9.000 | 9.000 | 1.000 | 81.000 | 69.000 |
| 24 | $\{1,6,8\}$ | 6.333 | 6.000 | 0.947 | 54.000 | 65.368 |
| 25 | $\{1,6,9\}$ | 7.000 | 5.000 | 0.714 | 45.000 | 49.286 |
| 26 | $\{1,7,8\}$ | 9.000 | 11.000 | 1.222 | 99.000 | 84.333 |
| 27 | $\{1,7,9\}$ | 9.667 | 10.000 | 1.034 | 90.000 | 71.379 |
| 28 | $\{1,8,9\}$ | 7.000 | 7.000 | 1.000 | 63.000 | 69.000 |
| 29 | $\{2,3,4\}$ | 10.000 | 12.333 | 1.233 | 111.000 | 85.100 |
| 30 | $\{2,3,5\}$ | 7.333 | 9.333 | 1.273 | 84.000 | 87.818 |
| 31 | $\{2,3,6\}$ | 7.000 | 7.000 | 1.000 | 63.000 | 69.000 |
| 32 | $\{2,3,7\}$ | 9.667 | 12.000 | 1.241 | 108.000 | 85.655 |
| 33 | $\{2,3,8\}$ | 7.000 | 9.000 | 1.286 | 81.000 | 88.714 |
| 34 | $\{2,3,9\}$ | 7.667 | 8.000 | 1.043 | 72.000 | 72.000 |
| 35 | $\{2,4,5\}$ | 7.667 | 10.667 | 1.391 | 96.000 | 96.000 |
| 36 | $\{2,4,6\}$ | 7.333 | 8.333 | 1.136 | 75.000 | 78.409 |
| 37 | $\{2,4,7\}$ | 10.000 | 13.333 | 1.333 | 120.000 | 92.000 |
| 38 | $\{2,4,8\}$ | 7.333 | 10.333 | 1.409 | 93.000 | 97.227 |
| 39 | $\{2,4,9\}$ | 8.000 | 9.333 | 1.167 | 84.000 | 80.500 |
| 40 | $\{2,5,6\}$ | 4.667 | 5.333 | 1.143 | 48.000 | 78.857 |

| Sample Number | Sample $\mathcal{S}$ | $\bar{x}_{\mathcal{S}}$ | $\bar{y}_{\mathcal{S}}$ | $\hat{B}$ | $\hat{t}_{\text{SRS}}$ | $\hat{t}_{yr}$ |
|---|---|---|---|---|---|---|
| 41 | $\{2,5,7\}$ | 7.333 | 10.333 | 1.409 | 93.000 | 97.227 |
| 42 | $\{2,5,8\}$ | 4.667 | 7.333 | 1.571 | 66.000 | 108.429 |
| 43 | $\{2,5,9\}$ | 5.333 | 6.333 | 1.188 | 57.000 | 81.938 |
| 44 | $\{2,6,7\}$ | 7.000 | 8.000 | 1.143 | 72.000 | 78.857 |
| 45 | $\{2,6,8\}$ | 4.333 | 5.000 | 1.154 | 45.000 | 79.615 |
| 46 | $\{2,6,9\}$ | 5.000 | 4.000 | 0.800 | 36.000 | 55.200 |
| 47 | $\{2,7,8\}$ | 7.000 | 10.000 | 1.429 | 90.000 | 98.571 |
| 48 | $\{2,7,9\}$ | 7.667 | 9.000 | 1.174 | 81.000 | 81.000 |
| 49 | $\{2,8,9\}$ | 5.000 | 6.000 | 1.200 | 54.000 | 82.800 |
| 50 | $\{3,4,5\}$ | 9.000 | 12.667 | 1.407 | 114.000 | 97.111 |
| 51 | $\{3,4,6\}$ | 8.667 | 10.333 | 1.192 | 93.000 | 82.269 |
| 52 | $\{3,4,7\}$ | 11.333 | 15.333 | 1.353 | 138.000 | 93.353 |
| 53 | $\{3,4,8\}$ | 8.667 | 12.333 | 1.423 | 111.000 | 98.192 |
| 54 | $\{3,4,9\}$ | 9.333 | 11.333 | 1.214 | 102.000 | 83.786 |
| 55 | $\{3,5,6\}$ | 6.000 | 7.333 | 1.222 | 66.000 | 84.333 |
| 56 | $\{3,5,7\}$ | 8.667 | 12.333 | 1.423 | 111.000 | 98.192 |
| 57 | $\{3,5,8\}$ | 6.000 | 9.333 | 1.556 | 84.000 | 107.333 |
| 58 | $\{3,5,9\}$ | 6.667 | 8.333 | 1.250 | 75.000 | 86.250 |
| 59 | $\{3,6,7\}$ | 8.333 | 10.000 | 1.200 | 90.000 | 82.800 |
| 60 | $\{3,6,8\}$ | 5.667 | 7.000 | 1.235 | 63.000 | 85.235 |
| 61 | $\{3,6,9\}$ | 6.333 | 6.000 | 0.947 | 54.000 | 65.368 |
| 62 | $\{3,7,8\}$ | 8.333 | 12.000 | 1.440 | 108.000 | 99.360 |
| 63 | $\{3,7,9\}$ | 9.000 | 11.000 | 1.222 | 99.000 | 84.333 |
| 64 | $\{3,8,9\}$ | 6.333 | 8.000 | 1.263 | 72.000 | 87.158 |
| 65 | $\{4,5,6\}$ | 6.333 | 8.667 | 1.368 | 78.000 | 94.421 |
| 66 | $\{4,5,7\}$ | 9.000 | 13.667 | 1.519 | 123.000 | 104.778 |
| 67 | $\{4,5,8\}$ | 6.333 | 10.667 | 1.684 | 96.000 | 116.211 |
| 68 | $\{4,5,9\}$ | 7.000 | 9.667 | 1.381 | 87.000 | 95.286 |
| 69 | $\{4,6,7\}$ | 8.667 | 11.333 | 1.308 | 102.000 | 90.231 |
| 70 | $\{4,6,8\}$ | 6.000 | 8.333 | 1.389 | 75.000 | 95.833 |
| 71 | $\{4,6,9\}$ | 6.667 | 7.333 | 1.100 | 66.000 | 75.900 |
| 72 | $\{4,7,8\}$ | 8.667 | 13.333 | 1.538 | 120.000 | 106.154 |
| 73 | $\{4,7,9\}$ | 9.333 | 12.333 | 1.321 | 111.000 | 91.179 |
| 74 | $\{4,8,9\}$ | 6.667 | 9.333 | 1.400 | 84.000 | 96.600 |
| 75 | $\{5,6,7\}$ | 6.000 | 8.333 | 1.389 | 75.000 | 95.833 |
| 76 | $\{5,6,8\}$ | 3.333 | 5.333 | 1.600 | 48.000 | 110.400 |
| 77 | $\{5,6,9\}$ | 4.000 | 4.333 | 1.083 | 39.000 | 74.750 |
| 78 | $\{5,7,8\}$ | 6.000 | 10.333 | 1.722 | 93.000 | 118.833 |
| 79 | $\{5,7,9\}$ | 6.667 | 9.333 | 1.400 | 84.000 | 96.600 |
| 80 | $\{5,8,9\}$ | 4.000 | 6.333 | 1.583 | 57.000 | 109.250 |
| 81 | $\{6,7,8\}$ | 5.667 | 8.000 | 1.412 | 72.000 | 97.412 |
| 82 | $\{6,7,9\}$ | 6.333 | 7.000 | 1.105 | 63.000 | 76.263 |
| 83 | $\{6,8,9\}$ | 3.667 | 4.000 | 1.091 | 36.000 | 75.273 |
| 84 | $\{7,8,9\}$ | 6.333 | 9.000 | 1.421 | 81.000 | 98.053 |
| Average | | 7.667 | 9.222 | 1.214 | 83.000 | 83.733 |
| Variance | | 3.767 | 6.397 | 0.044 | 518.169 | 208.083 |

(c)



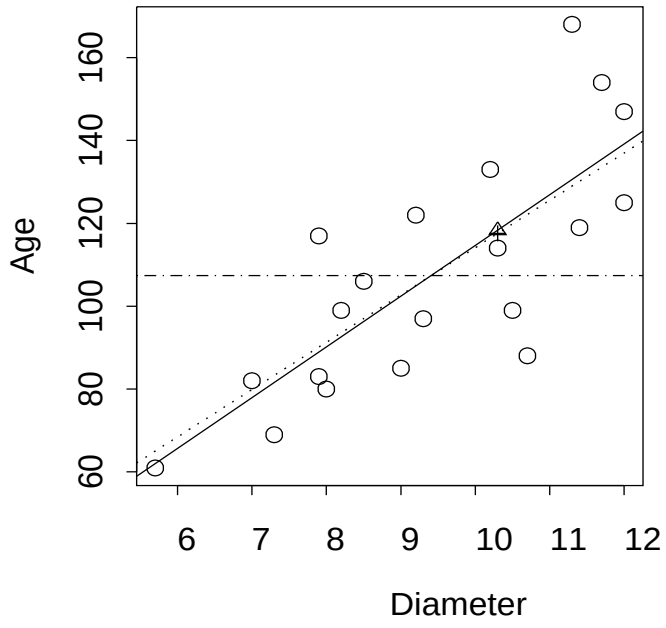

The shapes are actually quite similar; however, it appears that the histogram of the ratio estimator is a little less spread out.

(d) The mean of the sampling distribution of $\hat{t}_{yr}$ is 83.733; the variance is 208.083 and the bias is $83.733 - 83 = 0.733$. By contrast, the mean of the sampling distribution of $N\bar{y}$ is 83 and its variance is 518.169.

(e) From (4.6),

$$\text{Bias}\,(\hat{\bar{y}}_r) = 0.07073094.$$

**4.3** (a) The solid line is from regression estimation; the dashed line from ratio estimation; the dashed/dotted line has equation $y = 107.4$.




(b, c)

| Method | $\hat{\bar{y}}$ | SE $(\hat{\bar{y}})$ |
|---|---|---|
| SRS, $\bar{y}$ | 107.4 | 6.35 |
| Ratio | 117.6 | 4.35 |
| Regression | 118.4 | 3.96 |

For ratio estimation, $\hat{B} = 11.41946$; for regression estimation, $\hat{B}_0 = -7.808$ and $\hat{B}_1 = 12.250$. Note that the sample correlation of age and diameter is 0.78, so we would expect both ratio and regression estimation to improve precision.

To calculate $\hat{V}(\hat{\bar{y}}_r)$ using (4.9), note that $s_e^2 = 321.933$ so that

$$\hat{V}(\hat{\bar{y}}_r) = \left(1 - \frac{20}{1132}\right)\left(\frac{10.3}{9.405}\right)^2 \frac{321.933}{20} = 18.96$$

and $\mathrm{SE}(\hat{\bar{y}}) = 4.35$. For the regression estimator, we have $s_e^2 = 319.6$, so

$$\hat{V}(\hat{\bar{y}}) = \left(1 - \frac{20}{1132}\right)\frac{319.6}{20} = 15.7.$$

From a design-based perspective, it makes little difference which estimator is used. From a model-based perspective, though, a plot of residuals vs. predicted values exhibits a "funnel shape" indicating that the variability increases with $x$. Thus a model-based analysis for these data should incorporate the unequal variances. From that perspective, the ratio model might be more appropriate.

Note that the variances calculated using SAS PROC SURVEYREG are larger since they use $\hat{V}_2$ from Section 11.7.

Here is code for SAS:

```
 data trees;
    input treenum diam age @@;
sampwt = 1132/20;
datalines;
     1      12.0    125   11       5.7    61
     2      11.4    119   12       8.0    80
     3      7.9    83     13     10.3    114
     4      9.0    85     14     12.0    147
     5      10.5   99     15     9.2    122
     6      7.9    117    16     8.5    106
     7      7.3    69     17     7.0    82
     8      10.2   133    18     10.7   88
     9      11.7   154    19     9.3    97
    10      11.3   168    20     8.2    99
;
proc print data=trees;
run;

/* proc surveymeans will estimate ratios with keyword 'ratio'  */

proc surveymeans data=trees total=1132 mean stderr clm
                                        sum clsum ratio ;
   var diam age;  /* need both in var statement */
   ratio 'age/diameter' age/diam;
   weight sampwt;
   ods output Statistics=statsout Ratio=ratioout;
   run;

/* Can get ratio estimates of totals by taking output from
   proc surveymeans and multiplying by N */
```

```
data ratioout1;
   set ratioout;
   xmean = 10.3;
   ratiomean = ratio*xmean;
   semean = stderr*xmean;
   lowercls = lowercl*xmean;
   uppercls = uppercl*xmean;

proc print data = ratioout1;
run;

/* Can also calculate ratio estimate by hand */

data treesresid;
   set trees;
   resid = age -  11.419458*diam;
   resid2 = resid*(10.3/  9.405);

proc univariate data=treesresid;
run;

proc surveyreg data=trees total=1132;
   model age = diam / clparm solution ;
        /* fits the regression model */
   weight sampwt;
   estimate 'Mean age of trees' intercept 1 diam 10.3;
run;

proc gplot data=trees;
   plot age*diam;
run;

data trees2;
   set trees;
   resid = age - (-7.8080877 + 12.2496636*diam);
proc surveymeans data=trees2 total =1132;
   weight sampwt;
   var resid age diam;
run;
```

The output from proc surveyreg gives a larger standard error for the regression estimator:

```
                Analysis of Estimable Functions
```

```
                                        Standard
Parameter                 Estimate         Error    t Value    Pr > |t|

Mean age of trees     118.363449    5.20417420      22.74      <.0001
```

**4.5** There are 85 18-hole courses in the sample. For these 85 courses, the sample mean weekend greens fee is

$$\bar{y}_d = 34.829$$

and the sample variance is

$$s_d^2 = 395.498.$$

Using results from Section 4.3,

$$\text{SE}[\bar{y}_d] = \sqrt{\frac{395.498}{85}} = 2.16.$$

Here is SAS code:

```
filename golfsrs
data golfsrs;
   infile golfsrs delimiter="," dsd firstobs=2;
     /* The dsd option allows SAS to read the missing values between
        successive delimiters */
   input RN  state $ holes  type $ yearblt wkday18 wkday9 wkend18
      wkend9  backtee  rating  par  cart18  cart9  caddy $ pro $;
   sampwt = 14938/120;
   if holes = 18 then holes18 = 1;
   else holes18=0;

proc surveymeans data=golfsrs total = 14938;
   weight sampwt;
   var wkend18;
   domain holes18;
run;
```

```
            Data Summary

Number of Observations          120
Sum of Weights                14938


                        Statistics
```

| Variable | N | Mean | Std Error of Mean |
|---|---|---|---|
| wkend18 | 85 | 34.828824 | 2.148380 |

Statistics

| Variable | 95% CL for Mean | |
|---|---|---|
| wkend18 | 30.5565341 | 39.1011129 |

Domain Analysis: holes18

| holes18 | Variable | N | Mean | Std Error of Mean |
|---|---|---|---|---|
| 0 | wkend18 | 0 | . | . |
| 1 | wkend18 | 85 | 34.828824 | 2.144660 |

Domain Analysis: holes18

| holes18 | Variable | 95% CL for Mean | |
|---|---|---|---|
| 0 | wkend18 | . | . |
| 1 | wkend18 | 30.5639320 | 39.0937150 |

**4.6** As you can see from the plot of weekend greens fee vs. back-tee yardage, this is not a "classical" straight-line relationship. The variability in weekend greens fee appears to increase with the back-tee yardage. Nevertheless, we can estimate the slope and intercept, with

$$\hat{y} = -37.26 + 0.0113x.$$

(We'll discuss standard errors in Chapter 11.) For estimating the ratio, we have

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{34.83}{6392.29} = 0.00545.$$

Using (4.10), with $s_e^2$ the sample variance of the residuals,

$$\begin{aligned} \hat{V}(\hat{B}) &\approx \frac{s_e^2}{(85)(6292.29)^2} = \frac{362.578}{(85)(6392.29)^2} = 1.044 \times 10^{-7} \\ SE(\hat{B}) &= .00032. \end{aligned}$$

**4.7** (a) 88 courses have a golf professional. For these 88 courses, $\bar{y}_{d_1} = 23.5983$ and

$s^2_{d_1} = 387.7194$, so

$$\hat{V}(\bar{y}_{d_1}) = \frac{387.7194}{88} = 4.4059.$$

(b) For the 32 courses without a golf professional, $\bar{y}_{d2} = 10.6797$ and $s^2_{d2} = 19.146$, so

$$\hat{V}(\bar{y}_{d2}) = \frac{19.146}{32} = 0.5983.$$

**4.8** (a) (b) $\hat{B} = \bar{y}/\bar{x} = 297897/647.7467 = 459.8975$. Thus

$$\hat{t}_{yr} = t_x\hat{B} = (2087759)(459.8975) = 960{,}155{,}061.$$

The estimated variance of the residuals about the line $y = \hat{B}x$ is

$$s^2_e = 149{,}902{,}393{,}481.$$

Using (4.11), then, with farms87 as the auxiliary variable,

$$\text{SE}\,[\hat{t}_{yr}] = 3078\sqrt{1 - \frac{300}{3078}}\,\sqrt{\frac{s^2_e}{300}} = 65{,}364{,}822.$$

(c) The least squares regression equation is

$$\hat{y} = 267029.8 + 47.65325x$$

Then

$$\hat{\bar{y}}_{\text{reg}} = 267029.8 + 47.65325(647.7467) = 297897.04$$

and

$$\hat{t}_{y\text{reg}} = 3078\hat{\bar{y}}_{\text{reg}} = 916{,}927{,}075.$$

The estimated variance of the residuals from the regression is $s^2_e = 118{,}293{,}647{,}832$, which implies from (4.19) that

$$\text{SE}\,[\hat{t}_{y\text{reg}}] = 3078\sqrt{1 - \frac{300}{3078}}\,\sqrt{\frac{s^2_e}{300}} = 58{,}065{,}813.$$

(d) Clearly, for this response, it is better to use acres87 as an auxiliary variable. The correlation of farms87 with acres92 is only 0.06; using farms87 as an auxiliary variable does not improve on the SRS estimate $N\bar{y}$. The correlation of acres92 and acres87, however, exceeds 0.99. Here are the various estimates for the population total of acres92:

| Estimate | $\hat{t}$ | SE $[\hat{t}]$ |
|---|---|---|
| SRS, $N\bar{y}$ | 916,927,110 | 58,169,381 |
| Ratio, $x$ = acres87 | 951,513,191 | 5,344,568 |
| Ratio, $x$ = farms87 | 960,155,061 | 65,364,822 |
| Regression, $x$ = farms87 | 916,927,075 | 58,065,813 |

**Moral:** Ratio estimation can lead to greatly increased precision, but should not be used blindly. In this case, ratio estimation with auxiliary variable farms87 had larger standard error than if no auxiliary information were used at all. The regression estimate of $t$ is similar to $N\bar{y}$, because the regression slope is small relative to the magnitude of the data. The regression slope is not significantly different from 0; as can be seen from the picture in (a), the straight-line regression model does not describe the counties with few but large farms.

**4.9** We use results from Section 4.2. (a) Let $y_i$ = acres92 for county $i$, and $x_i$ = farms92 for county $i$. Define Then

$$\hat{t}_{y1} = N\bar{u} = 3078(161773.8) = 497{,}939{,}808$$

and

$$\text{SE}\,[\hat{t}_{y1}] = N\sqrt{1 - \frac{300}{3078}}\,\sqrt{\frac{s_u^2}{300}} = 3078\sqrt{1 - \frac{300}{3078}}\,\sqrt{\frac{109{,}710{,}284{,}064}{300}} = 55{,}919{,}525.$$

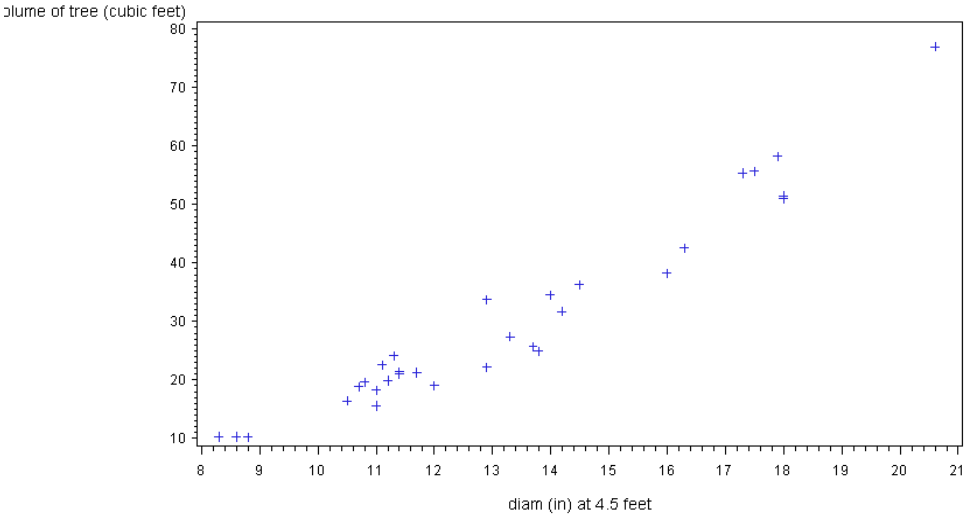
(b) Now

$$\hat{t}_{y2} = 3078(136123.2) = 418,987,302$$

and

$$\text{SE}\,[\hat{t}_{y2}] = 3078\sqrt{1 - \frac{300}{3078}}\,\sqrt{\frac{53{,}195{,}371{,}851}{300}} = 38{,}938{,}277.$$

**4.10** (a)




(b) Here is code and output from SAS:

```
filename cherries 'cherry.csv';
```

```
data cherry;
  infile cherries delimiter=',' firstobs=2;
  input diam height vol;
  sampwt = 2967/31;
  obsnum = _n_;
  label diam      = 'diam (in) at 4.5 feet'
        height    = 'height of tree (feet)'
        vol       = 'volume of tree (cubic feet)'
        sampwt    = 'sampling weight'
;
 /* Plot and print the data set */

proc print data = cherry;
   var diam height vol sampwt;

proc gplot data = cherry;
   plot vol *diam ;

proc surveymeans data = cherry total=2967 mean clm sum clsum ratio ;
   weight sampwt;
   var diam vol;
   ratio 'vol/diam' vol/diam;
   ods output Statistics=statsout Ratio=ratioout;
run;

data ratioout1;
   set ratioout;
   xtotal = 41835;
   ratiosum = ratio*xtotal;
   sesum = stderr*xtotal;
   lowercls = lowercl*xtotal;
   uppercls = uppercl*xtotal;

proc print data = ratioout1;
run;
```

Using this code, we obtain $\hat{t}_{yr} = 95272.16$ with 95% CI of [84098, 106,446].

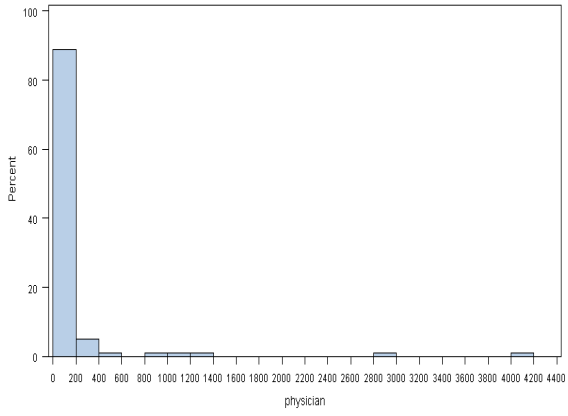
(c) SAS code and output follow:

```
proc surveyreg data=cherry total=100;
   model vol=diam / clparm solution;
   weight sampwt;
   estimate 'Total volume' intercept 2967 diam 41835;
```

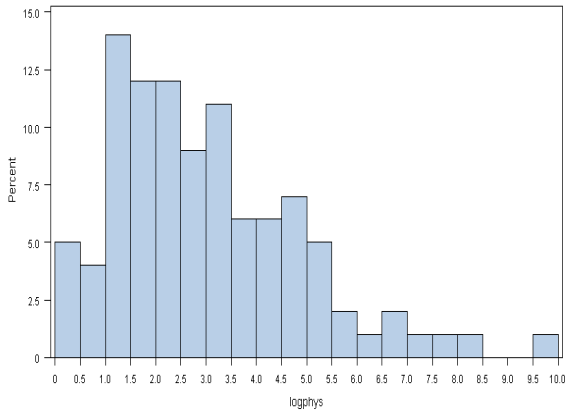
```
          /* substitute N for intercept, t_x for diam */
run;

                Analysis of Estimable Functions

                                    Standard
Parameter          Estimate          Error    t Value    Pr > |t|

Total volume     102318.860     2233.70776      45.81      <.0001

   Analysis of Estimable Functions

                     95% Confidence
Parameter               Interval

Total volume     97757.0204 106880.700
```

Note that the estimate from regression estimation is quite a bit higher than the estimate from ratio estimation. In addition, the CI for regression estimation is narrower than the CIs for $\hat{t}_{yr}$ or $N\bar{y}$. This is because the regression model is a better fit to the data than the ratio model.

**4.11** (a) The variable *number of physicians* has a skewed distribution. The first histogram excludes Cook County, Illinois (with $y_i = 15{,}153$) for slightly better visibility.




The next histogram, of all 100 counties, depicts the logarithm of (number of physicians + 1) (which is still skewed).




(b) $\bar{y} = 297.17$, $s_y^2$ =2,534,052.

Thus the estimated total number of physicians is

$$N\bar{y} = (3141)(297.17) = 933{,}411$$

with

$$\begin{aligned} \mathrm{SE}\,(N\bar{y}) &= N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n}} \\ &= 3141\sqrt{\left(1 - \frac{100}{3141}\right)\frac{2{,}534{,}052}{100}} \\ &= 3141\sqrt{24533.75} \\ &= 491{,}983. \end{aligned}$$

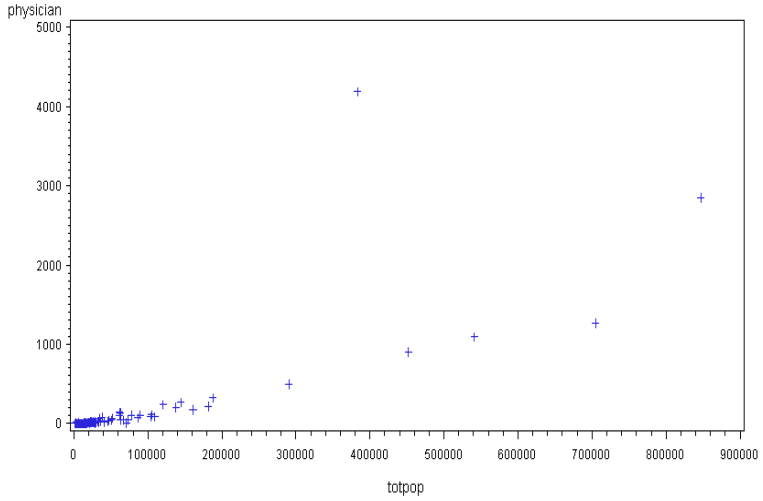
The standard error is large compared with the estimated total number of physicians.

The extreme skewness of the data makes us suspect that $N\bar{y}$ does not follow an approximate normal distribution, and that a confidence interval of the form $N\bar{y} \pm 1.96\ \mathrm{SE}\,(N\bar{y})$ would not have 95% coverage in practice. In fact, when we substitute sample quantities into (2.23), we obtain

$$n_{\min} = 28 + 25\,(6.04)^2 = 940$$

as the required minimum sample size for $\bar{y}$ to approximately follow a normal distribution.

(c) Again, we omit Cook County.




There appears to be increasing variance as *population* increases, so ratio estimation may be more appropriate.

(d) Ratio estimation:

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{297.17}{118531.2} = 0.002507.$$

$$\hat{t}_{yr} = \hat{B}t_x = (.002507)(255{,}077{,}536) = 639{,}506.$$

Using (4.11),

$$\begin{aligned} \text{SE}\,(\hat{t}_{yr}) &= N\sqrt{\left(1-\frac{n}{N}\right)\frac{s_e^2}{n}} \\ &= 3141\sqrt{\left(1-\frac{100}{3141}\right)\frac{(255077536)^2}{(372306374)^2}\frac{172268}{100}} \\ &= 87885. \end{aligned}$$

Regression estimation:

$$\hat{B}_0 = -54.23 \qquad \hat{B}_1 = 0.00296.$$

From (4.15) and (4.19),

$$\begin{aligned} \hat{\bar{y}}_{\text{reg}} &= \hat{B}_0 + \hat{B}_1\bar{x}_U \\ &= -54.23 + (0.00296)\left(\frac{255{,}077{,}536}{3141}\right) \\ &= 186.52 \end{aligned}$$

and

$$\begin{aligned} \mathrm{SE}\left(\hat{\bar{y}}_{\mathrm{reg}}\right) &= \sqrt{\left(1-\frac{n}{N}\right)\frac{s_e^2}{n}} \\ &= \sqrt{\left(1-\frac{100}{3141}\right)\frac{114644.1}{100}} \\ &= 33.316. \end{aligned}$$

Consequently,

$$\hat{t}_{y\mathrm{reg}} = N\hat{\bar{y}}_{\mathrm{reg}} = 3141(186.52) = 585871$$

and

$$\mathrm{SE}\left(\hat{t}_{y\mathrm{reg}}\right) = N\ \mathrm{SE}\left(\hat{\bar{y}}_{\mathrm{reg}}\right) = 3141(33.316) = 104645.$$

The standard error from proc surveyreg is smaller and equals 92535.

(e) Ratio estimation and regression estimation both lead to a smaller standard error, and an estimate that is closer to the true value.

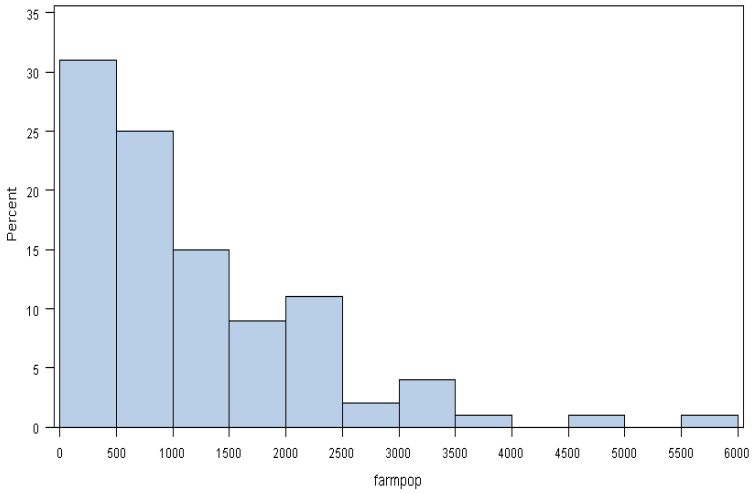
Here is SAS code for performing these analyses:

```
data counties;
    infile counties firstobs=2 delimiter=",";
    input  RN State County landarea totpop physician enroll
      percpub civlabor unemp farmpop numfarm farmacre fedgrant
      fedciv milit veterans percviet ;
   sampwt = 3141/100;
   logphys = log(physician+1);

/* The following histogram is really really skewed
    because of Cook County */
proc univariate data=counties ;
  var physician;
  histogram / endpoints = 0 to 16000 by 500;
run;

data countnoCook;
   set counties;
   if physician gt 10000 then delete;

proc univariate data=countnoCook;
  var physician ;
  histogram / endpoints = 0 to 4400 by 200;
run;

proc univariate data=counties;
  var logphys ;
```

```
  histogram / endpoints = 0 to 10 by .5;
run;

proc surveymeans data=counties total = 3141 mean clm sum clsum;
  weight sampwt;
  var physician;
run;

proc gplot data=countnoCook;
   plot physician *  totpop;
run;

proc surveymeans data=counties total=3141 mean clm sum clsum ratio;
   var physician totpop;  /* need both in var statement */
   ratio 'physician/totpop' physician/totpop;
   weight sampwt;
   ods output Statistics=statsout Ratio=ratioout;
   run;

/* Can get ratio estimates of totals by taking output from
   proc surveymeans and multiplying by t_x */

data ratioout1;
   set ratioout;
   xtot = 255077536;
   ratiotot = ratio*xtot;
   setot = stderr*xtot;
   lowercls = lowercl*xtot;
   uppercls = uppercl*xtot;

proc print data = ratioout1;
run;

/* Can also calculate ratio estimate by hand */

data resid;
   set counties;
   resid = physician -0.002507*totpop;
   resid2 = resid*(255077536/   372306374);
       /* Use g-weights in SE formula*/

proc surveymeans data=resid total=3141 mean clm sum clsum;
   weight sampwt;
   var resid resid2;
run;
```

```
proc surveyreg data=counties total=3141;
   model physician=totpop / clparm solution ;
       /* fits the regression model */
   weight sampwt;
   estimate 'Total number of physicians' intercept 3141
                                          totpop 255077536;
   estimate 'Average number of physicians' intercept 3141
                                       totpop 255077536/divisor = 3141;
run;
```

**4.12** (a) The distribution appears to be skewed, but not quite as skewed as the distribution in Exercise 4.11.




(b) $\bar{y} = 1146.87$, $s_y^2$ =1,138,630

Thus the estimated total farm population, using $N\bar{y}$, is

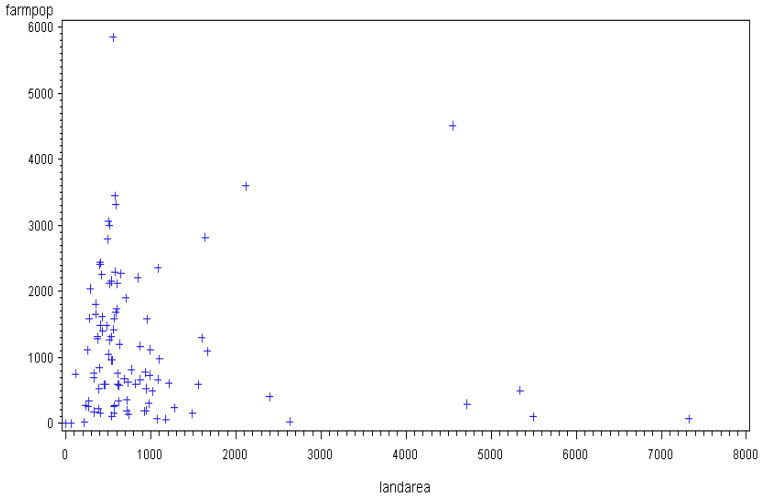
$$N_{\bar{y}} = (3141)(1146.87) = 3{,}602{,}319$$

with

$$\text{SE}\,(N\bar{y}) = 3141\sqrt{\left(1 - \frac{100}{3141}\right)\frac{1{,}138{,}630}{100}} = 329{,}787.$$

SAS proc surveymeans gives the same result.

(c) Note that corr(farmpop,landarea) $= -0.058$. We would not expect ratio or regression estimation to do well here.




(d) Ratio estimation:

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{1146.87}{944.92} = 1.21372.$$

$$s_e^2 = 3{,}297{,}971$$

$$\hat{t}_{yr} = \hat{B}_{t_x} = (1.21372)(3{,}536{,}278) = 4{,}292{,}058$$

$$\mathrm{SE}\,(\hat{t}_{yr}) = 3141\sqrt{\left(1 - \frac{100}{3141}\right)\frac{(3{,}536{,}278)^2}{(2967994)^2}\frac{3{,}297{,}971}{100}} = 668\,727.$$

Note that the SE for the ratio estimate is higher than that for $N\bar{y}$.

Regression estimation:

$$\hat{B}_0 = 1197.35, \quad \hat{B}_1 = -.05342218, \quad s_e^2 = 1{,}134{,}785.$$

From (4.15) and (4.19),

$$\begin{aligned}\hat{\bar{y}}_{\mathrm{reg}} &= \hat{B}_0 + \hat{B}_1\bar{x}_U \\ &= 1197.35 - 0.0534(3{,}536{,}278) \\ &= 1137.2\end{aligned}$$

$$\mathrm{SE}\,(\hat{\bar{y}}_{\mathrm{reg}}) = \sqrt{\left(1 - \frac{100}{3141}\right)\frac{1{,}134{,}785}{100}} = 104.82.$$

Consequently,

$$\hat{t}_{y\mathrm{reg}} = N\hat{\bar{y}}_{\mathrm{reg}} = 3141(1137.2) = 3{,}571{,}960$$

$$\mathrm{SE}\,(\hat{t}_{y\mathrm{reg}}) = N\ \mathrm{SE}\,(\hat{\bar{y}}_{\mathrm{reg}}) = 3141(104.82) = 329{,}230.$$

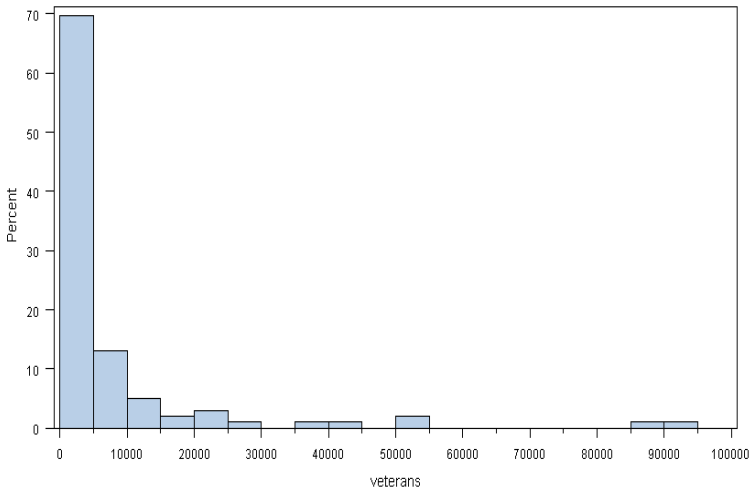
SAS proc surveyreg gives SE 350595.

(e) The "true" value is $t_y = 3{,}871{,}583$.

Here is SAS code that may be used to compute these estimates. See Exercise 4.11 solution for reading in the data.

```
proc univariate data=counties ;
  var farmpop;
  histogram  / endpoints = 0 to 6000 by 500;
run;

proc gplot data=counties;
   plot farmpop *  landarea;
run;

proc surveymeans data=counties total=3141 mean clm sum clsum ratio ;
   var farmpop landarea;  /* need both in var statement */
   ratio 'farmpop/landarea' farmpop/landarea;
   weight sampwt;
   ods output Statistics=statsout Ratio=ratioout;

data ratioout1;
   set ratioout;
   xtot = 3536278;
   ratiotot = ratio*xtot;
   setot = stderr*xtot;
   lowercls = lowercl*xtot;
   uppercls = uppercl*xtot;

proc print data = ratioout1;
run;

proc surveyreg data=counties total=3141;
   model farmpop=landarea / clparm solution ;
   weight sampwt;
   estimate 'Total farmpop' intercept 3141 landarea 3536278;
   estimate 'Average farmpop' intercept 3141
                                  landarea 3536278/divisor = 3141;
run;
```

**4.13** (a) As in Exercise 4.11, we omit Cook County, Illinois, from the histogram so we can see the other data points better. Cook County has 457,880 veterans. The distribution is very skewed; Cook County is an extreme outlier.

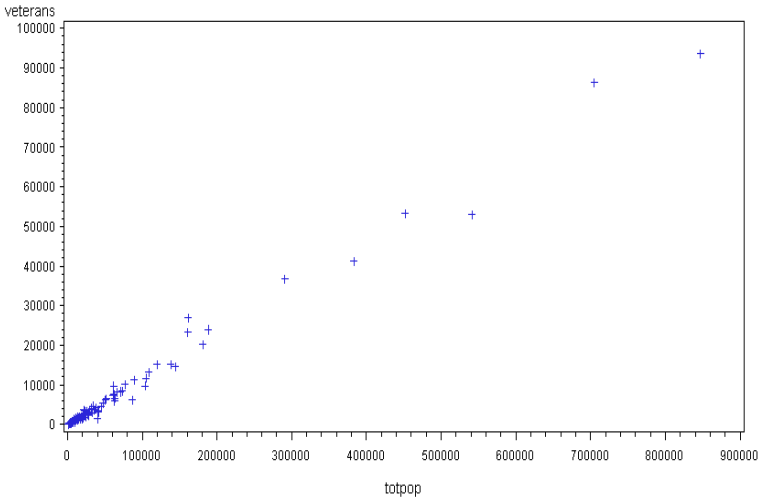



(b) $\bar{y} = 12249.71$, $s_y^2 = 2{,}263{,}371{,}150$.

$$N\bar{y} = (3141)(12249.71) = 38{,}476{,}339$$

$$\text{SE}\,(N\bar{y}) = 3141\sqrt{\left(1 - \frac{100}{3141}\right)\frac{2{,}263{,}371{,}150}{100}} = 14{,}703{,}478.$$

(c) Again, Cook County is omitted from this plot. These data appear very close to a straight line. We would expect ratio or regression estimation to help immensely.




(d) Ratio estimation:

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{12249.71}{118531.2} = 0.1033$$

$$\hat{t}_{yr} = \hat{B}t_x = (0.1033)(255{,}077{,}536) = 26{,}361{,}219.$$

$$s_e^2 = \frac{1}{99}\sum_{i\in\S}(y_i - \hat{B}x_i)^2 = 59{,}771{,}493$$

$$\text{SE}\,(\hat{t}_{yr}) = 3141\sqrt{\left(1 - \frac{100}{3141}\right)\frac{(255077536)^2}{(372306374)^2}\frac{59{,}771{,}493}{100}} = 1637046.$$

Regression estimation:

$$\hat{B}_0 = 1534.201, \quad \hat{B}_1 = 0.0904, \quad s_e^2 = 13{,}653{,}852$$

$$\begin{aligned}\hat{\bar{y}}_{\text{reg}} &= \hat{B}_0 + \hat{B}_1\bar{x}_U \\ &= (1534.2) + (0.0904)\left(\tfrac{255{,}077{,}536}{3141}\right) \\ &= 8875.7.\end{aligned}$$

$$\text{SE}\,(\hat{\bar{y}}_{\text{reg}}) = \sqrt{\left(1 - \frac{100}{3141}\right)\frac{13{,}653{,}852}{100}} = 363.58$$

$$\hat{t}_{y\text{reg}} = N\hat{\bar{y}}_{\text{reg}} = 27{,}878{,}564$$

$$\text{SE}\,(\hat{t}_{y\text{reg}}) = N\text{SE}\,(\hat{\bar{y}}_{\text{reg}}) = 1{,}142{,}010.$$

The SE from proc surveyreg is 1063699.

Here is SAS code. The data are read in as in Exercise 4.11.

```
/* The following histogram is skewed because of Cook County */
proc univariate data=counties ;
  var veterans;
  histogram / endpoints = 0 to 440000 by 10000;
run;

data countnoCook;
   set counties;
   if veterans gt 400000 then delete;

proc univariate data=countnoCook;
  var veterans ;
  histogram / endpoints = 0 to 100000 by 5000;
run;

proc gplot data=countnoCook;
   plot veterans *  totpop;
run;

proc surveymeans data=counties total=3141 mean clm sum clsum ratio;
```

```
   var veterans totpop;  /* need both in var statement */
   ratio 'veterans/totpop' veterans/totpop;
   weight sampwt;
   ods output Statistics=statsout Ratio=ratioout;
   run;

data ratioout1;
   set ratioout;
   xtot = 255077536;
   ratiotot = ratio*xtot;
   setot = stderr*xtot;
   lowercls = lowercl*xtot;
   uppercls = uppercl*xtot;

proc print data = ratioout1;
run;

/* Can also calculate ratio estimate by hand */

data resid;
   set counties;
   resid = veterans -0.002507*totpop;
   resid2 = resid*(255077536/   372306374);
          /* Use g-weights in SE formula*/

proc surveymeans data=resid total=3141 mean clm sum clsum;
   weight sampwt;
   var resid resid2;
run;

proc surveyreg data=counties total=3141;
   model veterans=totpop / clparm solution ;
         /* fits the regression model */
   weight sampwt;
   estimate 'Total number of veterans' intercept 3141
                                totpop 255077536;
   estimate 'Average number of veterans'
       intercept 3141 totpop 255077536/divisor = 3141;
run;
```

(e) Here, $t_y = 27{,}481{,}055$.

**4.15** (a) A 95% CI for the average concentration of *lead* is

$$127 \pm 1.96\frac{146}{\sqrt{121}} = [101.0, 153.0].$$

For *copper*, the corresponding interval is

$$35 \pm 1.96\frac{16}{\sqrt{121}} = [32.1, 37.85].$$

Note that we do not use an fpc here. Soil samples are collected at grid intersections, and we may assume that the amount of soil in the sample is negligible compared with that in the region.

(b) Because the samples are systematically taken on grid points, we know that $(N_h/N) = (n_h/n)$. Using (4.21),

$$\begin{aligned} \text{LEAD} : \bar{y}_{\text{post}} &= \frac{82}{121}71 + \frac{31}{121}259 + \frac{8}{121}189 = 127 \\ \text{COPPER} : \bar{y}_{\text{post}} &= \frac{82}{121}28 + \frac{31}{121}50 + \frac{8}{121}45 = 35. \end{aligned}$$

Not surprisingly, these are the same numbers from the first table. The variances and confidence intervals, however, differ. From (4.22),

$$\begin{aligned} \text{LEAD} : \hat{V}(\bar{y}_{\text{post}}) &= \left(\frac{82}{121}\right)\frac{28^2}{121} + \left(\frac{31}{121}\right)\frac{232^2}{121} + \left(\frac{8}{121}\right)\frac{79^2}{121} = 121.8 \\ \text{COPPER} : \hat{V}(\bar{y}_{\text{post}}) &= \left(\frac{82}{121}\right)\frac{9^2}{121} + \left(\frac{31}{121}\right)\frac{18^2}{121} + \left(\frac{8}{121}\right)\frac{15^2}{121} = 1.26. \end{aligned}$$

The corresponding 95% confidence intervals are

$$\begin{aligned} \text{LEAD} &: \quad 127 \pm 1.96\sqrt{121.8} = [105.4, 148.6] \\ \text{COPPER} &: \quad 35 \pm 1.96\sqrt{1.26} = [32.8, 37.2]. \end{aligned}$$

These confidence intervals are both smaller than the CIs in part (a); this indicates that stratified sampling would increase precision in future surveys.

The following table gives the estimated coefficients of variation for the SRS and poststratified estimates:

| | $\widehat{\text{CV}}(\bar{y})$ | |
|---|---|---|
| | SRS | Poststratified |
| Lead | 0.1045 | 0.0869 |
| Copper | 0.0416 | 0.0321 |

**4.18** As $d_i = y_i - Bx_i$, $\bar{d} = \bar{y} - B\bar{x}$. Then, using (A.10),

$$\begin{aligned} V(\bar{d}) &= V(\bar{y} - B\bar{x}) \\ &= V(\bar{y}) - 2\,\text{Cov}\,(\bar{y}, B\bar{x}) + B^2V(\bar{x}) \\ &= \left(1 - \frac{n}{N}\right)\left[\frac{S_y^2}{n} - 2B\frac{RS_xS_y}{n} + B^2\frac{S_x^2}{n}\right]. \end{aligned}$$

**4.21** From (4.6), the squared bias of $\hat{B}$ is approximately

$$\frac{1}{\overline{x}_U^4}\left(1 - \frac{n}{N}\right)^2\frac{1}{n^2}[BS_x^2 - RS_xS_y]^2$$

From (4.8),

$$E[(\hat{B}-B)^2] \approx \left(1-\frac{n}{N}\right)\frac{1}{n\bar{x}_U^2}(S_y^2 - 2BRS_xS_y + B^2S_x^2).$$

The approximate MSE is thus of order $1/n$, while the squared bias is of order $1/n^2$. Consequently, $\text{MSE}\,(\hat{B}) \approx V(\hat{B})$.

**4.22** A rigorous proof, showing that the lower order terms are negligible, is beyond the scope of this book. We give an argument for (4.6).

$$\begin{aligned}
E[\hat{B}-B] &= E\left[\frac{\bar{y}}{\bar{x}} - \frac{\bar{y}_U}{\bar{x}_U}\right] \\
&= E\left[\frac{\bar{y}}{\bar{x}_U}\left(\frac{\bar{x}_U}{\bar{x}}\right) - \frac{\bar{y}_U}{\bar{x}_U}\right] \\
&= E\left[\frac{\bar{y}}{\bar{x}_U}\left(1-\frac{\bar{x}-\bar{x}_U}{\bar{x}}\right) - \frac{\bar{y}_U}{\bar{x}_U}\right] \\
&= -E\left[\frac{\bar{y}(\bar{x}-\bar{x}_U)}{\bar{x}_U\bar{x}}\right] \\
&= E\left[\frac{\bar{y}(\bar{x}-\bar{x}_U)}{\bar{x}_U^2}\left(\frac{\bar{x}-\bar{x}_U}{\bar{x}} - 1\right)\right] \\
&= \frac{1}{\bar{x}_U^2}\{BV(\bar{x}) - \text{Cov}\,(\bar{x},\bar{y}) + E[(\hat{B}-B)(\bar{x}-\bar{x}_U)^2]\} \\
&\approx \frac{1}{\bar{x}_U^2}[BV(\bar{x}) - \text{Cov}\,(\bar{x},\bar{y})].
\end{aligned}$$

**4.24** (a) From (4.5),

$$\bar{y}_1 - \bar{y}_{U1} = \frac{1}{\bar{x}_U}\left(\bar{y} - \frac{t_u}{t_x}\bar{x}\right)\left(1-\frac{\bar{x}-\bar{x}_U}{\bar{x}}\right)$$

and

$$\bar{y}_2 - \bar{y}_{U2} = \frac{1}{1-\bar{x}_U}\left[\bar{y} - \bar{u} - \frac{t_y - t_u}{N-t_x}(1-\bar{x})\right]\left(1+\frac{\bar{x}-\bar{x}_U}{\bar{x}}\right).$$

The covariance follows because the expected value of terms involving $\bar{x}-\bar{x}_U$ are small compared with the other terms.

(b) Note that because $x_i$ takes on only values 0 and 1, $x_i^2 = x_i$, $x_iu_i = u_i$, and

$u_i y_i = u_i^2$. Now look at the numerator of (4.26).

$$\begin{aligned}
&\sum_{i=1}^{N}\left[u_i-\bar{u}_U-\frac{t_u}{t_x}(x_i-\bar{x}_U)\right]\left[y_i-\bar{y}_U-u_i+\bar{u}_U+\frac{t_y-t_u}{N-t_x}(x_i-\bar{x}_U)\right]\\
&= \sum_{i=1}^{N}\left[u_i-\bar{u}_U-\frac{t_u}{t_x}(x_i-\bar{x}_U)\right]\left[y_i-u_i+\frac{t_y-t_u}{N-t_x}x_i\right]\\
&= \sum_{i=1}^{N}\left[u_iy_i-u_i^2+\frac{t_y-t_u}{N-t_x}u_ix_i-\frac{t_u}{t_x}x_iy_i+\frac{t_u}{t_x}u_ix_i-\frac{t_u}{t_x}\frac{t_y-t_u}{N-t_x}x_i^2\right]\\
&\quad+\left(\frac{t_u}{t_x}\bar{x}_U-\bar{u}_U\right)\left(t_y-t_u+\frac{t_y-t_u}{N-t_x}t_x\right)\\
&= \sum_{i=1}^{N}\left[\frac{t_y-t_u}{N-t_x}u_i-\frac{t_u}{t_x}u_i+\frac{t_u}{t_x}u_i-\frac{t_u}{t_x}\frac{t_y-t_u}{N-t_x}x_i\right]+0\\
&= \frac{t_y-t_u}{N-t_x}t_u-\frac{t_u}{t_x}\frac{t_y-t_u}{N-t_x}t_x\\
&= 0.
\end{aligned}$$

Consequently,

$$\text{Cov}\left[\left(\bar{u}-\frac{t_u}{t_x}\bar{x}\right),\ \left\{\bar{y}-\bar{u}-\frac{t_y-t_u}{N-t_x}(1-\bar{x})\right\}\right]=0.$$

**4.25** We use the multivariate delta method (see for example Lehmann, 1999, p. 295). Let

$$g(a,b)=\bar{x}_U\frac{b}{a}$$

so that $g(\bar{x},\bar{y})=\hat{\bar{y}}_r$ and $g(\bar{x}_U,\bar{y}_U)=\bar{y}_U$. Then,

$$\frac{\partial g}{\partial a}=-\frac{\bar{x}_U b}{a^2},$$

and

$$\frac{\partial g}{\partial b}=\frac{\bar{x}_U}{a}.$$

Thus, the asymptotic distribution of

$$\sqrt{n}[g(\bar{x},\bar{y})-g(\bar{x}_U,\bar{y}_U)]=\sqrt{n}\left[\hat{\bar{y}}_r-\bar{y}_U\right]$$

is normal with mean 0 and variance

$$\begin{aligned}
V &= \left(\frac{\partial g}{\partial a}\right)^2_{\bar{x}_U,\bar{y}_U}S_x^2+2\left(\frac{\partial g}{\partial a}\right)\left(\frac{\partial g}{\partial b}\right)_{\bar{x}_U,\bar{y}_U}RS_xS_y+\left(\frac{\partial g}{\partial b}\right)^2_{\bar{x}_U,\bar{y}_U}S_y^2\\
&= B^2S_x^2+2BS_xS_y+S_y^2.
\end{aligned}$$

**4.26** Using (4.8) and (4.17),

$$\begin{aligned} V(\hat{\bar{y}}_r) - V(\hat{\bar{y}}_{\text{reg}}) &\approx \left(1-\frac{n}{N}\right)\frac{1}{n}(S_y^2 - 2BRS_xS_y + B^2S_x^2) \\ &\quad -\left(1-\frac{n}{N}\right)\frac{1}{n}S_y^2(1-R^2) \\ &= \left(1-\frac{n}{N}\right)\frac{1}{n}[-2BRS_xS_y + B^2S_x^2 + R^2S_y^2] \\ &= \left(1-\frac{n}{N}\right)\frac{1}{n}[RS_y - BS_x]^2 \\ &\geq 0. \end{aligned}$$

**4.28**

$$\begin{aligned} &\sum_{i=1}^{N}\frac{(y_i - \bar{y}_U - B_1[x_i - \bar{x}_U])^2}{N-1} \\ &= \sum_{i=1}^{N}\frac{[(y_i-\bar{y}_U)^2 - 2B_1(x_i-\bar{x}_U)(y_i-\bar{y}_U) + B_1^2(x_i-\bar{x}_U)^2]}{N-1} \\ &= S_y^2 - 2B_1RS_xS_y + B_1^2S_x^2 \\ &= S_y^2 - 2\frac{R^2S_y}{S_x}S_xS_y + \frac{R^2S_y^2}{S_x^2}S_x^2 \\ &= S_y^2(1-R^2). \end{aligned}$$

**4.29** From (4.15),

$$\begin{aligned} \hat{\bar{y}}_{\text{reg}} &= \bar{y} + \hat{B}_1(\bar{x}_U - \bar{x}) \\ &= \bar{y} + B_1(\bar{x}_U - \bar{x}) + (\hat{B}_1 - B_1)(\bar{x}_U - \bar{x}). \end{aligned}$$

Thus,

$$\begin{aligned} E[\hat{\bar{y}}_{\text{reg}} - \bar{y}_U] &= E[\bar{y} + B_1(\bar{x}_U - \bar{x}) + (\hat{B}_1 - B_1)(\bar{x}_U - \bar{x})] - \bar{y}_U \\ &= E[(\hat{B}_1 - B_1)(\bar{x}_U - \bar{x})] \\ &= -\text{Cov}\,(\hat{B}_1, \bar{x}). \end{aligned}$$

Now,

$$\hat{B}_1 = \frac{\sum_{i\in\mathcal{S}}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i\in\mathcal{S}}(x_i-\bar{x})^2}$$

and

$$\begin{aligned} (x_i-\bar{x})(y_i-\bar{y}) &= (x_i-\bar{x})(y_i-\bar{y}_U+\bar{y}_U-\bar{y}) \\ &= (x_i-\bar{x})[d_i + B_1(x_i-\bar{x}_U) + \bar{y}_U - \bar{y}] \\ &= (x_i-\bar{x})[d_i + B_1(x_i-\bar{x}) + B_1(\bar{x}-\bar{x}_U) + \bar{y}_U - \bar{y}] \end{aligned}$$

with

$$\sum_{i\in\mathcal{S}}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i\in\mathcal{S}}(x_i-\bar{x})d_i+B_1\sum_{i\in\mathcal{S}}(x_i-\bar{x})^2.$$

Thus,

$$\hat{B}_1=B_1+\frac{\sum_{i\in\mathcal{S}}(x_i-\bar{x}_U)d_i+(\bar{x}_U-\bar{x})\sum_{i\in\mathcal{S}}d_i}{\sum_{i\in\mathcal{S}}(x_i-\bar{x})^2}.$$

Let $q_i=d_i(x_i-\bar{x}_U)$. Then

$$\sum_{i=1}^{N}q_i=\sum_{i=1}^{N}(y_i-\bar{y}_U)(x_i-\bar{x}_U)-B_1\sum_{i=1}^{N}(x_i-\bar{x}_U)^2=0$$

by the definition of $B_1$, so $\bar{q}_U=0$. Consequently,

$$\begin{aligned}E[(\hat{B}_1-B_1)(\bar{x}_U-\bar{x})] &= E\left[\frac{n\bar{q}+n(\bar{x}_U-\bar{x})\bar{d}}{(n-1)s_x^2}(\bar{x}_U-\bar{x})\right]\\ &\approx \frac{1}{S_x^2}E[\bar{q}(\bar{x}_U-\bar{x})+\bar{d}(\bar{x}_U-\bar{x})^2]\\ &= \frac{1}{S_x^2}\{-\mathrm{Cov}\,(\bar{q},\bar{x})+E[\bar{d}(\bar{x}_U-\bar{x})^2]\}.\end{aligned}$$

Since

$$\begin{aligned}\mathrm{Cov}\,(\bar{q},\bar{x}) &= \frac{1}{N-1}\left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{i=1}^{N}(q_i-\bar{q}_U)(x_i-\bar{x}_U)\\ &= \frac{1}{N-1}\left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{i=1}^{N}q_i(x_i-\bar{x}_U)\end{aligned}$$

and $E[\bar{d}(\bar{x}_U-\bar{x})^2]$ is of smaller order than $\mathrm{Cov}\,(\bar{q},\bar{x})$, the approximation is shown.

**4.32** From linear models theory, if $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\varepsilon$, with $E[\varepsilon]=0$ and $\mathrm{Cov}[\varepsilon]=\sigma^2\mathbf{A}$, then the weighted least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}^T\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}^{-1}Y$$

with

$$V[\hat{\boldsymbol{\beta}}]=\sigma^2(\mathbf{X}^T\mathbf{A}^{-1}\mathbf{X})^{-1}.$$

This result may be found in any linear models book (for instance, Christensen, 1996, p. 31). In our case,

$$\mathbf{Y}=(Y_1,\ldots,Y_n)^T,\quad \mathbf{X}=(x_1,\ldots,x_n)^T,\ \text{and}\ \mathbf{A}=\mathrm{diag}(x_1,\ldots,x_n)$$

so

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}^T\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}^{-1}\mathbf{Y}=\frac{\sum_{i=1}^{n}Y_i}{\sum_{i=1}^{n}x_i}=\frac{\bar{y}}{\bar{x}}$$

and

$$V[\hat{\boldsymbol{\beta}}]=\sigma^2(\mathbf{X}^T\mathbf{A}^{-1}\mathbf{X})^{-1}=\frac{\sigma^2}{\sum_{i=1}^{n}x_i}.$$

**4.35** (a) No, because some values are 0.

(b) When the entire population is sampled, $\hat{t}_x = t_x$ and $\hat{t}_y = t_y$, so $\hat{B} = t_y/t_x$ and $t_x\hat{B} = t_y$.

(c) Answers will vary.

(d) Letting $Z_i$ be the indicator variable for inclusion in the sample,

$$\begin{aligned} E[\bar{b} - B] &= E\left[\frac{1}{n}\sum_{i=1}^{N} Z_i b_i - \frac{t_y}{t_x}\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} b_i - \frac{\sum_{i=1}^{N} b_i x_i}{t_x} \\ &= \frac{1}{N}\frac{(\sum_{i=1}^{N} b_i)(\sum_{j=1}^{N} x_i)}{t_x} - \frac{\sum_{i=1}^{N} b_i x_i}{t_x} \\ &= -\frac{1}{t_x}\left[\sum_{i=1}^{N} b_i x_i - N\bar{b}_U\bar{x}_u\right] \\ &= -\frac{(N-1)S_{bx}}{t_x}. \end{aligned}$$

(e) From linear models theory, if $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$, with $E[\varepsilon] = 0$ and $\text{Cov}[\varepsilon] = \sigma^2\mathbf{A}$, then the weighted least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (X^T\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}^{-1}\mathbf{Y}$$

with

$$V[\hat{\boldsymbol{\beta}}] = \sigma^2(X^T\mathbf{A}^{-1}\mathbf{X})^{-1}.$$

Here, $\mathbf{A} = \text{diag}(x_i^2)$, so

$$\mathbf{X}^T\mathbf{A}^{-1}\mathbf{X} = \sum_{i\in\mathcal{S}} x_i\frac{1}{x_i^2}x_i = n$$

and

$$\mathbf{X}^T\mathbf{A}^{-1}\mathbf{Y} = \sum_{i\in\mathcal{S}} x_i\frac{1}{x_i^2}y_i = \sum_{i\in\mathcal{S}} Y_i/x_i.$$

**4.42**

(a)

```
proc surveymeans  data = vius mean sum clm clsum;
   weight tabtrucks;
   strata stratum;
   var miles_annl;
```

```
   domain business;
run;
```

| Business in which vehicle was most often used during 2002 | Variable | Mean | Std Error of Mean |
|---|---|---|---|
| For-hire tra | MILES_ANNL | 56452 | 1246.317720 |
| Vehicle leas | MILES_ANNL | 23306 | 790.981413 |
| Agriculture, | MILES_ANNL | 10768 | 415.312729 |
| Mining | MILES_ANNL | 19210 | 1126.305405 |
| Utilities | MILES_ANNL | 15081 | 830.112607 |
| Construction | MILES_ANNL | 16714 | 381.173223 |
| Manufacturin | MILES_ANNL | 19650 | 1018.600046 |
| Wholesale tr | MILES_ANNL | 23052 | 1184.715817 |
| Retail trade | MILES_ANNL | 17948 | 561.147469 |
| Information | MILES_ANNL | 14927 | 1396.653144 |
| Waste manage | MILES_ANNL | 14410 | 726.842687 |
| Arts, entert | MILES_ANNL | 9536.588165 | 1267.485898 |
| Accommodatio | MILES_ANNL | 20461 | 1300.857231 |
| Other servic | MILES_ANNL | 16818 | 600.239019 |

Domain Analysis: Business in which vehicle was most often used

| Business in which vehicle was most often used during 2002 | Variable | 95% CL for Mean | | Sum |
|---|---|---|---|---|
| For-hire tra | MILES_ANNL | 54009.5907 | 58895.1519 | 72272793289 |
| Vehicle leas | MILES_ANNL | 21755.6067 | 24856.2511 | 20024589014 |
| Agriculture, | MILES_ANNL | 9954.2886 | 11582.3131 | 24119946651 |
| Mining | MILES_ANNL | 17002.3551 | 21417.4684 | 3411543277 |
| Utilities | MILES_ANNL | 13454.3175 | 16708.3561 | 10244675655 |
| Construction | MILES_ANNL | 15966.8537 | 17461.0515 | 75906142636 |
| Manufacturin | MILES_ANNL | 17653.3448 | 21646.2535 | 15384530602 |
| Wholesale tr | MILES_ANNL | 20729.7496 | 25373.8316 | 16963450921 |
| Retail trade | MILES_ANNL | 16848.3582 | 19048.0543 | 27470445448 |
| Information | MILES_ANNL | 12189.9160 | 17664.7915 | 5622014452 |

| | | | | |
|---|---|---|---|---|
| Waste manage | MILES_ANNL | 12985.5076 | 15834.7285 | 10709275945 |
| Arts, entert | MILES_ANNL | 7052.3180 | 12020.8583 | 1784083855 |
| Accommodatio | MILES_ANNL | 17911.2857 | 23010.6416 | 5816313888 |
| Other servic | MILES_ANNL | 15641.1955 | 17994.1304 | 35776203775 |

| Business in which vehicle was most often used during 2002 | Variable | Std Dev | 95% CL for Sum | |
|---|---|---|---|---|
| For-hire tra | MILES_ANNL | 1608230919 | 6.91207E10 | 7.54249E10 |
| Vehicle leas | MILES_ANNL | 1213307392 | 1.76465E10 | 2.24027E10 |
| Agriculture, | MILES_ANNL | 1354386330 | 2.14654E10 | 2.67745E10 |
| Mining | MILES_ANNL | 360265917 | 2705422697 | 4117663856 |
| Utilities | MILES_ANNL | 942933274 | 8396528057 | 1.20928E10 |
| Construction | MILES_ANNL | 2821651145 | 7.03757E10 | 8.14366E10 |
| Manufacturin | MILES_ANNL | 1399406209 | 1.26417E10 | 1.81274E10 |
| Wholesale tr | MILES_ANNL | 1348917090 | 1.43196E10 | 1.96073E10 |
| Retail trade | MILES_ANNL | 1422261638 | 2.46828E10 | 3.02581E10 |
| Information | MILES_ANNL | 923917751 | 3811137245 | 7432891659 |
| Waste manage | MILES_ANNL | 901989658 | 8941377763 | 1.24772E10 |
| Arts, entert | MILES_ANNL | 353650310 | 1090929855 | 2477237856 |
| Accommodatio | MILES_ANNL | 677802928 | 4487821313 | 7144806463 |
| Other servic | MILES_ANNL | 2201296141 | 3.14617E10 | 4.00907E10 |

(b)

```
proc surveymeans  data = vius mean clm ;
   weight tabtrucks;
   strata stratum;
   var mpg;
   domain transmssn;
run;
```

Domain Analysis: Type of Transmission

| Type of Transmission | Variable | Mean |
|---|---|---|
| Automatic | MPG | 16.665277 |
| Manual | MPG | 16.022122 |
| Semi-Automat | MPG | 14.846222 |
| Automated Ma | MPG | 16.732086 |

Domain Analysis: Type of Transmission

| Type of Transmission | Variable | Std Error of Mean | 95% CL for Mean | |
|---|---|---|---|---|
| Automatic | MPG | 0.043659 | 16.5797047 | 16.7508490 |
| Manual | MPG | 0.097739 | 15.8305539 | 16.2136901 |
| Semi-Automat | MPG | 1.012044 | 12.8626219 | 16.8298213 |
| Automated Ma | MPG | 1.599588 | 13.5969068 | 19.8672658 |

(c)

```
proc surveymeans  data = vius mean clm ;
   weight tabtrucks;
   strata stratum;
   var mpg;
   domain transmssn;
run;
```

The estimated ratio is 0.124410 with 95% CI [0.12258, 0.12624].

# Chapter 5

# Cluster Sampling with Equal Probabilities

**5.1** If the nonresponse can be ignored, then $\hat{p}$ is the ratio estimate of the proportion. The variance estimate given in the problem, though, assumes that an SRS of voters was taken. But this was a *cluster* sample—the sampling unit was a residential telephone number, not an individual voter. As we expect that voters in the same household are more likely to have similar opinions, the estimated variance using simple random sampling is probably too small.

**5.3** (a) This is a cluster sample because there are two levels of sampling units: the wetlands are the psus and the sites are the ssus.

(b) The analysis is not appropriate. A two-sample $t$ test assumes that all observations are independent. This is a cluster sample, however, and sites within the same wetland are expected to be more similar than sites selected at random from the population.

**5.4** (a) This is a cluster sample because the primary sampling unit is the journal, and the secondary sampling unit is an article in the journal from 1988.

(b) Let

$$M_i = \text{ number of articles in journal } i$$

and

$t_i =$ number of articles in journal $i$ that use non-probability sampling designs.

From the data file,

$$\sum_{i \in \mathcal{S}} t_i = 137$$

and

$$\sum_{i \in \mathcal{S}} M_i = 148.$$

Then, using (5.16),

$$\hat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}} t_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{137}{148} = 0.926.$$

The estimated variance of the residuals is

$$\frac{\sum_{i\in\mathcal{S}}(t_i - \hat{\bar{y}}_r M_i)^2}{n-1} = 0.993221;$$

using (5.18), with $\overline{M}_{\mathcal{S}}$ substituted for $\overline{M}_U$,

$$\text{SE}\,[\hat{\bar{y}}_r] = \sqrt{\left(1 - \frac{26}{1285}\right)\frac{1}{26(5.69)^2}(.993221)} = .034.$$

Here is SAS code:

```
data journal;
   infile journal delimiter=',' firstobs=2;
   input numemp prob nonprob ;
   sampwt = 1285/26;
   /* weight = N/n  since this is a one-stage cluster sample */

proc surveymeans data=journal total = 1285 mean clm sum clsum;
   weight sampwt;
   var numemp nonprob;
   ratio 'nonprob/(number of articles)' nonprob/numemp;
run;
```

**5.5**

```
options ls=78 nodate nocenter;
data spanish;
   infile spanish delimiter=',' firstobs=2;
   input class score trip;
   sampwt = 72/10;
   /* weight = N/n = 72/10 since one-stage cluster sample*/

proc print data=spanish;
run;

proc surveymeans data=spanish total = 72 mean clm sum clsum;
   weight sampwt;
   cluster class;
   var trip score;
run;
```

```
/* Construct a boxplot of the data */

proc sort data=spanish;
   by class;

proc boxplot data=spanish;
   plot score * class;
run;

/* Since this is a one-stage cluster sample, there is
   no contribution to variance from subsampling. */

proc surveymeans data=spanish mean clm sum clsum;
   weight sampwt;
   cluster class;
   var trip score;
run;
```

The SURVEYMEANS Procedure

Data Summary

| | |
|---|---|
| Number of Clusters | 10 |
| Number of Observations | 196 |
| Sum of Weights | 1411.2 |

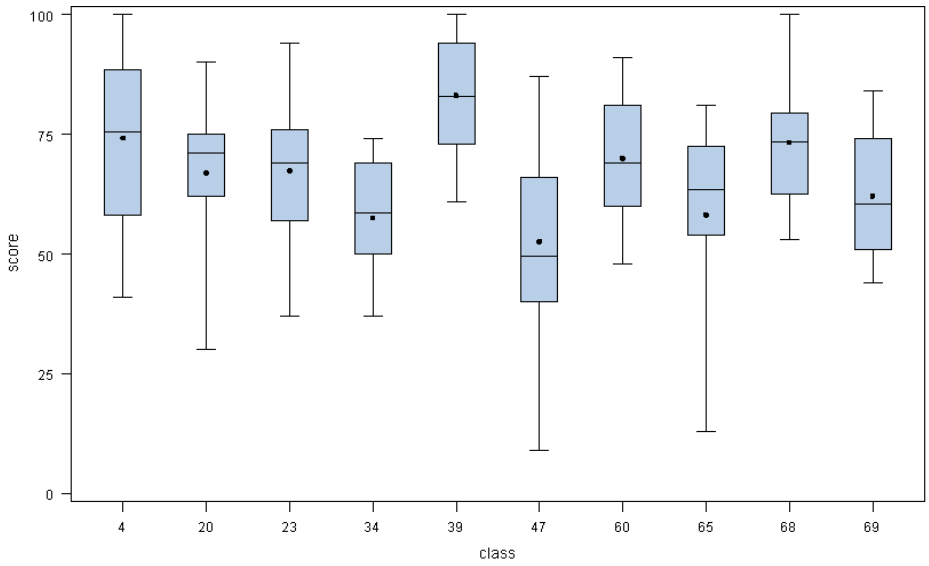
Statistics

| Variable | Mean | Std Error of Mean | 95% CL for Mean | |
|---|---|---|---|---|
| trip | 0.321429 | 0.079001 | 0.1427164 | 0.5001408 |
| score | 66.795918 | 2.919409 | 60.1917561 | 73.4000806 |

Statistics

| Variable | Sum | Std Dev | 95% CL for Sum | |
|---|---|---|---|---|
| trip | 453.600000 | 120.502946 | 181.0034 | 726.197 |
| score | 94262 | 7301.420092 | 77745.4402 | 110779.360 |

Here is a side-by-side boxplot for *score*.




**5.6** (a) The SAS code below was used to calculate summary statistics and the ANOVA table. The output is given below.

```
data worms;
  do case = 1 to 12;
   do can = 1 to 3;
      input worms @@;
  wt = (580/12)*(24/3);
      output;
   end;
  end;
 cards;
 1 5 7
 4 2 4
 0 1 2
 3 6 6
 4 9 8
 0 7 3
 5 5 1
 3 0 2
 7 3 5
 3 1 4
 4 7 9
 0 0 0
 ;
 proc print data=worms;
 run;
```

```
 proc glm data=worms;
    class case;
    model worms = case;
    mean case;
 run;

 /* SAS does not calculate the extra term for variance
    due to 2-stage sampling */
proc surveymeans data=worms total = 580;
    weight wt;
cluster case;
var worms;
run;
```

The GLM Procedure

Class Level Information

| Class | Levels | Values |
|---|---|---|
| case | 12 | 1 2 3 4 5 6 7 8 9 10 11 12 |

Number of Observations Read 36
Number of Observations Used 36

Dependent Variable: worms

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 11 | 149.6388889 | 13.6035354 | 3.00 | 0.0117 |
| Error | 24 | 108.6666667 | 4.5277778 | | |
| Corrected Total | 35 | 258.3055556 | | | |

| R-Square | Coeff Var | Root MSE | worms Mean |
|---|---|---|---|
| 0.579310 | 58.47547 | 2.127858 | 3.638889 |

| Level of | ----------worms---------- |
|---|---|

```
case          N             Mean          Std Dev

1             3       4.33333333       3.05505046
2             3       3.33333333       1.15470054
3             3       1.00000000       1.00000000
4             3       5.00000000       1.73205081
5             3       7.00000000       2.64575131
6             3       3.33333333       3.51188458
7             3       3.66666667       2.30940108
8             3       1.66666667       1.52752523
9             3       5.00000000       2.00000000
10            3       2.66666667       1.52752523
11            3       6.66666667       2.51661148
12            3       0.00000000       0.00000000
```

From the output, we see that

$$\hat{\bar{y}}_{\text{unb}} = 3.639$$

(this works here because $M_i = M$ for all $i$ and $m_i = m$ for all $i$; in this case, (5.26) reduces to the sample mean).

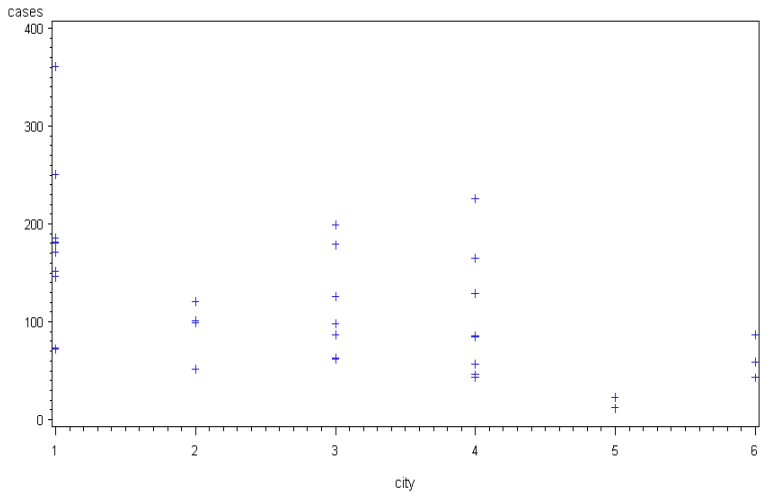
$$\begin{aligned} \text{SE}\left[\hat{\bar{y}}_{\text{unb}}\right] &= \sqrt{\left(1 - \frac{12}{580}\right)\frac{13.604}{(12)(3)} + \frac{1}{580}\left(1 - \frac{3}{24}\right)\frac{4.528}{3}} \\ &= \sqrt{0.3701 + 0.0023} \\ &= 0.61. \end{aligned}$$

Note that the second term contributes little to the standard error.

Here is the approximation from SAS, using PROC SURVEYMEANS:

```
The SURVEYMEANS Procedure

            Data Summary

Number of Clusters                12
Number of Observations            36
Sum of Weights                 13920


                                  Statistics

                              Std Error
Variable     N      Mean      of Mean       95% CL for Mean
-------------------------------------------------------------
worms       36  3.638889     0.614716    2.28590770 4.99187008
-------------------------------------------------------------
```

**5.7** We used SAS to obtain the mean and standard deviation for each city, and to plot the data.




```
The GLM Procedure

      Class Level Information

Class         Levels    Values

city               6    1 2 3 4 5 6


Number of Observations Read          34
Number of Observations Used          34


The GLM Procedure

Dependent Variable: cases

                            Sum of
Source          DF         Squares     Mean Square    F Value    Pr > F

Model            5      68339.9641      13667.9928       3.37    0.0166

Error           28     113570.6536       4056.0948

Corrected Total             33     181910.6176


R-Square     Coeff Var      Root MSE    cases Mean
```

```
0.375679      53.85163      63.68748      118.2647


The GLM Procedure

Level of            ------------cases------------
city          N              Mean          Std Dev

1            10       177.300000       83.5996411
2             4        93.250000       29.2389580
3             7       116.285714       54.5396317
4             8       104.625000       64.5930945
5             2        17.500000        7.7781746
6             3        63.000000       22.2710575

The SURVEYMEANS Procedure

            Data Summary

Number of Clusters                  6
Number of Observations             34
Sum of Weights                 1267.5


Statistics

                           Std Error
Variable          Mean       of Mean     95% CL for Mean
-------------------------------------------------------------
cases       120.688145     19.730137  69.9702138 171.406075
-------------------------------------------------------------

      Sum       Std Dev      95% CL for Sum
-----------------------------------------------
   152972         56601  7473.70477 298470.742
```

We can also use (5.18) and (5.24) for calculating the total number of cases sold, and (5.26) and (5.28) for calculating the average number of cases sold per supermarket. Summary quantities are given in the spreadsheet below.

| City | $M_i$ | $m_i$ | $\bar{y}_i$ | $s_i$ | $\hat{t}_i$ | $\hat{t}_i - M_i\hat{\bar{y}}_r$ | $(1-\frac{m_i}{M_i})M_i^2\frac{s_i^2}{m_i}$ |
|---|---|---|---|---|---|---|---|
| 1 | 52 | 10 | 177.30 | 83.60 | 9219.6 | 2943.8 | 1526376 |
| 2 | 19 | 4 | 93.25 | 29.24 | 1771.8 | -521.3 | 60913 |
| 3 | 37 | 7 | 116.29 | 54.54 | 4302.6 | -162.9 | 471682 |
| 4 | 39 | 8 | 104.63 | 64.59 | 4080.4 | -626.5 | 630534 |
| 5 | 8 | 2 | 17.50 | 7.78 | 140.0 | -825.5 | 1452 |
| 6 | 14 | 3 | 63.00 | 22.27 | 882.0 | -807.6 | 25461 |
| Sum | 169 | 34 | | | 20396.3 | | 2716418 |
| Var | | | | | 10952882 | 2138111 | |

From (5.18),

$$\hat{t}_{\text{unb}} = \frac{45}{6}(20396.3) = 152972.$$

Using (5.24),

$$\begin{aligned} \text{SE}\left[\hat{t}_{\text{unb}}\right] &= \sqrt{45^2\left(1-\frac{6}{45}\right)\frac{10952882}{6} + \frac{45}{6}(2716418)} \\ &= \sqrt{3{,}203{,}717{,}941 + 20{,}373{,}134} \\ &= 56,781. \end{aligned}$$

From (5.26) and (5.28),

$$\hat{\bar{y}}_r = \frac{20{,}396}{169} = 120.7$$

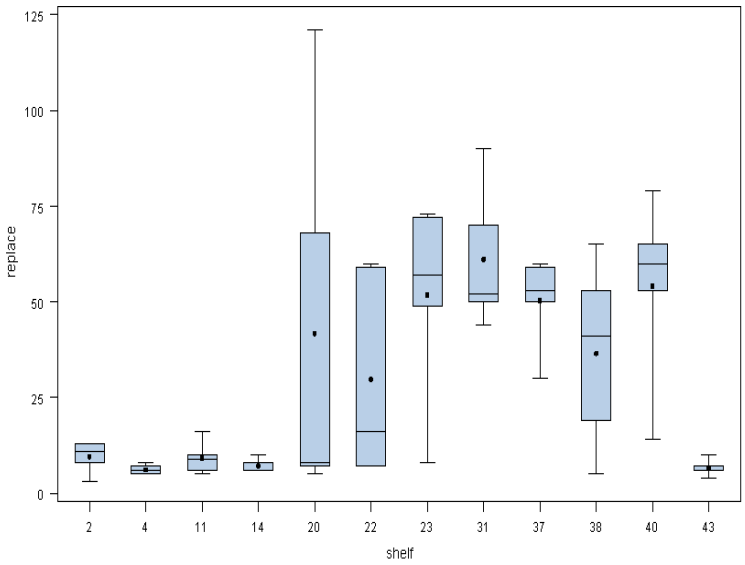
and

$$\begin{aligned} \text{SE}\left[\hat{\bar{y}}_r\right] &= \frac{1}{28.17}\sqrt{\left(1-\frac{6}{45}\right)\frac{2{,}138{,}111}{6} + \frac{1}{6(45)}(2716418)} \\ &= 21.05. \end{aligned}$$

**5.8** (a) The SAS code below computes estimates.

```
data books;
   infile booksdat delimiter=',' firstobs=2;
   input shelf Mi number purchase replace;
   sampwt = (44/12)*(Mi/5);
   /* The crucial part for estimating the total is correctly
      specifying the sampling weight. */

proc print data=books;
run;

/* Construct the plot for part (a) */

proc boxplot data=books;
   plot replace * shelf;
run;

/*Here is the with-replacement approximation to the variance */

proc surveymeans data=books mean clm sum clsum;
   weight sampwt;
   cluster shelf;
   var replace;
run;

/* Here is the without-replacement approximation to the variance
```

```
   using SAS. Note that this approximation does not include the
   second term in (5.21). */

proc surveymeans data=books total = 44 mean clm sum clsum;
   weight sampwt;
   cluster shelf;
   var replace;
run;
```




It appears that the means and variances differ quite a bit for different shelves.

(b) Quantities used in calculation are in the spreadsheet below.

| Shelf | $M_i$ | $m_i$ | $\bar{y}_i$ | $s_i^2$ | $\hat{t}_i = M_i\bar{y}_i$ | $(1-\frac{m_i}{M_i})M_i^2\frac{s_i^2}{m_i}$ | $(\hat{t}_i - M_i\hat{\bar{y}}_r)^2$ |
|---|---|---|---|---|---|---|---|
| 2 | 26 | 5 | 9.6 | 17.80 | 249.6 | 1943.76 | 132696.9 |
| 4 | 52 | 5 | 6.2 | 1.70 | 322.4 | 830.96 | 819661.7 |
| 11 | 70 | 5 | 9.2 | 18.70 | 644.0 | 17017.00 | 1017561.8 |
| 14 | 47 | 5 | 7.2 | 3.20 | 338.4 | 1263.36 | 594901.6 |
| 20 | 5 | 5 | 41.8 | 2666.70 | 209.0 | 0.00 | 8271.3 |
| 22 | 28 | 5 | 29.8 | 748.70 | 834.4 | 96432.56 | 30033.9 |
| 23 | 27 | 5 | 51.8 | 702.70 | 1398.6 | 83480.76 | 579293.8 |
| 31 | 29 | 5 | 61.2 | 353.20 | 1774.8 | 49165.44 | 1188301.2 |
| 37 | 21 | 5 | 50.4 | 147.30 | 1058.4 | 9898.56 | 316493.1 |
| 38 | 31 | 5 | 36.6 | 600.80 | 1134.6 | 96848.96 | 162144.0 |
| 40 | 14 | 5 | 54.2 | 595.70 | 758.8 | 15011.64 | 183399.3 |
| 43 | 27 | 5 | 6.6 | 4.80 | 178.2 | 570.24 | 210944.1 |
| | | | | | | | |
| sum | 377 | | | | 8901.2 | 372463.2 | 5243902.9 |
| var | | | | | 268531.00 | | |

To estimate the total replacement value of the book collection, we use the unbiased

estimator since $M_0$, the total number of books, is unknown.

$$\hat{t}_{\text{unb}} = \frac{N}{n}\sum_{i\in\mathcal{S}}\hat{t}_i = \frac{44}{12}8901.2 = 32637.73.$$

From the spreadsheet, $s_t^2 = 268{,}531$ and $\sum_{i\in\mathcal{S}} M_i^2(1-m_i/M_i)s_i^2/m_i = 372463.2$, so

$$\begin{aligned}
\text{SE}\left(\hat{t}_{\text{unb}}\right) &= N\sqrt{\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{1}{nN}\sum_{i\in\mathcal{S}} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}} \\
&= 44\sqrt{\left(1-\frac{12}{44}\right)\frac{268,531}{12} + \frac{1}{(12)(44)}(372463.2)} \\
&= 44\sqrt{16274.6 + 705.4} = 5733.52.
\end{aligned}$$

The standard error of $\hat{t}_{\text{unb}}$ is 5733.52, which is quite large when compared with the estimated total. The estimated coefficient of variation for $\hat{t}_{\text{unb}}$ is

$$\text{SE}\left(\hat{t}_{\text{unb}}\right)/\hat{t}_{\text{unb}} = \frac{5733.52}{32637.73} = 0.176.$$

Here is the approximation using SAS and the with-replacement variance:

```
                                Statistics

                             Std Error
Variable           Mean        of Mean      95% CL for Mean

replace       23.610610       6.344128   9.64727746 37.5739427


                    Statistics

Variable          Sum        Std Dev        95% CL for Sum

replace         32638    6582.020830    18150.8032 47124.6635
```

Note that the with-replacement variance is too large because the psu sampling fraction is large.

Here is the approximation using SAS and the without-replacement variance:

```
                                Statistics

                             Std Error
Variable           Mean        of Mean      95% CL for Mean
```

```
replace        23.610610        5.410291    11.7026400 35.5185801


                    Statistics

Variable           Sum         Std Dev         95% CL for Sum

replace          32638     5613.166224     20283.2378 44992.2289
```

Note that $5613 = 44\sqrt{16274.6}$; SAS calculates only the first term of the without-replacement variance.

(c) To find the average replacement cost per book with the information given, we use the ratio estimate in (5.28):

$$\hat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}} \hat{t}_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{8901.2}{377} = 23.61.$$

In the formula for variance in (5.29),

$$\hat{V}(\hat{\bar{y}}_r) = \left(1 - \frac{12}{44}\right)\frac{s_r^2}{n\overline{M}^2} + \frac{1}{(12)(44)\overline{M}^2}\sum_{i\in\mathcal{S}} M_i^2\left(1 - \frac{m_i}{M_i}\right)\frac{s_i^2}{m_i},$$

we use $\overline{M}_{\mathcal{S}} = 31.417$, the average of the $M_i$ in our sample. So we have

$$s_r^2 = \frac{\sum_{i\in\mathcal{S}}(\hat{t}_i - M_i\hat{\bar{y}}_r)^2}{n-1} = \frac{5243902.93}{11} = 476718.4455$$

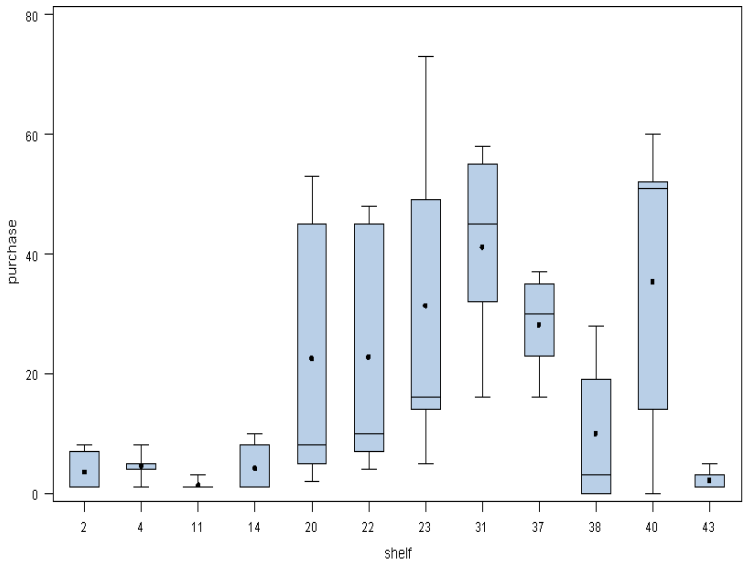
and

$$\begin{aligned}\mathrm{SE}\,(\hat{\bar{y}}_r) &= \frac{1}{31.417}\sqrt{\left(1 - \frac{12}{44}\right)\frac{476718.4455}{12} + 705.4} \\ &= \frac{1}{31.417}\sqrt{28892.0 + 705.4} = 5.476.\end{aligned}$$

The estimated coefficient of variation for $\hat{\bar{y}}_r$ is

$$\mathrm{SE}\,(\hat{\bar{y}}_r)/\hat{\bar{y}}_r = 5.476/23.61 = 0.232.$$

**5.9** (a)




It appears that the means and variances differ quite a bit for different shelves.

Here is SAS code and output for the purchase price of the books.

```
filename booksdat 'books.csv';

options ls=78 nodate nocenter;
data books;
   infile booksdat delimiter=',' firstobs=2;
   input shelf Mi number purchase replace;
   sampwt = (44/12)*(Mi/5);
   /* The crucial part for estimating the total is correctly
      specifying the sampling weight. */

proc print data=books;
run;

/* Construct the plot for part (a) */

proc boxplot data=books;
   plot purchase * shelf;
run;

/* Here is the with-replacement approx to the variance using SAS */

proc surveymeans data=books mean clm sum clsum;
   weight sampwt;
   cluster shelf;
   var purchase;
```

```
run;

/* Here is the without-replacement approximation to the variance
   using SAS. Note that this approximation does not include the
   second term in (5.21). */

proc surveymeans data=books total = 44 mean clm sum clsum;
   weight sampwt;
   cluster shelf;
   var purchase;
run;

The SURVEYMEANS Procedure

            Data Summary

Number of Clusters                12
Number of Observations            60
Sum of Weights            1382.33333


                                Statistics

                           Std Error
Variable           Mean        of Mean      95% CL for Mean

purchase      12.942706       3.630539   4.95194389 20.9334672

                    Statistics

Variable       Sum       Std Dev        95% CL for Sum

purchase     17891   3854.814474    9406.74388 26375.5228
```

**5.10** Here is the sample ANOVA table for the books data, calculated using SAS.

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 11 | 25570.98333 | 2324.635 | 4.76 | 0.0001 |
| Error | 48 | 23445.20000 | 488.442 | | |
| Corrected Total | 59 | 49016.18333 | | | |

We use (5.10) to estimate $R_a^2$: we have

$$\widehat{\text{MSW}} = 488.44167,$$

and we (with slight bias) estimate $S^2$ by

$$S^2 = \frac{\text{SSTO}}{\text{dfTO}} = \frac{49016.18333}{59} = 830.78,$$

so

$$\hat{R}_a^2 = 1 - \widehat{\text{MSW}}/\hat{S}^2 = 1 - 488.44/830.78 = 0.41.$$

The positive value of $\hat{R}_a^2$ indicates that books on the same shelf do tend to have more similar replacement costs.

**5.11** (a) Here, $N = 828$ and $n = 85$. We have the following frequencies for $t_i =$ number of errors in claim $i$:

| Number of errors | Frequency |
|---|---|
| 0 | 57 |
| 1 | 22 |
| 2 | 4 |
| 3 | 1 |
| 4 | 1 |

The 85 claims have a total of $\sum_{i\in\mathcal{S}} t_i = 37$ errors, so from (5.1),

$$\hat{t} = \frac{828}{85}(37) = 360.42$$

and

$$\begin{aligned} s_t^2 &= \frac{1}{84}\sum_{i\in\mathcal{S}}\left(t_i - \frac{37}{85}\right)^2 \\ &= \frac{1}{84}\left[57\left(0 - \frac{37}{85}\right)^2 + 22\left(1 - \frac{37}{85}\right)^2 + 4\left(2 - \frac{37}{85}\right)^2 \right. \\ &\qquad \left. + \left(3 - \frac{37}{85}\right)^2 + \left(4 - \frac{37}{85}\right)^2\right] \\ &= \frac{1}{84}[10.800 + 7.02 + 9.79 + 6.58 + 12.71] \\ &= 0.558263. \end{aligned}$$

Thus the error rate, using (5.4), is

$$\hat{\bar{y}} = \frac{\hat{t}}{NM} = \frac{360.42}{(828)(215)} = 0.002025.$$

From (5.6),

$$\text{SE}\,[\hat{\bar{y}}] = \frac{1}{215}\sqrt{\left(1 - \frac{85}{828}\right)\frac{s_t^2}{85}} = \frac{1}{215}(.07677) = .000357.$$

(b) From calculations in (a),

$$\hat{t} = 360.42$$

and, using (5.3),

$$\mathrm{SE}\,[\hat{t}] = 828\sqrt{\left(1 - \frac{85}{828}\right)\frac{s_t^2}{85}} = 63.565.$$

(c) If the same number of errors (37) had been obtained using an SRS of 18,275 of the 178,020 fields, the error rate would be

$$\frac{37}{18{,}275} = .002025$$

(the same as in (a)). But the estimated variance from an SRS would be

$$\hat{V}[\hat{p}_{\mathrm{srs}}] = \left(1 - \frac{18{,}275}{178{,}020}\right)\frac{\hat{p}_{\mathrm{srs}}(1-\hat{p}_{\mathrm{srs}})}{18{,}274} = 9.92 \times 10^{-8}.$$

The estimate varianced from (a) is

$$(\mathrm{SE}\,[\hat{\bar{y}}])^2 = 1.28 \times 10^{-7}.$$

Thus

$$\frac{\text{Estimated variance under cluster design}}{\text{Estimated variance under SRS}} = \frac{1.28 \times 10^{-7}}{9.92 \times 10^{-8}} = 1.29.$$

**5.12** Students should plot the data similarly to Figure 5.3.

$$\hat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}} \hat{t}_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{85478.56}{1757} = 48.65.$$

$s_r^2$ is the sample variance of the residuals $\hat{t}_i - M_i\hat{\bar{y}}_r$: here,

$$s_r^2 = \frac{\sum_{i\in\mathcal{S}}(\hat{t}_i - M_i\hat{\bar{y}}_r)^2}{183} = 280.0261.$$

Since $N$ is unknown, (5.29) yields

$$\mathrm{SE}\,[\hat{\bar{y}}_r] = \frac{1}{9.549}\sqrt{\frac{280.0261}{184}} = 0.129.$$

**5.13** (a) Cluster sampling is needed for this application because the household is the sampling unit. Yet the Arizona Statutes specify the statistic that must be used for estimating the error rate: it must be estimated by the sample mean. It is therefore important to make sure that a self-weighting sample is taken. In this application, a self-weighting sample will result if an SRS of $n$ households is taken from the population of $N$ households in the county, and if all individuals in the household are measured. It makes sense in this example to take a one-stage cluster sample.

(b) We know that if we were taking an SRS of persons, we would calculate $n_0 = (1.96)^2(0.1)(0.9)/(0.03)^2 = 385$ and

$$n = \frac{n_0}{1 + n_0/N} = 310.$$

Since we obtain at least some additional information by sampling everyone in the household, we need at most 310 households. To calculate the number of households to be sampled, we need an idea of the measure of homogeneity within the households. We know from the equation following (5.11) that

$$\frac{V(\hat{t}_{\text{cluster}})}{V(\hat{t}_{\text{SRS}})} = 1 + \frac{N(M-1)}{N-1} R_a^2$$

so we can calculate a sample size by multiplying the number of persons needed for an SRS (310) by the ratio of variances, then dividing by $M$. The following table gives some sample sizes for different values of $R_a^2$:

| M | $R_a^2$ | sample size |
|---|---|---|
| 1 | 0.1 | 310 |
| 2 | 0.1 | 171 |
| 3 | 0.1 | 124 |
| 4 | 0.1 | 101 |
| 5 | 0.1 | 87 |
| 1 | 0.5 | 310 |
| 2 | 0.5 | 233 |
| 3 | 0.5 | 207 |
| 4 | 0.5 | 194 |
| 6 | 0.5 | 181 |
| 1 | 0.8 | 310 |
| 2 | 0.8 | 279 |
| 3 | 0.8 | 269 |
| 4 | 0.8 | 264 |
| 5 | 0.8 | 260 |

**5.14** (a) Treating the proportions as means, and letting $M_i$ and $m_i$ be the number of female students in the school and interviewed, respectively, we have the following summaries.

| School | $M_i$ | $m_i$ | smokers | $\bar{y}_i$ | $s_i^2$ | $\hat{t}_i$ | $\hat{t}_i - M_i\hat{\bar{y}}_r$ | $(1 - \frac{m_i}{M_i})\frac{M_i^2 s_i^2}{m_i}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 792 | 25 | 10 | 0.4 | .010 | 316.8 | -24.7 | 243 |
| 2 | 447 | 15 | 3 | 0.2 | .011 | 89.4 | -103.3 | 147 |
| 3 | 511 | 20 | 6 | 0.3 | .011 | 153.3 | -67.0 | 139 |
| 4 | 800 | 40 | 27 | 0.675 | .006 | 540.0 | 195.1 | 86 |
| Sum | 2550 | 100 | | | | 1099.5 | | 614 |
| Var | | | | | | | 17943 | |

Then,

$$\begin{aligned}
\hat{\bar{y}}_r &= \frac{1099.5}{2550} = 0.43 \\
\text{SE}\left[\hat{\bar{y}}_r\right] &= \frac{1}{637.5}\sqrt{\left(1-\frac{4}{29}\right)\frac{17943}{4} + \frac{1}{4(29)}(614)} \\
&= \frac{1}{637.5}\sqrt{3867.0 + 5.3} \\
&= 0.098.
\end{aligned}$$

(b) We construct a sample ANOVA table from the summary information in part (a). Note that

$$\text{ssw} = \sum_{i=1}^{4}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{4}(m_i - 1)s_i^2,$$

and

| Source | df | SS | MS |
|---|---|---|---|
| Between psus | 3 | | |
| Within psus | 96 | 0.837 | 0.0087 |
| Total | 99 | | |

**5.15** (a) A cluster sample was used for this study because Arizona has no list of all elementary school teachers in the state. All schools would have to be contacted to construct a sampling frame of teachers, and this would be expensive. Taking a cluster sample also makes it easier to distribute surveys. It's possible that distributing questionnaires through the schools might improve cooperation with the survey and give respondents more assurance that their data are kept confidential.

(b) The means and standard deviations, after eliminating records with missing values, are in the following table:

| School | Mean | Std. Dev. |
|---|---|---|
| 11 | 33.99 | 1.953 |
| 12 | 36.12 | 4.942 |
| 13 | 34.58 | 0.722 |
| 15 | 36.76 | 0.746 |
| 16 | 36.84 | 1.079 |
| 18 | 35.00 | 0.000 |
| 19 | 34.87 | 0.231 |
| 20 | 36.36 | 2.489 |
| 21 | 35.41 | 3.154 |
| 22 | 35.68 | 4.983 |
| 23 | 35.17 | 3.392 |
| 24 | 31.94 | 0.860 |
| 25 | 31.25 | 0.668 |
| 28 | 31.46 | 2.627 |
| 29 | 29.11 | 2.440 |
| 30 | 35.79 | 1.745 |
| 31 | 34.52 | 1.327 |
| 32 | 35.46 | 1.712 |
| 33 | 26.82 | 0.380 |
| 34 | 27.42 | 0.914 |
| 36 | 36.98 | 2.961 |
| 38 | 37.66 | 1.110 |
| 41 | 36.88 | 0.318 |

There appears to be large variation among the school means. An ANOVA table for the data is shown below.

| | Df | Sum of Sq | Mean Sq | F Value | Pr(F) |
|---|---|---|---|---|---|
| school | 22 | 1709.187 | 77.69034 | 13.76675 | 0 |
| Residuals | 224 | 1264.106 | 5.64333 | | |

(c) There appears to be a wider range of standard deviations for schools with higher means.

(d) Calculations are in the following table.

| School | $M_i$ | $m_i$ | $\bar{y}_i$ | $s_i^2$ | $M_i\bar{y}_i$ | resids | $M_i^2(1-\frac{m_i}{M_i})\frac{s_i^2}{m_i}$ |
|---|---|---|---|---|---|---|---|
| 11 | 33 | 10 | 33.990 | 3.814 | 1121.670 | 5.501 | 289.508 |
| 12 | 16 | 13 | 36.123 | 24.428 | 577.969 | 36.797 | 90.196 |
| 13 | 22 | 3 | 34.583 | 0.521 | 760.833 | 16.721 | 72.569 |
| 15 | 24 | 24 | 36.756 | 0.557 | 882.150 | 70.391 | 0.000 |
| 16 | 27 | 24 | 36.840 | 1.165 | 994.669 | 81.440 | 3.933 |
| 18 | 18 | 2 | 35.000 | 0.000 | 630.000 | 21.181 | 0.000 |
| 19 | 16 | 3 | 34.867 | 0.053 | 557.867 | 16.694 | 3.698 |
| 20 | 12 | 8 | 36.356 | 6.197 | 436.275 | 30.396 | 37.185 |
| 21 | 19 | 5 | 35.410 | 9.948 | 672.790 | 30.147 | 529.234 |
| 22 | 33 | 13 | 35.677 | 24.835 | 1177.338 | 61.170 | 1260.867 |
| 23 | 31 | 16 | 35.175 | 11.506 | 1090.425 | 41.903 | 334.393 |
| 24 | 30 | 9 | 31.944 | 0.740 | 958.333 | -56.366 | 51.776 |
| 25 | 23 | 8 | 31.250 | 0.446 | 718.750 | -59.186 | 19.252 |
| 28 | 53 | 17 | 31.465 | 6.903 | 1667.630 | -125.005 | 774.766 |
| 29 | 50 | 8 | 29.106 | 5.955 | 1455.313 | -235.852 | 1563.270 |
| 30 | 26 | 22 | 35.791 | 3.045 | 930.564 | 51.158 | 14.393 |
| 31 | 25 | 18 | 34.525 | 1.761 | 863.125 | 17.543 | 17.118 |
| 32 | 23 | 16 | 35.456 | 2.932 | 815.494 | 37.558 | 29.503 |
| 33 | 21 | 5 | 26.820 | 0.145 | 563.220 | -147.069 | 9.710 |
| 34 | 33 | 7 | 27.421 | 0.835 | 904.907 | -211.261 | 102.333 |
| 36 | 25 | 4 | 36.975 | 8.769 | 924.375 | 78.793 | 1150.953 |
| 38 | 38 | 10 | 37.660 | 1.231 | 1431.080 | 145.795 | 130.978 |
| 41 | 30 | 2 | 36.875 | 0.101 | 1106.250 | 91.551 | 42.525 |
| Sum | 628 | | | | 21241.026 | | 6528.159 |
| Variance | | | | | | 9349.2524 | |

$$\begin{aligned}
\hat{\bar{y}}_r &= 33.82 \\
V(\hat{\bar{y}}_r) &= \frac{1}{(27.30)^2}\left[\left(1-\frac{23}{245}\right)\frac{9349.252}{23}+\frac{6528.159}{(23)(245)}\right] \\
&= \frac{1}{745.53}[368.33+1.16] \\
&= 0.50.
\end{aligned}$$

**5.16** (a) Summary quantities for estimating $\bar{y}$ and its variance are given in the table below. Here, $k_i$ denotes the number sampled in school $i$. We use the number of respondents in school $i$ as $m_i$.

| School | $M_i$ | $k_i$ | $m_i$ | Return | $\bar{y}_i$ | $\hat{t}_i$ | $\hat{t}_i - M_i$ | $M_i^2(1-\frac{m_i}{M_i})\frac{s_i^2}{m_i}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 78 | 40 | 38 | 19 | 0.5000 | 39.0000 | -6.1580 | 21.0811 |
| 2 | 238 | 38 | 36 | 19 | 0.5278 | 125.6111 | -12.1786 | 342.3401 |
| 3 | 261 | 19 | 17 | 13 | 0.7647 | 199.5882 | 48.4828 | 716.1696 |
| 4 | 174 | 30 | 30 | 18 | 0.6000 | 104.4000 | 3.6630 | 207.3600 |
| 5 | 236 | 30 | 26 | 12 | 0.4615 | 108.9231 | -27.7087 | 492.6675 |
| 6 | 188 | 25 | 24 | 13 | 0.5417 | 101.8333 | -7.0089 | 332.8031 |
| 7 | 113 | 23 | 22 | 15 | 0.6818 | 77.0455 | 11.6243 | 106.2293 |
| 8 | 170 | 43 | 36 | 21 | 0.5833 | 99.1667 | 0.7455 | 158.1944 |
| 9 | 296 | 38 | 35 | 23 | 0.6571 | 194.5143 | 23.1456 | 511.9485 |
| 10 | 207 | 21 | 17 | 7 | 0.4118 | 85.2353 | -34.6070 | 595.3936 |
| Sum | 1961 | 307 | 281 | 160 | | 1135.3175 | | 3484.1873 |
| var | | | | | | | 581.79702 | |

(b) Using (5.26) and (5.28),

$$\hat{\bar{y}}_r = \frac{1135.3175}{1961} = 0.5789$$

$$\begin{aligned} \hat{V}(\hat{\bar{y}}_r) &= \frac{1}{(196.1)^2}\left[\left(1-\frac{10}{46}\right)\frac{581.797}{10} + \frac{3484.1873}{(10)(46)}\right] \\ &= \frac{1}{(196.1)^2}[45.532 + 7.574] \\ &= 0.001381. \end{aligned}$$

An approximate 95% confidence interval for the percentage of parents who returned the questionnaire is

$$0.5789 \pm 1.96\sqrt{0.001381} = [0.506, 0.652].$$

(c) If the clustering were (incorrectly!) ignored, we would have had $\hat{p} = 160/281 = .569$ with $\hat{V}(\hat{p}) = .569(1-.569)/280 = .000876$.

**5.17** (a) The following table gives summary quantities; the column $\bar{y}_i$ gives the estimated proportion of children who had previously had measles in each school.

| School | $M_i$ | $m_i$ | Hadmeas | $\bar{y}_i$ | $\hat{t}_i$ | $\hat{t}_i - M_i$ | $M_i^2(1-\frac{m_i}{M_i})\frac{s_i^2}{m_i}$ |
|---|---|---|---|---|---|---|---|
| 1 | 78 | 40 | 32 | 0.8000 | 62.4000 | 28.0573 | 12.1600 |
| 2 | 238 | 38 | 10 | 0.2632 | 62.6316 | -42.1576 | 249.4572 |
| 3 | 261 | 19 | 12 | 0.6316 | 164.8421 | 49.9262 | 816.4986 |
| 4 | 174 | 30 | 19 | 0.6333 | 110.2000 | 33.5894 | 200.6400 |
| 5 | 236 | 30 | 16 | 0.5333 | 125.8667 | 21.9581 | 417.2408 |
| 6 | 188 | 25 | 6 | 0.2400 | 45.1200 | -37.6546 | 232.8944 |
| 7 | 113 | 23 | 11 | 0.4783 | 54.0435 | 4.2906 | 115.3497 |
| 8 | 170 | 43 | 23 | 0.5349 | 90.9302 | 16.0808 | 127.8864 |
| 9 | 296 | 38 | 5 | 0.1316 | 38.9474 | -91.3787 | 235.8449 |
| 10 | 207 | 21 | 11 | 0.5238 | 108.4286 | 17.2884 | 480.1837 |
| Sum | 1961 | 307 | 145 | | 863.41 | | 2888.1556 |
| Var | | | | | | 1890.1486 | |

(b)

$$\begin{aligned}\hat{\bar{y}}_r &= \frac{863.41}{1961} = 0.4403 \\ \hat{V}(\hat{\bar{y}}_r) &= \frac{1}{(196.1)^2}\left[\left(1-\frac{10}{46}\right)\frac{1890.15}{10} + \frac{2888.16}{(10)(46)}\right] \\ &= \frac{1}{(196.1)^2}[147.92 + 6.28] \\ &= 0.004\end{aligned}$$

An approximate 95% CI is

$$0.4403 \pm 1.96\sqrt{0.004} = [0.316, 0.564]$$

**5.18** With the different costs, the last two columns of Table 5.4 change. The table now becomes:

| Number of Stems Sampled per Site | $\hat{\bar{y}}$ | SE($\hat{\bar{y}}$) | Cost to Sample One Field | Relative Net Precision |
|---|---|---|---|---|
| 1 | 1.12 | 0.15 | 50 | 0.15 |
| 2 | 1.01 | 0.10 | 70 | 0.14 |
| 3 | 0.96 | 0.08 | 90 | 0.13 |
| 4 | 0.91 | 0.07 | 110 | 0.12 |
| 5 | 0.91 | 0.06 | 130 | 0.12 |

Now the relative net precision is highest when one stem is sampled per site.

**5.19** (a) Here is the sample ANOVA table from SAS PROC GLM:

```
Dependent Variable: acres92

                                 Sum of
Source                DF        Squares   Mean Square  F Value   Pr > F

Model                 41   2.0039161E13  488760018795     8.16   <.0001

Error                258   1.5456926E13   59910564633

Corrected Total      299   3.5496086E13


R-Square     Coeff Var      Root MSE    acres92 Mean

0.564546      82.16474      244766.3        297897.0
```

We estimate $R_a^2$ by

$$1 - \frac{59910564633}{3.5496086 \times 10^{13}/299} = 0.4953455.$$

This value is greater than 0, indicating that there is a clustering effect.

(b) Using (5.32), with $c_1 = 15c_2$ and $R_a^2 = 0.5$, and using the approximation $M(N-1) \approx NM - 1$, we have

$$\bar{m}_{\text{opt}} = \sqrt{\frac{15(1-.5)}{.5}} = 3.9.$$

Taking $\bar{m} = 4$, we would have $n = 300/4 = 75$.

**5.20** Answers will vary.

**5.21** Answers will vary.

**5.22** First note that

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{M}(y_{ij}-\bar{y}_U)(y_{ik}-\bar{y}_U) &= \sum_{i=1}^{N}\left[\sum_{j=1}^{M}(y_{ij}-\bar{y}_U)\right]^2 \\ &= \sum_{i=1}^{N}[M(\bar{y}_{iU}-\bar{y}_U)]^2 \\ &= M(\text{SSB}).\end{aligned}$$

Thus

$$\begin{aligned}\text{ICC} &= \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}^{M}(y_{ij}-\bar{y}_U)(y_{ik}-\bar{y}_U)}{(NM-1)(M-1)S^2} \\ &= \frac{M(\text{SSB})-\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_U)^2}{(NM-1)(M-1)S^2} \\ &= \frac{M(\text{SSB})-\text{SSTO}}{(M-1)\text{SSTO}} \\ &= \frac{M(\text{SSTO}-\text{SSW})-\text{SSTO}}{(M-1)\text{SSTO}} \\ &= 1-\frac{M}{M-1}\,\frac{\text{SSW}}{\text{SSTO}},\end{aligned}$$

proving the result.

**5.23** From (5.8),

$$\text{ICC} = 1-\frac{M}{M-1}\,\frac{\text{SSW}}{\text{SSTO}}.$$

Rewriting, we have

$$\begin{aligned}\text{MSW} &= \frac{1}{N(M-1)}\text{SSTO}\,\frac{M-1}{M}(1-\text{ICC}) \\ &= \frac{NM-1}{NM}S^2(1-\text{ICC}).\end{aligned}$$

Using Table 5.1, we know that SSW + SSB = SSTO, so from (5.8),

$$\begin{aligned} \text{ICC} &= 1 - \frac{M}{M-1}\frac{\text{SSTO} - (N-1)\text{MSB}}{\text{SSTO}} \\ &= -\frac{1}{M-1} + \frac{M(N-1)\text{MSB}}{(M-1)(NM-1)S^2} \end{aligned}$$

and

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

**5.24** (a) From (5.24), we have that

$$\begin{aligned} V(\hat{\bar{y}}_{\text{unb}}) &= \frac{1}{(NM)^2}\left[N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\sum_{i=1}^{N}\left(1-\frac{m}{M}\right)M^2\frac{S_i^2}{m}\right] \\ &= \left(1-\frac{n}{N}\right)\frac{S_t^2}{nM^2} + \left(1-\frac{m}{M}\right)\frac{1}{Nnm}\sum_{i=1}^{N}S_i^2. \end{aligned}$$

The equation above (5.7) states that

$$S_t^2 = M(\text{MSB});$$

the definition of $S_i^2$ in Section 5.1 implies

$$\sum_{i=1}^{N}S_i^2 = \frac{1}{M-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{iU})^2 = \frac{\text{SSW}}{M-1} = N(\text{MSW}).$$

Thus,

$$\begin{aligned} V(\hat{\bar{y}}_{\text{unb}}) &= \frac{1}{(NM)^2}\left[N^2\left(1-\frac{n}{N}\right)\frac{M(\text{MSB})}{n} + \frac{N}{n}\left(1-\frac{m}{M}\right)\frac{M^2}{m}N(\text{MSW})\right] \\ &= \left(1-\frac{n}{N}\right)\frac{\text{MSB}}{nM} + \left(1-\frac{m}{M}\right)\frac{\text{MSW}}{nm}. \end{aligned}$$

(b) The first equality follows directly from (5.9). Because SSTO = SSB + SSW,

$$\begin{aligned} \text{MSB} &= \frac{1}{N-1}[\text{SSTO} - N(M-1)\text{MSW}] \\ &= \frac{1}{N-1}[(NM-1)S^2 - N(M-1)S^2(1-R_a^2)] \\ &= \frac{1}{N-1}[(N-1)S^2 + N(M-1)S^2R_a^2] \\ &= S^2\left[\frac{N(M-1)R_a^2}{N-1}+1\right]. \end{aligned}$$

(c)
$$V(\hat{\bar{y}}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{nM}\left[\frac{N(M-1)R_a^2}{N-1} + 1\right] + \left(1 - \frac{m}{M}\right)\frac{S^2}{nm}(1 - R_a^2).$$

(d) The coefficient of $S^2R_a^2$ in (c) is

$$\left(1 - \frac{n}{N}\right)\frac{N(M-1)}{nM(N-1)} - \left(1 - \frac{m}{M}\right)\frac{1}{nm} = \frac{(N-n)(M-1)m - (M-m)(N-1)}{Mnm(N-1)}$$

But if $N(m-1) > nm$, then

$$(N-n)(M-1)m - (M-m)(N-1) = M[N(m-1) - nm + 1] + m(n-1) > 0,$$

so $V(\hat{\bar{y}})$ is an increasing function of $R_a^2$.

**5.25** (a) If $M_i = M$ for all $i$, and $m_i = m$ for all $i$, then from (5.15),

$$\hat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}} M\bar{y}_i}{nM} = \frac{1}{NM}\,\hat{t}_{\text{unb}} = \hat{\bar{y}}_{\text{unb}}.$$

(b) In the following, let $\hat{\bar{y}} = \hat{\bar{y}}_r = \hat{\bar{y}}_{\text{unb}}$

| Source | df | SS |
|---|---|---|
| Between clusters | $n-1$ | $\displaystyle\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(\bar{y}_i - \hat{\bar{y}})^2$ |
| Within clusters | $n(m-1)$ | $\displaystyle\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(y_{ij} - \bar{y}_i)^2$ |
| Total | $nm-1$ | $\displaystyle\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(y_{ij} - \hat{\bar{y}})^2$ |

(c) Let
$$Z_i = \begin{cases} 1 & \text{if psu } i \text{ in sample} \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned}
\widehat{\text{SSW}} &= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(y_{ij} - \bar{y}_i)^2 \\
E[\widehat{\text{SSW}}] &= E\left[\sum_{i=1}^{N} Z_i \sum_{j\in\mathcal{S}_i}(y_{ij} - \bar{y}_i)^2\right] \\
&= E\left\{\sum_{i=1}^{N} Z_i E\left[\sum_{j\in\mathcal{S}_i}(y_{ij} - \bar{y}_i)^2 \mid \mathbf{Z}\right]\right\} \\
&= E\left\{\sum_{i=1}^{N} Z_i(m-1)E[s_i^2 \mid \mathbf{Z}]\right\} \\
&= E\left\{\sum_{i=1}^{N} Z_i(m-1)S_i^2\right\} \\
&= (m-1)\frac{n}{N}\sum_{i=1}^{N} S_i^2
\end{aligned}$$

Thus,

$$E[\text{msw}] = \frac{1}{N}\sum_{i=1}^{N} S_i^2 = \text{MSW}.$$

Since $M_i = M$ and $m_i = m$ for all $i$,

$$\hat{\bar{y}}_{\text{unb}} = \frac{1}{n}\sum_{i\in\mathcal{S}} \bar{y}_i = \frac{1}{n}\sum_{i=1}^{N} Z_i \bar{y}_i$$

$$\widehat{\text{SSB}} = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} (\bar{y}_i - \hat{\bar{y}}_{\text{unb}})^2 = m\sum_{i\in\mathcal{S}} (\bar{y}_i - \hat{\bar{y}}_{\text{unb}})^2$$

$$
\begin{aligned}
E[\widehat{\mathrm{SSB}}] &= mE\Big[\sum_{i\in\mathcal{S}}(\bar{y}_i^2 - 2\bar{y}_i\hat{\bar{y}}_{\mathrm{unb}} + \hat{\bar{y}}_{\mathrm{unb}}^2)\Big] \\
&= mE\Big[\sum_{i\in\mathcal{S}}\bar{y}_i^2 - n\hat{\bar{y}}_{\mathrm{unb}}^2\Big] \\
&= mE\bigg(E\Big[\sum_{i=1}^{N} Z_i\bar{y}_i^2 \mid Z_1,\ldots,Z_n\Big]\bigg) - mnE[\hat{\bar{y}}_{\mathrm{unb}}^2] \\
&= mE\Big[\sum_{i=1}^{N} Z_i\{V(\bar{y}_i \mid \mathbf{Z}) + \bar{y}_{iU}^2\}\Big] - mn[V(\hat{\bar{y}}_{\mathrm{unb}}) + \bar{y}_U^2] \\
&= mE\Big[\sum_{i=1}^{N} Z_i\Big\{\Big(1-\frac{m}{M}\Big)\frac{S_i^2}{m} + \bar{y}_{iU}^2\Big\}\Big] \\
&\quad -mn\Big[\frac{1}{M^2}\Big(1-\frac{n}{N}\Big)\frac{S_t^2}{n} + \frac{1}{nN}\sum_{i=1}^{N}\Big(1-\frac{m}{M}\Big)\frac{S_i^2}{m} + \bar{y}_U^2\Big] \\
&= \frac{mn}{N}\sum_{i=1}^{N}\Big[\Big(1-\frac{m}{M}\Big)\frac{S_i^2}{m} + \bar{y}_{iU}^2\Big] - \frac{m}{M^2}\Big(1-\frac{n}{N}\Big)S_t^2 \\
&\quad -\frac{m}{N}\sum_{i=1}^{N}\Big(1-\frac{m}{M}\Big)\frac{S_i^2}{m} - mn\bar{y}_U^2 \\
&= \frac{m(n-1)}{N}\Big(1-\frac{m}{M}\Big)\sum_{i=1}^{N}\frac{S_i^2}{m} + mn\Big[\frac{1}{N}\sum_{i=1}^{N}\bar{y}_{iU}^2 - \bar{y}_U^2\Big] \\
&\quad -\frac{m}{M}\Big(1-\frac{n}{N}\Big)(\mathrm{MSB}) \\
&= \frac{m(n-1)}{N}\Big(1-\frac{m}{M}\Big)\sum_{i=1}^{N}\frac{S_i^2}{m} + \frac{mn}{N}\sum_{i=1}^{N}(\bar{y}_{iU} - \bar{y}_U)^2 \\
&\quad -\frac{m}{M}\Big(1-\frac{n}{N}\Big)\frac{1}{N-1}\mathrm{SSB} \\
&= \frac{m(n-1)}{N}\Big(1-\frac{m}{M}\Big)\sum_{i=1}^{N}\frac{S_i^2}{m} + \Big[\frac{mn}{NM} - \frac{m}{M(N-1)}\Big(1-\frac{n}{N}\Big)\Big]\mathrm{SSB} \\
&= \frac{m(n-1)}{N}\Big(1-\frac{m}{M}\Big)\sum_{i=1}^{N}\frac{S_i^2}{m} + \frac{(N-1)mn - m(N-n)}{NM(N-1)}\mathrm{SSB} \\
&= (n-1)\Big(1-\frac{m}{M}\Big)\mathrm{MSW} + (n-1)\frac{m}{M}\mathrm{MSB}.
\end{aligned}
$$

Thus

$$
E[\mathrm{msb}] = \Big(1-\frac{m}{M}\Big)\mathrm{MSW} + \frac{m}{M}\mathrm{MSB}.
$$

(d) From (c),

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

(e) From (5.22) and (5.23),

$$\begin{aligned} s_t^2 &= \frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N}\right)^2 \\ &= \frac{1}{n-1}\sum_{i\in\mathcal{S}}(M\bar{y}_i - M\hat{\bar{y}})^2 \\ &= \frac{1}{n-1}\frac{M^2}{m}\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(\bar{y}_i - \hat{\bar{y}})^2 \\ &= \frac{M^2}{m}\text{msb} \end{aligned}$$

and

$$\begin{aligned} \textstyle\sum_{i\in\mathcal{S}} s_i^2 &= \frac{1}{m-1}\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(y_{ij} - \bar{y}_i)^2 \\ &= n\text{msw}. \end{aligned}$$

Then, using (5.24),

$$\begin{aligned} \hat{V}(\hat{\bar{y}}) &= \frac{1}{(NM)^2}\left[N^2\left(1 - \frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\left(1 - \frac{m}{M}\right)\frac{M^2}{m}\sum_{i\in\mathcal{S}} s_i^2\right] \\ &= \left(1 - \frac{n}{N}\right)\frac{\text{msb}}{nm} + \frac{1}{N}\left(1 - \frac{m}{M}\right)\frac{\text{msw}}{m}. \end{aligned}$$

**5.26** (a) From Exercise 5.25,

$$\text{msto} = \frac{1}{nm-1}\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(y_{ij} - \hat{\bar{y}})^2.$$

Now,

$$
\begin{aligned}
& E\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(y_{ij}-\hat{\hat{y}})^2\right] \\
&= E[\widehat{\mathrm{SSW}}] + E[\widehat{\mathrm{SSB}}] \\
&= (m-1)\frac{n}{N}\sum_{i=1}^{N} S_i^2 + (n-1)\left(1-\frac{m}{M}\right)\mathrm{MSW} + (n-1)\frac{m}{M}\mathrm{MSB} \\
&= \left[(m-1)n + (n-1)\left(1-\frac{m}{M}\right)\right]\mathrm{MSW} + (n-1)\frac{m}{M}\mathrm{MSB} \\
&= \left[nm-1-\frac{m}{M}(n-1)\right]\mathrm{MSW} + (n-1)\frac{m}{M}\mathrm{MSB} \\
&= \frac{1}{N(M-1)}\left[nm-1-\frac{m}{M}(n-1)\right]\mathrm{SSW} + \frac{(n-1)}{N-1}\frac{m}{M}\mathrm{SSB} \\
&= \frac{nm-1}{NM-1}\left[\frac{NM-1}{N(M-1)}\mathrm{SSW} + \frac{(n-1)m}{(N-1)M}\frac{NM-1}{nm-1}\mathrm{SSB}\right] - \frac{m(n-1)}{NM(M-1)}\mathrm{SSW} \\
&= \frac{nm-1}{NM-1}\left[\left\{1+\frac{N-1}{N(M-1)}\right\}\mathrm{SSW}\right. \\
&\qquad \left. + \left\{1+\frac{nm(M-1)-(m-1)NM-M+m}{(N-1)M(nm-1)}\right\}\mathrm{SSB}\right] \\
&\quad -\frac{m(n-1)}{NM(M-1)}\mathrm{SSW} \\
&= \frac{nm-1}{NM-1}\mathrm{SSTO} + \frac{nm-1}{NM-1}\frac{nm(M-1)-(m-1)NM-M+m}{(N-1)M(nm-1)}\mathrm{SSB} \\
&\quad + \left[\frac{nm-1}{NM-1}\frac{N-1}{N(M-1)} - \frac{m(n-1)}{NM(M-1)}\right]\mathrm{SSW} \\
&= \frac{nm-1}{NM-1}\left[\mathrm{SSTO} + O\left(\frac{1}{n}\right)\mathrm{SSB} + O\left(\frac{1}{n}\right)\mathrm{SSW}\right].
\end{aligned}
$$

The notation $O(1/n)$ denotes a term that tends to 0 as $n \to \infty$.

(b) Follows from the last line of (a).

(c) From Exercise 5.25, $E[\mathrm{msw}] = \mathrm{MSW}$ and

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

Consequently,

$$\begin{aligned}
E[\hat{S}^2] &= \frac{M(N-1)}{m(NM-1)}E[\text{msb}] + \frac{(m-1)NM+M-m}{m(NM-1)}E[\text{msw}] \\
&= \frac{M(N-1)}{m(NM-1)}\left[\frac{m}{M}\text{MSB} + \left(1-\frac{m}{M}\right)\text{MSW}\right] \\
&\quad + \frac{(m-1)NM+M-m}{m(NM-1)}\text{MSW} \\
&= \frac{1}{NM-1}\text{SSB} \\
&\quad + \frac{1}{NM-1}\left[(N-1)(M-m)+(m-1)NM+M-m\right]\text{MSW} \\
&= \frac{1}{NM-1}\text{SSB} + \frac{N(M-1)}{NM-1}\text{MSW} \\
&= S^2.
\end{aligned}$$

**5.27** The cost constraint implies that $n = C/(c_1 + c_2 m)$; substituting into (5.30), we have:

$$\begin{aligned}
g(m) &= V(\hat{\bar{y}}_{\text{unb}}) = \frac{(c_1+c_2m)\text{MSB}}{CM} - \frac{\text{MSB}}{NM} + \left(1-\frac{m}{M}\right)\frac{(c_1+c_2m)\text{MSW}}{C_m} \\
\frac{dg}{dm} &= \frac{c_2\text{MSB}}{CM} - \frac{c_1\text{MSW}}{Cm^2} - \frac{c_2\text{MSW}}{CM}.
\end{aligned}$$

Setting the derivative equal to zero and solving for $m$, we have

$$m = \sqrt{\frac{c_1 M(\text{MSW})}{c_2(\text{MSB}-\text{MSW})}}.$$

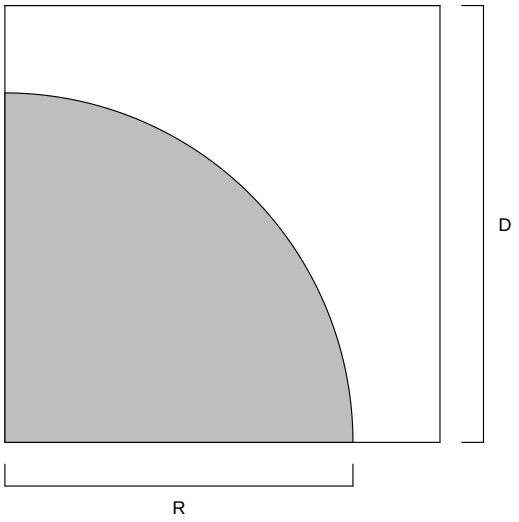
Using Exercise 5.24,

$$m = \sqrt{\frac{c_1 M S^2(1-R_a^2)}{c_2 S^2[N(M-1)R_a^2/(N-1)+1-(1-R_a^2)]}} = \sqrt{\frac{c_1 M(N-1)(1-R_a^2)}{c_2(NM-1)R_a^2}}.$$

**5.28** This exercise does not rely on methods developed in this chapter (other than for a general knowledge of systematic sampling), but represents the type of problem a sampling practitioner might encounter. (A good sampling practitioner must be versatile.)

(a) For all three cases, $P$ (detect the contaminant) = $P$(distance between container and nearest grid point is less than $R$). We can calculate the probability by using simple geometry and trigonometry.

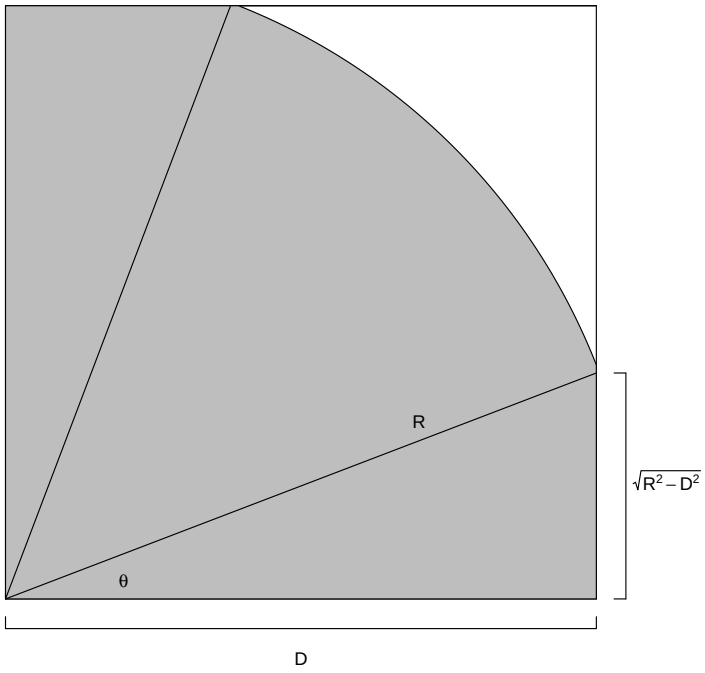
**Case 1:** $R < D$.




Since we assume that the waste container is equally likely to be anywhere in the square relative to the nearest grid point, the probability is the ratio

$$\frac{\text{area of shaded part}}{\text{area of square}} = \frac{\pi R^2}{4D^2}.$$

**Case 2:** $D \leq R \leq \sqrt{2}D$




The probability is again

$$\frac{\text{area of shaded part}}{\text{area of square}}.$$

The area of the shaded part is

$$2\left(\frac{1}{2}D\sqrt{R^2-D^2}\right)+\left(\frac{\pi/2-2\theta}{\pi/2}\right)\left(\frac{\pi R^2}{4}\right)=D\sqrt{R^2-D^2}+\left(1-\frac{4\theta}{\pi}\right)\left(\frac{\pi R^2}{4}\right),$$

where $\cos(\theta)=D/R$.

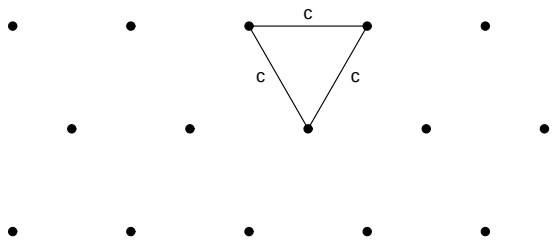
The probability is thus

$$\sqrt{\left(\frac{R}{D}\right)^2-1}+\left(1-\frac{4\theta}{\pi}\right)\left(\frac{\pi R^2}{4D^2}\right).$$

**Case 3:** $R>\sqrt{2}D$.

The probability of detection is 1.

(b) Even though the square grid is commonly used in practice, we can increase the probability of detecting a contaminant by staggering the rows.




**5.30** (a) Because the $A_i$'s and the $\varepsilon_{ij}$'s are independent,

$$\begin{aligned} V_{\mathrm{M1}}\Big[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} b_{ij}Y_{ij}\Big] &= V_{\mathrm{M1}}\Big[\sum_{i\in\mathcal{S}} A_i\Big(\sum_{j\in\mathcal{S}_i} b_{ij}\Big)\Big] + V_{M1}\Big[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} b_{ij}\varepsilon_{ij}\Big] \\ &= \sum_{i\in\mathcal{S}}\Big(\sum_{j\in\mathcal{S}_i} b_{ij}\Big)^2\sigma_A^2 + \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} b_{ij}^2\sigma^2. \end{aligned}$$

(b) Let

$$c_{ij} = \begin{cases} b_{ij}-1 & \text{if } i\in\mathcal{S} \text{ and } j\in\mathcal{S}_i \\ -1 & \text{otherwise.} \end{cases}$$

Then

$$\hat{T} - T = \sum_{i=1}^{N}\sum_{j=1}^{M_i} c_{ij}y_{ij}.$$

Using the same argument as in part (a),

$$\begin{aligned}
V_{\mathrm{M1}}[\hat{T}-T] &= \sum_{i=1}^{N}\left(\sum_{j=1}^{M_i} c_{ij}\right)^2\sigma_A^2 + \sum_{i=1}^{N}\sum_{j=1}^{M_i} c_{ij}^2\sigma^2 \\
&= \sum_{i\in\mathcal{S}}\left[\left(\sum_{j\in\mathcal{S}_i} c_{ij} + \sum_{j\notin\mathcal{S}_i} c_{ij}\right)^2\right]\sigma_A^2 + \sum_{i\notin\mathcal{S}}\left(\sum_{j=1}^{M} c_{ij}\right)^2\sigma_A^2 \\
&\quad + \sum_{i\in\mathcal{S}}\left(\sum_{j\in\mathcal{S}_i} c_{ij}^2 + \sum_{j\notin\mathcal{S}_i} c_{ij}^2\right)\sigma^2 + \sum_{i\notin\mathcal{S}}\sum_{j=1}^{M_i} c_{ij}^2\sigma^2 \\
&= \textstyle\sum_{i\in\mathcal{S}}\left[\left\{\sum_{j\in\mathcal{S}_i}(b_{ij}-1) - (M_i-m_i)\right\}^2\right]\sigma_A^2 + \sum_{i\notin\mathcal{S}} M_i^2\sigma_A^2 \\
&\quad + \textstyle\sum_{i\in\mathcal{S}}\left[\sum_{j\in\mathcal{S}_i}(b_{ij}-1)^2 + (M_i-m_i)\right]\sigma^2 + \sum_{i\notin\mathcal{S}} M_i\sigma^2 \\
&= \textstyle\sum_{i\in\mathcal{S}}\left[\sum_{j\in\mathcal{S}_i} b_{ij} - M_i\right]^2\sigma_A^2 + \sum_{i\notin\mathcal{S}} M_i^2\sigma_A^2 \\
&\quad + \textstyle\sum_{i\in\mathcal{S}}\left[\sum_{j\in\mathcal{S}_i}(b_{ij}^2 - 2b_{ij}) + M_i\right]\sigma^2 + \sum_{i\notin\mathcal{S}} M_i\sigma^2
\end{aligned}$$

**5.32** Recall that

$$\hat{T}_r = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}} b_{ij}y_{ij},$$

with

$$b_{ij} = M_0\frac{M_i}{m_i\sum_{k\in\mathcal{S}} M_k}.$$

Then, from (5.36),

$$\begin{aligned}
V_{M1}[\hat{T}_r - T] &= \sigma_A^2\left[\sum_{i\in\mathcal{S}}\left(\sum_{j\in\mathcal{S}_i}\frac{M_0M_i}{m_i\sum_{k\in\mathcal{S}}M_k} - M_i\right)^2 + \sum_{i\notin\mathcal{S}} M_i^2\right] \\
&\quad + \sigma^2\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}\left(\left(\frac{M_0M_i}{m_i\sum_{k\in\mathcal{S}}M_k}\right)^2 - 2\frac{M_0M_i}{m_i\sum_{k\in\mathcal{S}}M_k}\right) + M_0\right] \\
&= \sigma_A^2\left[\sum_{i\in\mathcal{S}} M_i^2\left(\frac{M_0}{\sum_{k\in\mathcal{S}}M_k} - 1\right)^2 + \sum_{i\notin\mathcal{S}} M_i^2\right] \\
&\quad + \sigma^2\sum_{i\in\mathcal{S}}\left[\frac{M_0^2M_i^2}{m_i(\sum_{k\in\mathcal{S}}M_k)^2} - 2\frac{M_0M_i}{\sum_{k\in\mathcal{S}}M_k}\right] + \sigma^2M_0,
\end{aligned}$$

which is minimized when

$$\sum_{i\in\mathcal{S}} M_i^2/m_i$$

is minimized. Let

$$g(m_1,\ldots,m_N,\lambda) = \sum_{i\in\mathcal{S}}\frac{M_i^2}{m_i} - \lambda\left(i - \sum_{i\in\mathcal{S}} m_i\right).$$

Then

$$\frac{\partial g}{\partial \lambda} = \sum_{i \in \mathcal{S}} m_i - L$$

and, for $k \in \mathcal{S}$,

$$\frac{\partial g}{\partial m_k} = -\frac{M_k^2}{m_k^2} + \lambda.$$

Setting the partial derivatives equal to zero, we have that $M_k/m_k = \sqrt{\lambda}$ is constant for all $k$; that is, $m_k$ is proportional to $M_k$.

# Chapter 6

# Sampling with Unequal Probabilities

**6.2** (a) Instead of creating columns of cumulative $M_i$ range as in Example 6.2, we create columns for the cumulative $\psi_i$ range. Then draw 10 random numbers between 0 and 1 to select the psu's with replacement. A table giving the cumulative $\psi_i$ range for each psu follows:

| psu | $\psi_i$ | Cumulative $\psi_i$ | |
|---|---|---|---|
| 1 | 0.000110 | 0.000000 | 0.000110 |
| 2 | 0.018556 | 0.000110 | 0.018666 |
| 3 | 0.062999 | 0.018666 | 0.081665 |
| 4 | 0.078216 | 0.081665 | 0.159881 |
| 5 | 0.075245 | 0.159881 | 0.235126 |
| 6 | 0.073983 | 0.235126 | 0.309109 |
| 7 | 0.076580 | 0.309109 | 0.385689 |
| 8 | 0.038981 | 0.385689 | 0.424670 |
| 9 | 0.040772 | 0.424670 | 0.465442 |
| 10 | 0.022876 | 0.465442 | 0.488318 |
| 11 | 0.003721 | 0.488318 | 0.492039 |
| 12 | 0.024917 | 0.492039 | 0.516956 |
| 13 | 0.040654 | 0.516956 | 0.557610 |
| 14 | 0.014804 | 0.557610 | 0.572414 |
| 15 | 0.005577 | 0.572414 | 0.577991 |
| 16 | 0.070784 | 0.577991 | 0.648775 |
| 17 | 0.069635 | 0.648775 | 0.718410 |
| 18 | 0.034650 | 0.718410 | 0.753060 |
| 19 | 0.069492 | 0.753060 | 0.822552 |
| 20 | 0.036590 | 0.822552 | 0.859142 |
| 21 | 0.033853 | 0.859142 | 0.892995 |
| 22 | 0.016959 | 0.892995 | 0.909954 |
| 23 | 0.009066 | 0.909954 | 0.919020 |
| 24 | 0.021795 | 0.919020 | 0.940815 |
| 25 | 0.059185 | 0.940815 | 1.000000 |

(Note: the numbers in the "Cumulative $\psi_i$" columns were rounded to fit in the table.)

Ten random numbers I generated between 0 and 1 were:

{0.46242032, 0.34980142, 0.35083063, 0.55868338, 0.62149246,

0.03779992, 0.88290415, 0.99612658, 0.02660724, 0.26350658}.

Using these ten random numbers would result in psu's 9, 7, 7, 14, 16, 3, 21, 25, 3, and 6 being the sample.

(b) Here $\max\{\psi_i\} = 0.078216$. To use Lahiri's method, we select two random numbers for each draw—the first is a random integer between 1 and 25, and the second is a random uniform between 0 and 0.08 (or any other number larger than $\max\{\psi_i\}$). Thus, if our pair of random number is (20, 0.054558), we reject the pair and try again because $0.054 > \psi_{20} = 0.03659$. If the next pair is (8, 0.028979), we include psu 8 in the sample.

**6.3** Calculate $\hat{t}_{\psi\mathcal{S}} = t_i/\psi_i$ for each sample.

| Store | $\psi_i$ | $t_i$ | $\hat{t}_{\psi\mathcal{S}}$ | $(\hat{t}_{\psi\mathcal{S}} - t)^2$ |
|---|---|---|---|---|
| A | 1/16 | 75 | 1200 | 810,000 |
| B | 2/16 | 75 | 600 | 90,000 |
| C | 3/16 | 75 | 400 | 10,000 |
| D | 10/16 | 75 | 120 | 32,400 |

$$\begin{aligned} E[\hat{t}_\psi] &= \frac{1}{16}(1200) + \frac{2}{16}(600) + \frac{3}{16}(400) + \frac{10}{16}(120) \\ &= 300. \\ V[\hat{t}_\psi] &= \frac{1}{16}(810{,}000) + \cdots + \frac{10}{16}(32{,}400) \\ &= 84{,}000. \end{aligned}$$

**6.4**

| Store | $\psi_i$ | $t_i$ | $\hat{t}_{\psi\mathcal{S}}$ | $(\hat{t}_{\psi\mathcal{S}} - t)^2$ |
|---|---|---|---|---|
| A | 7/16 | 11 | 25.14 | 75546.4 |
| B | 3/16 | 20 | 106.67 | 37377.8 |
| C | 3/16 | 24 | 128.00 | 29584.0 |
| D | 3/16 | 245 | 1306.67 | 1013377.8 |

As shown in (6.3) for the general case,

$$\begin{aligned} E[\hat{t}_\psi] &= \frac{7}{16}(25.14) + \frac{3}{16}(106.67) + \frac{3}{16}(128) + \frac{3}{16}(1306.67) \\ &= 300. \end{aligned}$$

Using (6.4),

$$\begin{aligned} V[\hat{t}_\psi] &= \frac{7}{16}(75546.4) + \frac{3}{16}(37377.8) + \frac{3}{16}(29584) + \frac{3}{16}(1013377.8) \\ &= 235615.2. \end{aligned}$$

This is a poor sampling design. Store A, with the smallest sales, is sampled with the largest probability, while Store D is sampled with a smaller probability.

The $\psi_i$ used in this exercise produce a higher variance than simple random sampling.

**6.5** We use (6.5) to calculate $\hat{t}_\psi$ for each sample. So for sample (A,A),

$$\hat{t}_\psi = \frac{1}{2}\left(2\frac{11}{\psi_A}\right) = \frac{11}{\psi_A} = 176.$$

For sample (A,B),

$$\hat{t}_\psi = \frac{1}{2}\left[\frac{11}{\psi_A} + \frac{20}{\psi_B}\right] = \frac{1}{2}[176 + 160] = 168.$$

The results are given in the following table:

| Sample, $\mathcal{S}$ | $P(\mathcal{S})$ | $t_1/\psi_1$ | $t_2/\psi_2$ | $\hat{t}_\psi$ | $P(\mathcal{S})(\hat{t}-t)^2$ |
|---|---|---|---|---|---|
| (A,A) | 1/256 | 176 | 176 | 176 | 60.06 |
| (A,B) | 2/256 | 176 | 160 | 168 | 136.13 |
| (A,C) | 3/256 | 176 | 128 | 152 | 256.69 |
| (A,D) | 10/256 | 176 | 392 | 284 | 10.00 |
| (B,A) | 2/256 | 160 | 176 | 168 | 136.13 |
| (B,B) | 4/256 | 160 | 160 | 160 | 306.25 |
| (B,C) | 6/256 | 160 | 128 | 144 | 570.38 |
| (B,D) | 20/256 | 160 | 392 | 276 | 45.00 |
| (C,A) | 3/256 | 128 | 176 | 152 | 256.69 |
| (C,B) | 6/256 | 128 | 160 | 144 | 570.38 |
| (C,C) | 9/256 | 128 | 128 | 128 | 1040.06 |
| (C,D) | 30/256 | 128 | 392 | 260 | 187.50 |
| (D,A) | 10/256 | 392 | 176 | 284 | 10.00 |
| (D,B) | 20/256 | 392 | 160 | 276 | 45.00 |
| (D,C) | 30/256 | 392 | 128 | 260 | 187.50 |
| (D,D) | 100/256 | 392 | 392 | 392 | 3306.25 |
| Total | 1 | | | | 7124.00 |

$$\begin{aligned} E[\hat{t}_\psi] &= \frac{1}{256}(176) + \cdots + \frac{100}{256}(392) = 300 \\ V[\hat{t}_\psi] &= \frac{1}{256}(176-300)^2 + \cdots + \frac{100}{256}(392-300)^2 = 7124. \end{aligned}$$

Of course, an easier solution is to note that (6.5) and (6.6) imply that $E[\hat{t}_\psi] = t$, and that $V[\hat{t}_\psi]$ will be half of the variance found when taking a sample of one psu in Section 6.2; i.e., $V[\hat{t}_\psi] = 14248/2 = 7124$.

**6.6** (a) The following table does the calculations, using (6.4) to find the variance.

| name | $t_i$ | $\psi_i$ | $\hat{t}_\psi = \frac{t_i}{\psi_i}$ | $\psi_i(\hat{t}_\psi - t)^2$ | $\hat{t}_{\text{SRS}}$ | $\frac{1}{13}(\hat{t}_{\text{SRS}} - t)^2$ |
|---|---|---|---|---|---|---|
| Apache | 31621 | 0.0572 | 553292.1 | 20477223 | 411073 | 1997590608 |
| Cochise | 51126 | 0.0969 | 527405.6 | 194693778 | 664638 | 656992453 |
| Coconino | 53443 | 0.0958 | 558108.6 | 19071085 | 694759 | 1155043188 |
| Gila | 28189 | 0.0423 | 667034.6 | 379902891 | 366457 | 3256832592 |
| Graham | 11430 | 0.0276 | 414597.1 | 684957758 | 148590 | 13804863397 |
| Greenlee | 3744 | 0.0070 | 532113.6 | 11318255 | 48672 | 21084888877 |
| La Paz | 15133 | 0.0162 | 932417.7 | 2105687838 | 196729 | 10845710928 |
| Mohave | 80062 | 0.1276 | 627317.4 | 387423351 | 1040806 | 16890146325 |
| Navajo | 47413 | 0.0802 | 590892.8 | 27974547 | 616369 | 149926608 |
| Pinal | 81154 | 0.1480 | 548502.8 | 83232656 | 1055002 | 17929037997 |
| Santa Cruz | 13036 | 0.0316 | 412582.0 | 805214838 | 169468 | 12477690693 |
| Yavapai | 81730 | 0.1379 | 592659.0 | 57604012 | 1062490 | 18489514797 |
| Yuma | 74140 | 0.1317 | 562787.3 | 11723897 | 963820 | 11796136677 |
| Sum | 572221 | 1.0000 | | 4789282131 | | 1.30534E+11 |

The $\hat{t}_\psi$'s are given in column 5, and $V(\hat{t}_\psi)$ =4,789,282,131.

(b) Using $\psi_i = 1/13$ for each $i$, we get the SRS estimators in column 7 with

$V(\hat{t}_\psi)$ =130,534,375,140.

The unequal-probability sample is more efficient because $t_i$ and $M_i$ are highly correlated: the correlation is 0.9905. This means that the quantity $t_i/\psi_i$ does not vary much from sample to sample.

**6.7** From (6.6),

$$\hat{V}(\hat{t}_\psi) = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}\left(\frac{t_i}{\psi_i} - \hat{t}_\psi\right)^2.$$

In an SRSWR, $\psi_i = 1/N$, so we have $\hat{t}_\psi = N\bar{t}$ and

$$\hat{V}(\hat{t}_\psi) = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}(Nt_i - N\bar{t})^2 = \frac{N^2}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}(t_i - \bar{t})^2.$$

**6.8** We use (6.13) and (6.14), along with the following calculations from a spreadsheet.

| Academic Unit | $M_i$ | $\psi_i$ | $y_{ij}$ | $\bar{y}_i$ | $\hat{t}_i$ | $\hat{t}_i/\psi_i$ |
|---|---|---|---|---|---|---|
| 14 | 65 | 0.0805452 | 3 0 0 4 | 1.75 | 113.75 | 1412.25 |
| 23 | 25 | 0.0309789 | 2 1 2 0 | 1.25 | 31.25 | 1008.75 |
| 9 | 48 | 0.0594796 | 0 0 1 0 | 0.25 | 12.00 | 201.75 |
| 14 | 65 | 0.0805452 | 2 0 1 0 | 0.75 | 48.75 | 605.25 |
| 16 | 2 | 0.0024783 | 2 0 | 1.00 | 2.00 | 807.00 |
| 6 | 62 | 0.0768278 | 0 2 2 5 | 2.25 | 139.50 | 1815.75 |
| 14 | 65 | 0.0805452 | 1 0 0 3 | 1.00 | 65.00 | 807.00 |
| 19 | 62 | 0.0768278 | 4 1 0 0 | 1.25 | 77.50 | 1008.75 |
| 21 | 61 | 0.0755886 | 2 2 3 1 | 2.00 | 122.00 | 1614.00 |
| 11 | 41 | 0.0508055 | 2 5 12 3 | 5.50 | 225.50 | 4438.50 |
| | | | | | | |
| | | | average | | | 1371.90 |
| | | | std. dev. | | | 1179.47 |

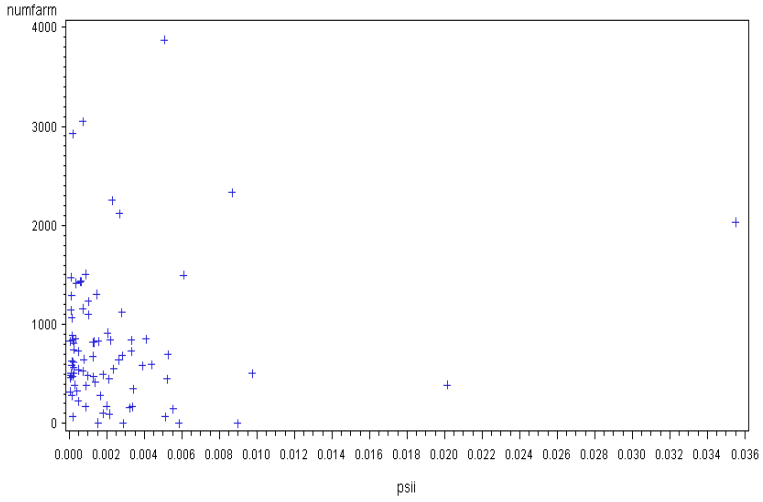
Thus $\hat{t}_\psi = 1371.90$ and $\mathrm{SE}(\hat{t}_\psi) = (1/\sqrt{10})(1179.47) = 372.98$.

Here is SAS code for calculating these estimates. Note that unit 14 appears 3 times; in SAS, you have to give each of these repetitions a different unit number. Otherwise, SAS will just put all of the observations in the same psu for calculations.

```
data faculty;
   input unit $ Mi psi y1 y2 y3 y4;
   array yarray{4} y1 y2 y3 y4;
   sampwt = (1/( 10*psi))*(Mi/4);
   if unit = 16 then sampwt =  (1/( 10*psi));
   do i = 1 to 4;
      y  = yarray{i};
```

```
  if y ne . then output;
end;
   datalines;
   /* Note: label the 3 unit 14s with different psu numbers */
14a 65 0.0805452 3 0 0 4
23 25 0.0309789 2 1 2 0
9 48 0.0594796 0 0 1 0
14b 65 0.0805452 2 0 1 0
16 2 0.0024783 2 0 .   .
6 62 0.0768278 0 2 2 5
14c 65 0.0805452 1 0 0 3
19 62 0.0768278 4 1 0 0
21 61 0.0755886 2 2 3 1
11 41 0.0508055 2 5 12 3
;

proc surveymeans data=faculty mean clm sum clsum;
   cluster unit;
   weight sampwt;
   var y;
run;
```

**6.12** (a) The correlation between $\psi_i$ and $t_i$ (= number of farms in county) is 0.26. We expect some benefit from pps sampling, but not a great deal—sampling with probability proportional to population works better for quantities highly correlated with population, such as number of physicians as in Example 6.5.




(b) As in Example 6.5, we form a new column $t_i/\psi_i$. The mean of this column is

1,896,300, and the standard deviation of the column is 3,674,225. Thus

$$\hat{t}_\psi = 1{,}896{,}300$$

and

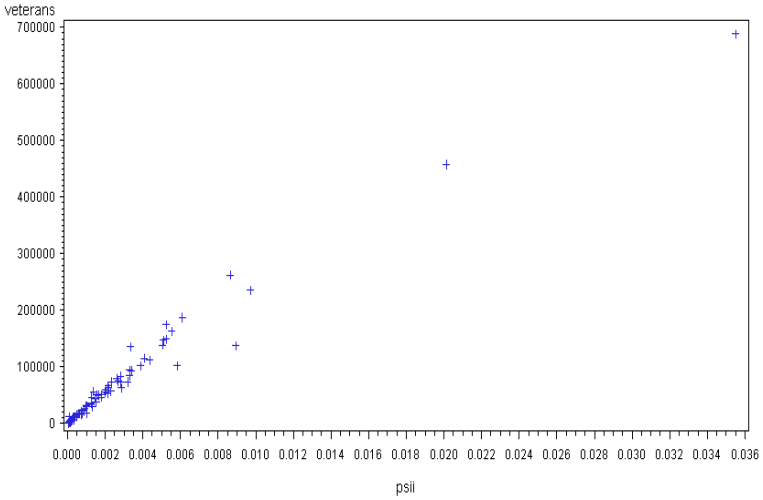
$$\text{SE}\,[\hat{t}_\psi] = \frac{3{,}674{,}225}{\sqrt{100}} = 367{,}423.$$

A histogram of the $t_i/\psi_i$ exhibits strong skewness, however, so a confidence interval using the normal approximation may not be appropriate.

Here is SAS code for producing these estimates. Note that we do not use the cluster statement in proc surveymeans since the observations are psu totals.

```
data statepop;
   infile statepop delimiter=',' firstobs=2;
   input state $ county $ landarea popn physicns farmpop
        numfarm farmacre veterans percviet;
   psii = popn/255077536;
   wt = 1/(100*psii);  /* weight = 1/(n \psi_i) */

/* Be careful when constructing your dataset that if a psu is
   selected multiple times, then it appears that many times
   in the data.  Otherwise, you will get the wrong estimate.*/
proc print data=statepop;
proc corr data=statepop;

proc gplot data=statepop;
   plot numfarm*psii;
run;

/*Omit the 'total' option because sampling is with replacement*/

proc surveymeans data=statepop nobs mean sum clm clsum;
   var numfarm;
   weight wt;
run;
```

**6.13** (a) Corr (population, number of veterans) = .99. We expect the unequal probability sampling to be very efficient here.




(b)

$$\hat{t}_\psi = \frac{1}{100}\sum \frac{t_i}{\psi_i} = 27{,}914{,}180$$
$$\mathrm{SE}\,[\hat{t}_\psi] = \frac{10874532}{\sqrt{100}} = 1{,}087{,}453$$

Note that Allen County, Ohio appears to be an outlier among the $t_i/\psi_i$: it has population 26,405 and 12,642 veterans.

(c) For each county,

$$\mathrm{Vietvet}_i = \mathrm{veterans}_i * \mathrm{percviet}_i/100.$$

We then form a column with $i$th entry $\mathrm{Vietvet}_i/\psi_i$, and find the mean (= 8050477) and standard deviation (= 3273372) of that column. Then

$$\begin{aligned} \hat{t}_\psi &= 8{,}050{,}477 \\ \mathrm{SE}\,[\hat{t}_\psi] &= 327{,}337 \end{aligned}$$

Here is SAS code for calculating these estimates:

```
data statepop;
   infile statepop delimiter=',' firstobs=2;
   input state $ county $ landarea popn physicns farmpop
         numfarm farmacre veterans percviet;
   vietvet = veterans*percviet/100;
   psii = popn/255077536;
   wt = 1/(100*psii);  /* weight = 1/(n \psi_i) */

proc print data=statepop; run;
```

```
proc corr data=statepop; run;

proc gplot data=statepop;
   plot veterans*psii;

proc surveymeans data=statepop nobs mean sum clm clsum;
   var veterans vietvet;
   weight wt;
run;
```

**6.14** (a) We use (6.28) and (6.29) to calculate the variances. We have

| Class | $\hat{t}_i$ | $\hat{V}(\hat{t}_i)$ |
|---|---|---|
| 4 | 110.00 | 16.50 |
| 10 | 106.25 | 185.94 |
| 1 | 154.00 | 733.33 |
| 9 | 195.75 | 2854.69 |
| 14 | 200.00 | 1200.00 |

From Table 6.7, and calculating

$$\hat{V}(\hat{t}_i) = M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{s_i^2}{m_i},$$

we have the second term in (6.28) and (6.29) is

$$\sum_{i \in \mathcal{S}} \frac{\hat{V}(\hat{t}_i)}{\pi_i} = 11355.$$

To calculate the first term of (6.28), note that if we set $\pi_{ii} = \pi_i$, we can write

$$\sum_{i \in \mathcal{S}} (1 - \pi_i) \frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i \in \mathcal{S}} \sum_{\substack{k \in \mathcal{S} \\ k \neq i}} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik}} \frac{\hat{t}_i}{\pi_i} \frac{\hat{t}_k}{\pi_k} = \sum_{i \in \mathcal{S}} \sum_{k \in \mathcal{S}} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik}} \frac{\hat{t}_i}{\pi_i} \frac{\hat{t}_k}{\pi_k}.$$

We obtain that the first term of (6.28) is

$$\sum_{i \in \mathcal{S}} \sum_{k \in \mathcal{S}} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik}} \frac{\hat{t}_i}{\pi_i} \frac{\hat{t}_k}{\pi_k} = 6059.6,$$

so

$$\hat{V}_{\mathrm{HT}}(\hat{t}_{\mathrm{HT}}) = 6059.6 + 11355 = 17414.46.$$

Similarly,

$$\hat{V}_{\mathrm{SYG}}(\hat{t}_{\mathrm{HT}}) = 6059.6 + 11355 = 66139.41.$$

Note how widely these values differ because of the instability of the estimators with this small sample size ($n = 5$).

(b) Here is the pseudo-fpc calculated by SAS.

```
                              Statistics

                                    Std Error
Variable            Mean              of Mean      95% CL for Mean
-------------------------------------------------------------------
y               3.450000             0.393436   2.35764732 4.54235268
-------------------------------------------------------------------

                    Statistics

Variable            Sum        Std Dev          95% CL for Sum
-----------------------------------------------------------------
y           2232.150000     254.552912     1525.39781 2938.90219
-----------------------------------------------------------------
```

This standard error is actually pretty close to the SYG SE of 257.

**6.15** (a) Let $J$ be an integer greater than $\max\{M_i\}$. Let $U_1, U_2, \ldots$, be independent discrete uniform $\{1, \ldots, N\}$ random variables, and let $V_1, V_2, \ldots$ be independent discrete uniform $\{1, \ldots, J\}$ random variables. Assume that all $U_i$ and $V_j$ are independent. Then, on any given iteration of the procedure,

$$
\begin{aligned}
P(\text{select psu } i) &= P(\text{select psu } i \text{ with first pair of random numbers}) \\
&\quad + P(\text{select psu } i \text{ with second pair}) \\
&\quad + \cdots \\
&= P(U_1 = i \quad \text{and} \quad V_1 \le M_i) \\
&\quad + P(U_2 = i, V_2 \le M_i) P\Big( \bigcup_{j=1}^{N} \{U_1 = j, V_2 > M_j\} \Big) \\
&\quad + \cdots + P(U_k = i, V_k \le M_i) \prod_{l=1}^{k-1} P\Big( \bigcup_{j=1}^{N} \{U_l = j, V_l > M_j\} \Big) \\
&= \frac{1}{N}\frac{M_i}{J} + \frac{1}{N}\frac{M_i}{J}\left[\frac{1}{N}\sum_{j=1}^{N}\frac{J - M_j}{J}\right] \\
&\quad + \cdots + \frac{1}{N}\frac{M_i}{J}\left[\frac{1}{N}\sum_{j=1}^{N}\frac{J - M_j}{J}\right]^{k-1} + \cdots \\
&= \frac{1}{N}\frac{M_i}{J}\sum_{k=0}^{\infty}\left[\frac{1}{N}\sum_{j=1}^{N}\frac{J - M_j}{J}\right]^{k} \\
&= \frac{1}{N}\frac{M_i}{J}\frac{1}{1 - \sum_{j=1}^{N}(J - M_j)/(JN)} \\
&= M_i \Big/ \sum_{j=1}^{N} M_j.
\end{aligned}
$$

(b) Let $W$ represent the number of pairs of random numbers that must be generated

to obtain the first valid psu. Since sampling is done with replacement, and hence all psu's are selected independently, we will have that $E[X] = nE[W]$. But $W$ has a geometric distribution with success probability

$$p = P(U_1 = i, V_1 \leq M_i \text{ for some } i) = \sum_{i=1}^{N} \frac{1}{N} \frac{M_i}{J}.$$

Then

$$P(W = k) = (1-p)^{k-1} p$$

and

$$E[W] = \frac{1}{p} = NJ \Big/ \sum_{i=1}^{N} M_i.$$

Hence,

$$E[X] = nNJ \Big/ \sum_{i=1}^{N} M_i.$$

**6.16** The random variables $Q_1, \ldots, Q_N$ have a joint multinomial distribution with probabilities $\psi_1, \ldots, \psi_N$. Consequently,

$$\begin{aligned} \textstyle\sum_{i=1}^{n} Q_i &= n, \\ E[Q_i] &= n\psi_i, \\ V[Q_i] &= n\psi_i(1-\psi_i), \end{aligned}$$

and

$$\text{Cov}\,(Q_i, Q_k) = -n\psi_i\psi_k \quad \text{for } i \neq k.$$

(a) Using successive conditioning,

$$\begin{aligned} E[\hat{t}_\psi] &= E\left[\frac{1}{n} \sum_{i=1}^{N} \sum_{j=1}^{Q_i} \frac{\hat{t}_{ij}}{\psi_i}\right] \\ &= E\left\{ E\left[\frac{1}{n} \sum_{i=1}^{N} \sum_{j=1}^{Q_i} \frac{\hat{t}_{ij}}{\psi_i} \,\middle|\, Q_1, \ldots, Q_N\right]\right\} \\ &= E\left[\frac{1}{n} \sum_{i=1}^{N} Q_i \frac{t_i}{\psi_i}\right] \\ &= \frac{1}{n} \sum_{i=1}^{N} n\psi_i \frac{t_i}{\psi_i} = t. \end{aligned}$$

Thus $\hat{t}_\psi$ is unbiased for $t$.

(b) To find $V(\hat{t}_\psi)$, note that

$$\begin{aligned} V(\hat{t}_\psi) &= V[E(\hat{t}_\psi \mid Q_1, \ldots, Q_N)] + E[V(\hat{t}_\psi \mid Q_1, \ldots, Q_N)] \\ &= V\left[\frac{1}{n} \sum_{i=1}^{N} Q_i \frac{t_i}{\psi_i}\right] + E\left[V\left(\frac{1}{n} \sum_{i=1}^{N} \sum_{j=1}^{Q_i} \frac{\hat{t}_{ij}}{\psi_i} \,\middle|\, Q_1, \ldots, Q_N\right)\right]. \end{aligned}$$

Looking at the two terms separately,

$$\begin{aligned}
V\left[\frac{1}{n}\sum_{i=1}^{N} Q_i \frac{t_i}{\psi_i}\right] &= \mathrm{Cov}\left[\frac{1}{n}\sum_{i=1}^{N} Q_i \frac{t_i}{\psi_i}, \frac{1}{n}\sum_{k=1}^{N} Q_k \frac{t_k}{\psi_k}\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{N}\sum_{k=1}^{N} \frac{t_i}{\psi_i}\frac{t_k}{\psi_k}\ \mathrm{Cov}\left(Q_i, Q_k\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{N}\left(\frac{t_i}{\psi_i}\right)^2 n\psi_i(1-\psi_i) \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{N}\sum_{k=1,k\neq i}^{N} \frac{t_i}{\psi_i}\frac{t_k}{\psi_k}(-n\psi_i\psi_k) \\
&= \frac{1}{n}\sum_{i=1}^{N}\left(\frac{t_i}{\psi_i}\right)^2 [\psi_i(1-\psi_i)+\psi_i^2] \\
&\quad - \frac{1}{n}\sum_{i=1}^{N}\sum_{k=1}^{N} t_i t_k \\
&= \frac{1}{n}\sum_{i=1}^{N}\psi_i\left(\frac{t_i}{\psi_i} - t\right)^2.
\end{aligned}$$

$$\begin{aligned}
E\left[V\left(\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{\hat{t}_{ij}}{\psi_i} \mid Q_1,\ldots,Q_N\right)\right] &= E\left[\frac{1}{n^2}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{V(\hat{t}_{ij})}{\psi_i^2}\right] \\
&= E\left[\frac{1}{n^2}\sum_{i=1}^{N} Q_i\frac{V_i}{\psi_i^2}\right] \\
&= \frac{1}{n}\sum_{i=1}^{N}\frac{V_i}{\psi_i}.
\end{aligned}$$

This equality uses the assumptions that $V(\hat{t}_{ij}) = V_i$ for any $j$, and that the estimates $\hat{t}_{ij}$ are independent. Thus,

$$V[\hat{t}_\psi] = \frac{1}{n}\sum_{i=1}^{N}\psi_i\left(\frac{t_i}{\psi_i} - t\right)^2 + \frac{1}{n}\sum_{i=1}^{N}\frac{V_i}{\psi_i}.$$

(c) To show that (6.9) is an unbiased estimator of the variance, note that

$$\begin{aligned}
E[\hat{V}(\hat{t}_\psi)] &= E\left[\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{(\hat{t}_{ij}/\psi_i - \hat{t}_\psi)^2}{n-1}\right] \\
&= E\left\{\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{1}{n-1}\left[\left(\frac{\hat{t}_{ij}}{\psi_i}\right)^2 - 2\frac{\hat{t}_{ij}}{\psi_i}\hat{t}_\psi + \hat{t}_\psi^2\right]\right\} \\
&= E\left[\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{1}{n-1}\left(\frac{\hat{t}_{ij}}{\psi_i}\right)^2 - \frac{\hat{t}_\psi^2}{n-1}\right] \\
&= E\left\{\frac{1}{n}E\left[\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{1}{n-1}\left(\frac{\hat{t}_{ij}}{\psi_i}\right)^2\middle| Q_1,\ldots,Q_N\right]\right\} - \frac{1}{n-1}[t^2 + V(\hat{t}_\psi)] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{N}\frac{Q_i}{n-1}\frac{t_i^2+V_i}{\psi_i^2}\right] - \frac{1}{n-1}[t^2 + V(\hat{t}_\psi)] \\
&= \frac{1}{n-1}\sum_{i=1}^{N}\frac{t_i^2+V_i}{\psi_i} - \frac{1}{n-1}[t^2 + V(\hat{t}_\psi)] \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{N}\psi_i\frac{t_i^2}{\psi_i^2} - t^2\right) + \frac{1}{n-1}\left(\sum_{i=1}^{N}\frac{V_i}{\psi_i} - V(\hat{t}_\psi)\right) \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{N}\psi_i\left(\frac{t_i}{\psi_i} - t\right)^2 + \sum_{i=1}^{N}\frac{V_i}{\psi_i}\right] - \frac{1}{n-1}V(\hat{t}_\psi) \\
&= \frac{n}{n-1}V(\hat{t}_\psi) - \frac{1}{n-1}V(\hat{t}_\psi) = V(\hat{t}_\psi).
\end{aligned}$$

**6.17** It is sufficient to show that (6.22) and (6.23) are equivalent. When an SRS of psus is selected, then $\pi_i = n/N$ and $\pi_{ik} = \dfrac{n(n-1)}{N(N-1)}$ for all $i$ and $k$. So, starting with the SYG form,

$$\begin{aligned}
&\frac{1}{2}\sum_{i\in\mathcal{S}}\sum_{\substack{k\in\mathcal{S}\\k\neq i}}\frac{\pi_i\pi_k - \pi_{ik}}{\pi_{ik}}\left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k}\right)^2 \\
&= \frac{1}{2}\frac{1}{\pi_1^2}\frac{\pi_1^2 - \pi_{12}}{\pi_{12}}\sum_{i\in\mathcal{S}}\sum_{\substack{k\in\mathcal{S}\\k\neq i}}(t_i - t_k)^2 \\
&= \frac{1}{2}\frac{1}{\pi_1^2}\frac{\pi_1^2 - \pi_{12}}{\pi_{12}}\sum_{i\in\mathcal{S}}\sum_{\substack{k\in\mathcal{S}\\k\neq i}}(t_i^2 + t_k^2 - 2t_it_k) \\
&= \frac{1}{\pi_1^2}\frac{\pi_1^2 - \pi_{12}}{\pi_{12}}\sum_{i\in\mathcal{S}}(n-1)t_i^2 - \frac{1}{\pi_1^2}\frac{\pi_1^2 - \pi_{12}}{\pi_{12}}\sum_{i\in\mathcal{S}}\sum_{\substack{k\in\mathcal{S}\\k\neq i}}t_it_k.
\end{aligned}$$

But in an SRS,

$$\frac{1}{\pi_1^2}\frac{\pi_1^2-\pi_{12}}{\pi_{12}}(n-1)=\left(\frac{N}{n}\right)^2\frac{\frac{n}{N}-\frac{n-1}{N-1}}{\frac{n-1}{N-1}}(n-1)=\left(\frac{N}{n}\right)^2\left(1-\frac{n}{N}\right),$$

which proves the result.

**6.18** To show the results from stratified sampling, we treat the $H$ strata as the $N$ psus. Note that since all strata are subsampled, we have $\pi_i = 1$ for each stratum. Thus, (6.25) becomes

$$\hat{t}_{\mathrm{HT}}=\sum_{i=1}^{H}\hat{t}_i=\sum_{i=1}^{H}N_i\bar{y}_i.$$

For (3.3), since all strata are sampled, either (6.26) or (6.27) gives

$$V(\hat{t}_{\mathrm{HT}})=\sum_{i=1}^{H}V(\hat{t}_i).$$

Result (3.4) follows from (6.29) similarly.

**6.19** (a) We use the method on page 239 to calculate the probabilities.

| | $\pi_{jk}$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Unit | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | $\pi_i$ |
| 1 | — | 0.049 | 0.080 | 0.088 | 0.007 | 0.021 | 0.080 | 0.021 | 0.035 | 0.381 |
| 2 | 0.049 | — | 0.041 | 0.045 | 0.003 | 0.010 | 0.041 | 0.010 | 0.018 | 0.218 |
| 3 | 0.080 | 0.041 | — | 0.073 | 0.006 | 0.017 | 0.067 | 0.017 | 0.029 | 0.330 |
| 4 | 0.088 | 0.045 | 0.073 | — | 0.006 | 0.019 | 0.073 | 0.019 | 0.032 | 0.356 |
| 5 | 0.007 | 0.003 | 0.006 | 0.006 | — | 0.001 | 0.006 | 0.001 | 0.002 | 0.033 |
| 6 | 0.021 | 0.010 | 0.017 | 0.019 | 0.001 | — | 0.017 | 0.004 | 0.007 | 0.097 |
| 7 | 0.080 | 0.041 | 0.067 | 0.073 | 0.006 | 0.017 | — | 0.017 | 0.029 | 0.330 |
| 8 | 0.021 | 0.010 | 0.017 | 0.019 | 0.001 | 0.004 | 0.017 | — | 0.007 | 0.097 |
| 9 | 0.035 | 0.018 | 0.029 | 0.032 | 0.002 | 0.007 | 0.029 | 0.007 | — | 0.159 |
| $\pi_i$ | 0.381 | 0.218 | 0.330 | 0.356 | 0.033 | 0.097 | 0.330 | 0.097 | 0.159 | 2.000 |

(b) Using (6.21), $V(\hat{t}_{\mathrm{HT}}) = 1643$. Using (6.46) (we only need the first term), $V_{WR}(\hat{t}_\psi) = 1867$.

**6.20** (a) We write

$$\begin{aligned}
\text{Cov}\,(\hat{t}_x, \hat{t}_y) &= \text{Cov}\left(\sum_{i=1}^{N} Z_i \frac{t_{ix}}{\pi_i}, \sum_{k=1}^{N} Z_k \frac{t_{ky}}{\pi_k}\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{N} \frac{t_{ix}}{\pi_i}\frac{t_{ky}}{\pi_k}\text{Cov}\ (Z_i, Z_k) \\
&= \sum_{i=1}^{N} \frac{t_{ix}}{\pi_i}\frac{t_{iy}}{\pi_i}\pi_i(1-\pi_i) + \sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} \frac{t_{ix}}{\pi_i}\frac{t_{ky}}{\pi_k}(\pi_{ik} - \pi_i\pi_k) \\
&= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_{ix}t_{iy} + \sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} (\pi_{ik} - \pi_i\pi_k)\frac{t_{ix}}{\pi_i}\frac{t_{ky}}{\pi_k}.
\end{aligned}$$

(b) If the design is an SRS,

$$\begin{aligned}
\text{Cov}\,(\hat{t}_x, \hat{t}_y) &= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_{ix}t_{iy} + \sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} (\pi_{ik} - \pi_i\pi_k)\frac{t_{ix}}{\pi_i}\frac{t_{ky}}{\pi_k} \\
&= \sum_{i=1}^{N} \frac{N}{n}\left(1-\frac{n}{N}\right) t_{ix}t_{iy} \\
&\quad + \left(\frac{N}{n}\right)^2 \sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} \left[\frac{n}{N}\frac{(n-1)}{(N-1)} - \left(\frac{n}{N}\right)^2\right] t_{ix}t_{iy} \\
&= \sum_{i=1}^{N} \frac{N}{n}\left(1-\frac{n}{N}\right) t_{ix}t_{iy} + \frac{N}{n}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} \left[\frac{(n-1)}{(N-1)} - \frac{n}{N}\right] t_{ix}t_{ky} \\
&= \frac{N}{n}\left(1-\frac{n}{N}\right)\sum_{i=1}^{N} t_{ix}t_{iy} - \frac{N}{n}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} \frac{N-n}{N(N-1)} t_{ix}t_{ky} \\
&= \frac{N}{n}\left(1-\frac{n}{N}\right)\left[\sum_{i=1}^{N} t_{ix}t_{iy} - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} t_{ix}t_{ky}\right] \\
&= \frac{N}{n}\left(1-\frac{n}{N}\right)\left[\sum_{i=1}^{N} t_{ix}t_{iy} - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{k=1}^{N} t_{ix}t_{ky} + \frac{1}{N-1}\sum_{i=1}^{N} t_{ix}t_{iy}\right] \\
&= \frac{N}{n}\left(1-\frac{n}{N}\right)\left[\frac{N}{N-1}\sum_{i=1}^{N} t_{ix}t_{iy} - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{k=1}^{N} t_{ix}t_{ky}\right] \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[\sum_{i=1}^{N} t_{ix}t_{iy} - \frac{t_x t_y}{N}\right]
\end{aligned}$$

**6.21** Note that

$$\hat{\bar{y}} = \frac{\sum_{i\in\mathcal{S}}\sum_{j=1}^{M} w_{ij}y_{ij}}{\sum_{i\in\mathcal{S}}\sum_{j=1}^{M} w_{ij}} = \frac{\hat{t}_y}{NM}$$

Thus,

$$\begin{aligned}
&(NM)^2\text{Cov}\left[\left(\hat{\bar{u}} - \frac{t_u}{t_x}\hat{\bar{x}}\right),\ \left\{\hat{\bar{y}} - \hat{\bar{u}} - \frac{t_y - t_u}{N - t_x}(1 - \hat{\bar{x}})\right\}\right] \\
&= \text{Cov}\left[\left(\hat{t}_u - \frac{t_u}{t_x}\hat{t}_x\right),\ \left\{\hat{t}_y - \hat{t}_u - \frac{t_y - t_u}{N - t_x}(NM - \hat{t}_x)\right\}\right] \\
&= \text{Cov}\left[\hat{t}_u, \hat{t}_y - \hat{t}_u + \frac{t_y - t_u}{N - t_x}\hat{t}_x\right] - \frac{t_u}{t_x}\text{Cov}\left[\hat{t}_x, \hat{t}_y - \hat{t}_u + \frac{t_y - t_u}{N - t_x}\hat{t}_x\right] \\
&= \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}\left[\sum_{i=1}^{N}\left\{t_{iu}t_{iy} - t_{iu}^2 + \frac{t_y - t_u}{N - t_x}t_{iu}t_{ix}\right.\right. \\
&\qquad\qquad \left.\left. - \frac{t_u}{t_x}\left(t_{ix}t_{iy} - t_{ix}t_{iu} + \frac{t_y - t_u}{N - t_x}t_{ix}^2\right)\right\}\right] \\
&\quad -\frac{N}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}\left[t_u t_y - t_u^2 + \frac{t_y - t_u}{N - t_x}t_u t_x\right. \\
&\qquad\qquad \left. - \frac{t_u}{t_x}\left(t_x t_y - t_x t_u + \frac{t_y - t_u}{N - t_x}t_x^2\right)\right] \\
&= \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}\left[\sum_{i=1}^{N}\left\{t_{iu}t_{iy} - t_{iu}^2 + \frac{t_y - t_u}{N - t_x}t_{iu}t_{ix}\right.\right. \\
&\qquad\qquad \left.\left. - \frac{t_u}{t_x}\left(t_{ix}t_{iy} - t_{ix}t_{iu} + \frac{t_y - t_u}{N - t_x}t_{ix}^2\right)\right\}\right].
\end{aligned}$$

(b) Now,

$$t_{iy} = \sum_{j=1}^{M} y_{ij}$$

and

$$t_{iu} = \sum_{j=1}^{M} x_{ij}y_{ij}.$$

So if psu $i$ is in domain 1 then $x_{ij} = 1$ for every ssu in psu $i$, $t_{ix} = M$, and $t_{iu} = t_{iy}$; if psu $i$ is in domain 2 then $x_{ij} = 0$ for every ssu in psu $i$, $t_{ix} = 0$, and $t_{iu} = 0$. We

may then write the covariance as

$$\begin{aligned}
&\sum_{i=1}^{N}\left\{t_{iu}t_{iy}-t_{iu}^2+\frac{t_y-t_u}{N-t_x}t_{iu}t_{ix}-\frac{t_u}{t_x}\left(t_{ix}t_{iy}-t_{ix}t_{iu}+\frac{t_y-t_u}{N-t_x}t_{ix}^2\right)\right\}\\
&= \sum_{i\in\text{domain 1}}\left\{t_{iu}t_{iy}-t_{iu}^2+\frac{t_y-t_u}{N-t_x}t_{iu}t_{ix}-\frac{t_u}{t_x}\left(t_{ix}t_{iy}-t_{ix}t_{iu}+\frac{t_y-t_u}{N-t_x}t_{ix}^2\right)\right\}\\
&\quad+\sum_{i\in\text{domain 2}}\left\{t_{iu}t_{iy}-t_{iu}^2+\frac{t_y-t_u}{N-t_x}t_{iu}t_{ix}-\frac{t_u}{t_x}\left(t_{ix}t_{iy}-t_{ix}t_{iu}+\frac{t_y-t_u}{N-t_x}t_{ix}^2\right)\right\}\\
&= \sum_{i\in\text{domain 1}}\left\{t_{iy}t_{iy}-t_{iy}^2+\frac{t_y-t_u}{N-t_x}t_{iy}M-\frac{t_u}{t_x}\left(Mt_{iy}-Mt_{iy}+\frac{t_y-t_u}{N-t_x}M^2\right)\right\}\\
&= \sum_{i\in\text{domain 1}}\left\{\frac{t_y-t_u}{N-t_x}t_{iy}M-\frac{t_u}{t_x}\frac{t_y-t_u}{N-t_x}M^2\right\}\\
&= \frac{t_y-t_u}{N-t_x}t_uM-N\frac{t_x}{NM}\frac{t_u}{t_x}\frac{t_y-t_u}{N-t_x}M^2\\
&= 0.
\end{aligned}$$

(c) Almost any example will work, as long as some psus have units from both domains.

**6.22** (a) Since $E(Z_i)=\pi_i$,

$$E(\hat{t}_y)=\sum_{i=1}^{N}\frac{1}{\pi_i}u_iE(Z_i)=\sum_{i=1}^{N}u_i=\sum_{i=1}^{N}\sum_{k=1}^{M}\frac{\ell_{ik}y_k}{L_k}=\sum_{k=1}^{M}y_k.$$

The last equality follows since $\sum_{i=1}^{N}\ell_{ik}=L_k$.

The variance given is the variance of the one-stage Horvitz-Thompson estimator.

(b) Note that

$$\hat{t}_y=\sum_{i=1}^{N}\frac{Z_i}{\pi_i}\sum_{k=1}^{M}\frac{\ell_{ik}y_k}{L_k}=\sum_{k=1}^{M}\frac{1}{L_k}\sum_{i=1}^{N}\frac{Z_i}{\pi_i}\ell_{ik}y_k.$$

But the sum is over all $k$ from 1 to $M$, not just the units in $\mathcal{S}^B$. We need to show that $\sum_{i=1}^{N}\frac{Z_i}{\pi_i}\ell_{ik}=0$ for $k\notin\mathcal{S}^B$.

$$w_k^*=\frac{\displaystyle\sum_{i=1}^{N}\frac{Z_i}{\pi_i}\ell_{ik}}{\displaystyle\sum_{i=1}^{N}\ell_{ik}}.$$

But a student is in $\mathcal{S}^B$ if and only if s/he is linked to one of the sampled units in $\mathcal{S}^A$. In other words, $k\in\mathcal{S}^B$ if and only if $\sum_{i\in\mathcal{S}^A}\ell_{ik}>0$. For $k\notin\mathcal{S}^B$, we must have $\ell_{ik}=0$ for each $i\in\mathcal{S}^A$.

(c) Suppose $L_k = 1$ for all $k$. Then,

$$\hat{t}_y = \sum_{i=1}^{N} \frac{Z_i}{\pi_i} \sum_{k=1}^{M} \ell_{ik} y_k = \sum_{i=1}^{N} \frac{Z_i}{\pi_i} y_i$$

because $\sum_{k=1}^{M} \ell_{ik} y_k = y_i$.

(d) The values of $u_i$ are:

| | | Element $k$ from $U^B$ | | |
|---|---|---|---|---|
| $\ell_{ik}$ | | 1 | 2 | $u_i$ |
| Unit $i$ | 1 | 1 | 0 | $4/2 = 2$ |
| from $U^A$ | 2 | 1 | 1 | $4/2 + 6/2 = 5$ |
| | 3 | 0 | 1 | $6/2 = 3$ |

Here are the three SRSs from $U^A$:

| Sample | $\hat{t}_y$ |
|---|---|
| {1,2} | $\frac{3}{2}(2+5) = \frac{21}{2}$ |
| {1,3} | $\frac{3}{2}(2+3) = \frac{15}{2}$ |
| {2,3} | $\frac{3}{2}(5+3) = \frac{24}{2}$ |

Consequently,

$$E[\hat{t}_y] = \frac{1}{3}\left(\frac{21}{2} + \frac{15}{2} + \frac{24}{2}\right) = 10,$$

so it is unbiased. But

$$\begin{aligned} V[\hat{t}_y] &= \frac{1}{3}\left[\left(\frac{21}{2} - 10\right)^2 + \left(\frac{15}{2} - 10\right)^2 + \left(\frac{24}{2} - 10\right)^2\right] \\ &= \frac{1}{3}\left[0.25 + 6.25 + 4\right] \\ &= 3.5. \end{aligned}$$

(e) We construct the variable $u_i = \sum_{k=1}^{M} \ell_{ik} y_k / L_k$ for each adult in the sample, where $L_k =$ (number of other adults +1). Using weight $w_i = 40{,}000/100 = 400$, we calculate $\hat{t}_u = \sum_{i \in \mathcal{S}} w_i u_i = 7200$ with

$$\hat{V}(\hat{t}_u) = \frac{N^2}{n} s_u^2 = 1{,}900{,}606.$$

This gives a 95% CI of [4464.5, 9935.5]. Note that the without-replacement variance estimator could also be used.

The following SAS code will compute the estimates.

```
data wtshare;
   infile wtshare delimiter="," firstobs=2;
   input id child preschool numadult;
   yoverLk = preschool/(numadult+1);

proc sort data=wtshare;
   by id;
run;

proc means data=wtshare sum noprint;
   by id;
   var yoverLk;
   output out=sumout sum = u;

data sumout;
   set sumout;
   sampwt = 40000/100;

proc surveymeans data=sumout mean sum clm clsum;
   weight sampwt;
   var u;
run;
```

**6.23** (a) We have $\pi_i$'s 0.50 0.25 0.50 0.75, $V(\hat{t}_{\text{HT}}) = 180.1147$, and $V(\hat{t}_\psi) = 101.4167$.

(b)

$$
\begin{aligned}
&\frac{1}{2n}\sum_{i=1}^{N}\sum_{k=1}^{N}\pi_i\pi_k\left(\frac{t_i}{\pi_i}-\frac{t_k}{\pi_k}\right)^2 \\
&\quad= \frac{1}{2n}\sum_{i=1}^{N}\sum_{k=1}^{N}\pi_i\pi_k\left[\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)-\left(\frac{t_k}{\pi_k}-\frac{t}{n}\right)\right]^2 \\
&\quad= \frac{1}{2n}\sum_{i=1}^{N}\sum_{k=1}^{N}\pi_i\pi_k\left[\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2+\left(\frac{t_k}{\pi_k}-\frac{t}{n}\right)^2-2\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)\left(\frac{t_k}{\pi_k}-\frac{t}{n}\right)\right] \\
&\quad= \frac{1}{n}\sum_{i=1}^{N}\sum_{k=1}^{N}\pi_i\pi_k\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2-\frac{1}{n}\left[\sum_{i=1}^{N}\pi_i\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)\right]^2 \\
&\quad= \frac{1}{n}\sum_{i=1}^{N}\sum_{k=1}^{N}n^2\psi_i\psi_k\left(\frac{t_i}{n\psi_i}-\frac{t}{n}\right)^2-\frac{1}{n}\left[\sum_{i=1}^{N}(t_i-\psi_i t)\right]^2 \\
&\quad= \frac{1}{n}\sum_{i=1}^{N}\psi_i\left(\frac{t_i}{\psi_i}-t\right)^2 \\
&\quad= V(\hat{t}_\psi).
\end{aligned}
$$

(c)

$$\begin{aligned}
V(\hat{t}_\psi) - V(\hat{t}_{HT}) &= \frac{1}{2n}\sum_{i=1}^{N}\sum_{k=1}^{N}\pi_i\pi_k\left(\frac{t_i}{\pi_i}-\frac{t_k}{\pi_k}\right)^2 \\
&\quad -\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}(\pi_i\pi_k-\pi_{ik})\left(\frac{t_i}{\pi_i}-\frac{t_k}{\pi_k}\right)^2 \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\left(\frac{\pi_i\pi_k}{n}-\pi_i\pi_k+\pi_{ik}\right)\left(\frac{t_i}{\pi_i}-\frac{t_k}{\pi_k}\right)^2 \\
&> \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\left[\pi_i\pi_k\left(\frac{1}{n}-1\right)-\frac{n-1}{n}\pi_i\pi_k\right]\left(\frac{t_i}{\pi_i}-\frac{t_k}{\pi_k}\right)^2 \\
&= 0.
\end{aligned}$$

(d) If $\pi_{ik} \geq (n-1)\pi_i\pi_k/n$ for all $i$ and $k$, then

$$\min\left(\frac{\pi_{ik}}{\pi_k}\right) \geq \frac{n-1}{n}\pi_i \text{ for all } i.$$

Consequently,

$$\sum_{i=1}^{N}\min\left(\frac{\pi_{ik}}{\pi_k}\right) \geq \frac{n-1}{n}\sum_{i=1}^{N}\pi_i = n-1.$$

(e) Suppose $V(\hat{t}_{HT}) \leq V(\hat{t}_\psi)$. Then from part (b),

$$\begin{aligned}
0 &\leq V(\hat{t}_\psi) - V(\hat{t}_{HT}) \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{k\neq i}^{N}\left(\pi_{ik}-\frac{n-1}{n}\pi_i\pi_k\right)\left(\frac{t_i}{\pi_i}-\frac{t_k}{\pi_k}\right)^2 \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{k\neq i}^{N}\left(\pi_{ik}-\frac{n-1}{n}\pi_i\pi_k\right)\left[\left(\frac{t_i}{\pi_i}\right)^2+\left(\frac{t_k}{\pi_k}\right)^2-2\frac{t_i}{\pi_i}\frac{t_k}{\pi_k}\right] \\
&= \sum_{i=1}^{N}\sum_{k\neq i}^{N}\left(\pi_{ik}-\frac{n-1}{n}\pi_i\pi_k\right)\left[\left(\frac{t_i}{\pi_i}\right)^2-\frac{t_i}{\pi_i}\frac{t_k}{\pi_k}\right] \\
&= \sum_{i=1}^{N}\sum_{k\neq i}^{N}\pi_{ik}\left(\frac{t_i}{\pi_i}\right)^2-\sum_{i=1}^{N}\sum_{k\neq i}^{N}\pi_{ik}\frac{t_i}{\pi_i}\frac{t_k}{\pi_k}-\frac{n-1}{n}\sum_{i=1}^{N}\pi_i(n-\pi_i)\left(\frac{t_i}{\pi_i}\right)^2 \\
&\quad +\frac{n-1}{n}\sum_{i=1}^{N}\sum_{k\neq i}^{N}t_it_k \\
&= \sum_{i=1}^{N}(n-1)\pi_i\left(\frac{t_i}{\pi_i}\right)^2-\sum_{i=1}^{N}\sum_{k\neq i}^{N}\pi_{ik}\frac{t_i}{\pi_i}\frac{t_k}{\pi_k}-(n-1)\sum_{i=1}^{N}\pi_i\left(\frac{t_i}{\pi_i}\right)^2 \\
&\quad +\frac{n-1}{n}\sum_{i=1}^{N}t_i^2+\frac{n-1}{n}\sum_{i=1}^{N}\sum_{k\neq i}^{N}t_it_k
\end{aligned}$$

Consequently,

$$\frac{n-1}{n}\sum_{i=1}^{N}\sum_{k=1}^{N}t_i t_k \geq \sum_{i=1}^{N}\sum_{k\neq i}^{N}\pi_{ik}\frac{t_i}{\pi_i}\frac{t_k}{\pi_k},$$

or

$$\sum_{i=1}^{N}\sum_{k=1}^{N}a_{ik}t_i t_k \geq 0,$$

where $a_{ii} = 1$ and

$$a_{ik} = 1 - \frac{n}{n-1}\frac{\pi_{ik}}{\pi_i\pi_k} \quad \text{if } i \neq k.$$

The matrix $\mathbf{A}$ must be nonnegative definite, which means that all principal submatrices must have determinant $\geq 0$. Using a $2 \times 2$ submatrix, we have $1 - a_{ik}^2 \geq 0$, which gives the result.

**6.24** (a) We have

| | | $k$ | | | | |
|---|---|---|---|---|---|---|
| | $\pi_{ik}$ | 1 | 2 | 3 | 4 | $\pi_i$ |
| | 1 | 0.00 | 0.31 | 0.20 | 0.14 | 0.65 |
| $i$ | 2 | 0.31 | 0.00 | 0.03 | 0.01 | 0.35 |
| | 3 | 0.20 | 0.03 | 0.00 | 0.31 | 0.54 |
| | 4 | 0.14 | 0.01 | 0.31 | 0.00 | 0.46 |
| | $\pi_k$ | 0.65 | 0.35 | 0.54 | 0.46 | 2.00 |

(b)

| Sample, $\mathcal{S}$ | $\hat{t}_{\text{HT}}$ | $\hat{V}_{\text{HT}}(\hat{t}_{\text{HT}})$ | $\hat{V}_{\text{SYG}}(\hat{t}_{\text{HT}})$ |
|---|---|---|---|
| $\{1,2\}$ | 9.56 | 38.10 | -0.9287681 |
| $\{1,3\}$ | 5.88 | -4.74 | 2.4710422 |
| $\{1,4\}$ | 4.93 | -3.68 | 8.6463858 |
| $\{2,3\}$ | 7.75 | -100.25 | 71.6674365 |
| $\{2,4\}$ | 21.41 | -165.72 | 323.3238494 |
| $\{3,4\}$ | 3.12 | 3.42 | -0.1793659 |

Note that $\sum_i \sum_{k>i} \pi_{ik}\hat{V}(\hat{t}_{\text{HT}}) = 6.74$ for each method.

**6.25** (a)

$$\begin{aligned}
&P\{\text{psu's } i \text{ and } j \text{ are in the sample}\} \\
&= P\{\text{psu } i \text{ drawn first and psu } j \text{ drawn second}\} \\
&\quad + P\{\text{psu } j \text{ drawn first and psu } i \text{ drawn second}\} \\
&= \frac{a_i}{\sum_{k=1}^{N} a_k}\frac{\psi_j}{1-\psi_i} + \frac{a_j}{\sum_{k=1}^{N} a_k}\frac{\psi_i}{1-\psi_j} \\
&= \frac{\psi_i(1-\psi_i)\psi_j}{\sum_{k=1}^{N} a_k(1-\psi_i)(1-\pi_i)} + \frac{\psi_j(1-\psi_j)\psi_i}{\sum_{k=1}^{N} a_k(1-\psi_j)(1-\pi_j)} \\
&= \frac{\psi_i\psi_j}{\sum_{k=1}^{N} a_k}\left(\frac{1}{1-\pi_i} + \frac{1}{1-\pi_j}\right)
\end{aligned}$$

(b) Using (a),

$$\begin{aligned}
P\{\text{psu } i \text{ in sample}\} &= \sum_{j=1, j\neq i}^{N} \pi_{ij} \\
&= \sum_{j=1, j\neq i}^{N} \frac{\psi_i\psi_j}{\sum_{k=1}^{N} a_k}\left(\frac{1}{1-\pi_i} + \frac{1}{1-\pi_j}\right) \\
&= \frac{\psi_i}{\sum_{k=1}^{N} a_k}\left\{\sum_{j=1}^{N}\psi_j\left(\frac{1}{1-\pi_i} + \frac{1}{1-\pi_j}\right) - \frac{2\psi_i}{1-\pi_i}\right\} \\
&= \frac{\psi_i}{\sum_{k=1}^{N} a_k}\left\{\frac{1}{1-\pi_i} + \sum_{j=1}^{N}\frac{\psi_j}{1-\pi_j} - \frac{\pi_i}{1-\pi_i}\right\} \\
&= \frac{\psi_i}{\sum_{k=1}^{N} a_k}\left\{1 + \sum_{j=1}^{N}\frac{\psi_j}{1-\pi_j}\right\}
\end{aligned}$$

In the third step above, we used the constraint that $\sum_{j=1}^{N}\pi_j = n = 2$, so $\sum_{j=1}^{N}\psi_j = 1$. Now note that

$$\begin{aligned}
2\textstyle\sum_{k=1}^{N} a_k &= 2\sum_{k=1}^{N}\frac{\psi_k(1-\psi_k)}{1-2\psi_k} \\
&= \sum_{k=1}^{N}\frac{\psi_k(1-2\psi_k+1)}{1-2\psi_k} \\
&= 1 + \sum_{k=1}^{N}\frac{\psi_k}{1-\pi_k}.
\end{aligned}$$

Thus $P\{\text{psu } i \text{ in sample}\} = 2\psi_i = \pi_i$.

(c) Using part (a),

$$\begin{aligned}\pi_i\pi_j - \pi_{ij} &= 4\psi_i\psi_j - \frac{\psi_i\psi_j}{\sum_{k=1}^{N} a_k}\left(\frac{1}{1-\pi_i} + \frac{1}{1-\pi_j}\right) \\ &= \frac{\psi_i\psi_j\left[4\left(\sum_{k=1}^{N} a_k\right)(1-\pi_i)(1-\pi_j) - (1-\pi_j+1-\pi_i)\right]}{\left(\sum_{k=1}^{N} a_k\right)(1-\pi_i)(1-\pi_j)}.\end{aligned}$$

Using the hint in part (b),

$$\begin{aligned}&4\left(\sum_{k=1}^{N} a_k\right)(1-\pi_i)(1-\pi_j) - (1-\pi_j+1-\pi_i) \\ &= 2\left[1+\sum_{k=1}^{N}\frac{\psi_k}{1-\pi_k}\right](1-\pi_i)(1-\pi_j) - 2 + \pi_i + \pi_j \\ &= 2(1-\pi_i)(1-\pi_j)\sum_{k=1}^{N}\frac{\psi_k}{1-\pi_k} + 2 - 2\pi_i - 2\pi_j + 2\pi_i\pi_j - 2 + \pi_i + \pi_j \\ &= 2(1-\pi_i)(1-\pi_j)\sum_{k=1}^{N}\frac{\psi_k}{1-\pi_k} - \pi_i - \pi_j + 2\pi_i\pi_j \\ &= \geq 2(1-\pi_i)\psi_j + 2(1-\pi_j)\psi_i - \pi_i - \pi_j + 2\pi_i\pi_j \\ &= (1-\pi_i)\pi_j + (1-\pi_j)\pi_i - \pi_i - \pi_j + 2\pi_i\pi_j \\ &= 0.\end{aligned}$$

Thus $\pi_i\pi_j - \pi_{ij} \geq 0$, and the SYG estimator of the variance is guaranteed to be nonnegative.

**6.26** The desired probabilities of inclusion are $\pi_i = 2M_i/\sum_{j=1}^{5} M_j$. We calculate $\psi_i = \pi_i/2$ and $a_i = \psi_i(1-\psi_i)/(1-\pi_i)$ for each psu in the following table:

| psu, $i$ | $M_i$ | $\pi_i$ | $\psi_i$ | $a_i$ |
|---|---|---|---|---|
| 1 | 5 | 0.40 | 0.20 | 0.26667 |
| 2 | 4 | 0.32 | 0.16 | 0.19765 |
| 3 | 8 | 0.64 | 0.32 | 0.60444 |
| 4 | 5 | 0.40 | 0.20 | 0.26667 |
| 5 | 3 | 0.24 | 0.12 | 0.13895 |
| Total | 25 | 2.00 | 1.00 | 1.47437 |

According to Brewer's method,

$$P(\text{select psu } i \text{ on 1st draw}) = a_i \Big/ \sum_{j=1}^{5} a_j$$

and

$$P(\text{psu } j \text{ on 2nd draw} \mid \text{psu } i \text{ on 1st draw}) = \psi_j \Big/ (1 - \psi_i).$$

Then

$$P\{\mathcal{S} = (1,2)\} = \frac{.26667}{1.47437}\frac{0.16}{0.8} = 0.036174,$$

$$P\{\mathcal{S} = (2,1)\} = \frac{.19765}{1.47437}\frac{0.2}{0.84} = 0.031918,$$

and

$$\pi_{12} = P\{\mathcal{S} = (1,2)\} + P\{\mathcal{S} = (2,1)\} = 0.068.$$

Continuing in like manner, we have the following table of $\pi_{ij}$.

| $i \backslash j$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | — | .068 | .193 | .090 | .049 |
| 2 | .068 | — | .148 | .068 | .036 |
| 3 | .193 | .148 | — | .193 | .107 |
| 4 | .090 | .068 | .193 | — | .049 |
| 5 | .049 | .036 | .107 | .049 | — |
| Sum | .400 | .320 | .640 | .400 | .240 |

We use (6.21) to calculate the variance of the Horvitz-Thompson estimator.

| $i$ | $j$ | $\pi_{ij}$ | $\pi_i$ | $\pi_j$ | $t_i$ | $t_j$ | $(\pi_i\pi_j - \pi_{ij})\left(\frac{t_i}{\pi_i} - \frac{t_j}{\pi_j}\right)^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 0.068 | 0.40 | 0.32 | 20 | 25 | 47.39 |
| 1 | 3 | 0.193 | 0.40 | 0.64 | 20 | 38 | 5.54 |
| 1 | 4 | 0.090 | 0.40 | 0.40 | 20 | 24 | 6.96 |
| 1 | 5 | 0.049 | 0.40 | 0.24 | 20 | 21 | 66.73 |
| 2 | 3 | 0.148 | 0.32 | 0.64 | 25 | 38 | 20.13 |
| 2 | 4 | 0.068 | 0.32 | 0.40 | 25 | 24 | 19.68 |
| 2 | 5 | 0.036 | 0.32 | 0.24 | 25 | 21 | 3.56 |
| 3 | 4 | 0.193 | 0.64 | 0.40 | 38 | 24 | 0.02 |
| 3 | 5 | 0.107 | 0.64 | 0.24 | 38 | 21 | 37.16 |
| 4 | 5 | 0.049 | 0.40 | 0.24 | 24 | 21 | 35.88 |
| Sum | | 1 | | | | | 243.07 |

Note that for this population, $t = 128$. To check the results, we see that

$$\Sigma P(\mathcal{S})\hat{t}_{HT\mathcal{S}} = 128 \quad \text{and} \quad \Sigma P(\mathcal{S})(\hat{t}_{HT\mathcal{S}} - 128)^2 = 243.07.$$

**6.27** A sequence of simple random samples with replacement (SRSWR) is drawn until the first SRSWR in which the two psu's are distinct. As each SRSWR in the sequence is selected independently, for the $l$th SRSWR in the sequence, and for

$i \neq j$,

$$\begin{aligned} & P\{\text{psu's } i \text{ and } j \text{ chosen in } l\text{th SRSWR}\} \\ &= P\{\text{psu } i \text{ chosen first and psu } j \text{ chosen second}\} \\ &\quad + P\{\text{psu } j \text{ chosen first and psu } i \text{ chosen second}\} \\ &= 2\psi_i\psi_j. \end{aligned}$$

The $(l+1)$st SRSWR is chosen to be the sample if each of the previous $l$ SRSWR's is rejected because the two psu's are the same. Now

$$P(\text{the two psu's are the same in an SRSWR}) = \sum_{k=1}^{N} \psi_k^2,$$

so because SRSWR's are drawn independently,

$$P(\text{reject first } l \text{ SRSWR's}) = \left( \sum_{k=1}^{N} \psi_k^2 \right)^l.$$

Thus

$$\begin{aligned} \pi_{ij} &= P\{\text{psu's } i \text{ and } j \text{ are in the sample}\} \\ &= \sum_{l=0}^{\infty} P\{\text{psu's } i \text{ and } j \text{ chosen in } (l+1)\text{st SRSWR,} \\ &\qquad\qquad \text{and the first } l \text{ SRSWR's are rejected}\} \\ &= \sum_{l=0}^{\infty} (2\psi_i\psi_j) \left( \sum_{k=1}^{N} \psi_k^2 \right)^l \\ &= \frac{2\psi_i\psi_j}{1 - \sum_{k=1}^{N} \psi_k^2}. \end{aligned}$$

Equation (6.18) implies

$$\begin{aligned} \pi_i &= \textstyle\sum_{j=1, j\neq i}^{N} \pi_{ij} \\ &= \textstyle\sum_{j=1, j\neq i}^{N} 2\psi_i\psi_j \Big/ \left(1 - \sum_{k=1}^{N} \psi_k^2\right) \\ &= 2\psi_i \Big/ \left(1 - \textstyle\sum_{k=1}^{N} \psi_k^2\right) - 2\psi_i^2 \Big/ \left(1 - \textstyle\sum_{k=1}^{N} \psi_k^2\right) \\ &= 2\psi_i(1 - \psi_i) \Big/ \left(1 - \textstyle\sum_{k=1}^{N} \psi_k^2\right). \end{aligned}$$

Note that, as (6.17) predicts,

$$\sum_{i=1}^{N} \pi_i = 2\left(1 - \sum_{i=1}^{N} \psi_i^2\right) \Big/ \left(1 - \sum_{k=1}^{N} \psi_k^2\right) = 2$$

**6.28** Note that, using the indicator variables,

$$\sum_{k=1}^{n} I_{ki} = 1 \quad \text{for all } i,$$

$$x_{ki} = \frac{M_i}{\sum_{j=1}^{N} I_{kj} M_j},$$

$$P(Z_i = 1 \mid I_{11}, \ldots, I_{nN}) = \sum_{k=1}^{n} \frac{I_{ki} M_i}{\sum_{j=1}^{N} I_{kj} M_j} = \sum_{k=1}^{n} I_{ki} x_{ki}$$

and

$$\hat{t}_{\text{RHC}} = \sum_{k=1}^{n} \frac{t_{\alpha(k)}}{x_{k,\alpha(k)}} = \sum_{k=1}^{n} \sum_{i=1}^{N} I_{ki} Z_i \frac{t_i}{x_{ki}}.$$

We show that $\hat{t}_{\text{RHC}}$ is conditionally unbiased for $t$ given the grouping

$$\begin{aligned} E[\hat{t}_{\text{RHC}} \mid I_{11}, \ldots, I_{nN}] &= E\left[ \sum_{k=1}^{n} \sum_{i=1}^{N} I_{ki} Z_i \frac{t_i}{x_{ki}} \mid I_{11}, \ldots, I_{nN} \right] \\ &= \sum_{k=1}^{n} \sum_{i=1}^{N} I_{ki} \frac{t_i}{x_{ki}} \sum_{l=1}^{N} I_{kl} x_{kl} \\ &= \sum_{k=1}^{n} \sum_{i=1}^{N} I_{ki} t_i \\ &= \sum_{i=1}^{N} t_i = t. \end{aligned}$$

Since $E[\hat{t}_{\text{RHC}} \mid I_{11}, \ldots, I_{nN}] = t$ for any random grouping of psu's, we have that $E[\hat{t}_{\text{RHC}}] = t$.

To find the variance, note that

$$V[\hat{t}_{\text{RHC}}] = E[V(\hat{t}_{\text{RHC}} \mid I_{11}, \ldots, I_{nN})] + V[E(\hat{t}_{\text{RHC}} \mid I_{11}, \ldots, I_{nN})].$$

Since $E[\hat{t}_{\text{RHC}} \mid I_{11}, \ldots, I_{nN}] = t$, however, we know that $V[E(\hat{t}_{\text{RHC}} \mid I_{11}, \ldots, I_{nN})] = 0$. Conditionally on the grouping, the $k$th term in $\hat{t}_{\text{RHC}}$ estimates the total of group $k$ using an unequal-probability sample of size one. We can thus use (6.4) within each group to find the conditional variance, noting that psu's in different groups are selected independently. (We can obtain the same result by using the indicator

variables directly, but it's messier.) Then

$$\begin{aligned} V(\hat{t}_{\text{RHC}} \mid I_{11}, \ldots, I_{nN}) &= \sum_{k=1}^{n}\sum_{i=1}^{N} I_{ki} x_{ki}\left(\frac{t_i}{x_{ki}} - \sum_{j=1}^{N} I_{kj} t_j\right)^2 \\ &= \sum_{k=1}^{n}\sum_{i=1}^{N} I_{ki}\frac{t_i^2}{x_{ki}} - \sum_{k=1}^{n}\sum_{i=1}^{N}\sum_{j=1}^{N} I_{ki} t_i I_{kj} t_j \\ &= \sum_{k=1}^{N}\sum_{i=1}^{N}\sum_{j=1}^{N} I_{ki} I_{kj}\left(\frac{M_j t_i^2}{M_i} - t_i t_j\right). \end{aligned}$$

item Now to find $E[V(\hat{t}_{\text{RHC}} \mid I_{11}, \ldots, I_{nN})]$, we need $E[I_{ki}]$ and $E[I_{ki}I_{kj}]$ for $i \neq j$. Let $N_k$ be the number of psu's in group $k$. Then

$$E[I_{ki}] = P\{\text{psu } i \text{ in group } k\} = \frac{N_k}{N}$$

and, for $i \neq j$,

$$E[I_{ki}I_{kj}] = P\{\text{psu's } i \text{ and } j \text{ in group } k\} = \frac{N_k}{N}\frac{N_k - 1}{N - 1}.$$

Thus, letting $\psi_i = M_i / \sum_{j=1}^{N} M_j$,

$$\begin{aligned} V[\hat{t}_{\text{RHC}}] &= E[V(\hat{t}_{\text{RHC}} \mid I_{11}, \ldots, I_{nN})] \\ &= E\left[\sum_{k=1}^{n}\sum_{i=1}^{N}\sum_{j=1}^{N} I_{ki} I_{kj}\left(\frac{M_j t_i^2}{M_i} - t_i t_j\right)\right] \\ &= \sum_{k=1}^{n}\sum_{i=1}^{N}\frac{N_k}{N}(t_i^2 - t_i^2) + \sum_{k=1}^{n}\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N}\frac{N_k}{N}\frac{N_k - 1}{N - 1}\left(\frac{M_j t_i^2}{M_i} - t_i t_j\right) \\ &= \left(\sum_{k=1}^{n}\frac{N_k}{N}\frac{N_k - 1}{N - 1}\right)\left(\sum_{i=1}^{N}\frac{t_i^2}{\psi_i} - t^2\right) \\ &= \left(\sum_{k=1}^{n}\frac{N_k}{N}\frac{N_k - 1}{N - 1}\right)\left(\sum_{i=1}^{N}\psi_i\left[\frac{t_i}{\psi_i} - t\right]^2\right). \end{aligned}$$

The second factor equals $nV(\hat{t}_\psi)$, with $V(\hat{t}_\psi)$ given in (6.46), assuming one stage cluster sampling.

What should $N_1, \ldots, N_n$ be in order to minimize $V[\hat{t}_{\text{RHC}}]$? Note that

$$\sum_{k=1}^{n} N_k(N_k - 1) = \sum_{k=1}^{n} N_k^2 - N$$

is smallest when all $N_k$'s are equal. If $N/n = L$ is an integer, take $N_k = L$ for $k = 1, 2, \ldots, n$. With this design,

$$\begin{aligned} V[\hat{t}_{\mathrm{RHC}}] &= \frac{L-1}{N-1}\sum_{i=1}^{N}\psi_i\left(\frac{t_i}{\psi_i} - t\right)^2 \\ &= \frac{N-n}{N-1}V(\hat{t}_\psi). \end{aligned}$$

**6.29** (a)

$$\begin{aligned} \sum_{k\neq i}\tilde{\pi}_{ik} &= \sum_{k\neq i}\pi_i\pi_k\left[1-(1-\pi_i)(1-\pi_k)/\sum_{j=1}^{N}c_j\right] \\ &= \pi_i\sum_{k\neq i}\pi_k - \frac{\pi_i(1-\pi_i)}{\sum_{j=1}^{N}c_j}\sum_{k\neq i}\pi_k(1-\pi_k) \\ &= \pi_i(n-\pi_i) - \frac{\pi_i(1-\pi_i)}{\sum_{j=1}^{N}c_j}\left[\sum_{k=1}^{N}\pi_k(1-\pi_k) - \pi_i(1-\pi_i)\right] \\ &= \pi_i(n-\pi_i) - \pi_i(1-\pi_i) + \frac{\pi_i^2(1-\pi_i)^2}{\sum_{j=1}^{N}c_j} \\ &= \pi_i(n-1) + \frac{\pi_i^2(1-\pi_i)^2}{\sum_{j=1}^{N}c_j}. \end{aligned}$$

(b) If an SRS is taken, $\pi_i = n/N$, so

$$\sum_{j=1}^{N}c_j = \sum_{j=1}^{N}\frac{n}{N}\left(1-\frac{n}{N}\right) = n\left(1-\frac{n}{N}\right)$$

and

$$\begin{aligned} \tilde{\pi}_{ik} &= \frac{n^2}{N^2}\left[1-\frac{(1-n/N)(1-n/N)}{n\left(1-\frac{n}{N}\right)}\right] \\ &= \frac{n}{N^2}\left[n-1+\frac{n}{N}\right] \\ &= \frac{n}{N}\left[\frac{n-1}{N}+\frac{n}{N^2}\right] \\ &= \frac{n}{N}\frac{n-1}{N-1}\frac{N-1}{n-1}\left[\frac{n-1}{N}+\frac{n}{N^2}\right] \\ &= \frac{n}{N}\frac{n-1}{N-1}\left[\frac{N-1}{N}+\frac{n(N-1)}{(n-1)N^2}\right] \\ &= \frac{n}{N}\frac{n-1}{N-1}\left[1+\frac{N-n}{(n-1)N^2}\right] \end{aligned}$$

(c) First note that

$$\pi_i\pi_k - \tilde{\pi}_{ik} = \pi_i\pi_k \frac{(1-\pi_i)(1-\pi_k)}{\sum_{j=1}^N c_j}.$$

Then, letting $B = \sum_{j=1}^N c_j$,

$$\begin{aligned}
V_{Haj}(\hat{t}_{HT}) &= \frac{1}{2}\sum_{i=1}^N\sum_{k=1}^N(\pi_i\pi_k - \tilde{\pi}_{ik})\left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k}\right)^2 \\
&= \frac{1}{2}\sum_{i=1}^N\sum_{k=1}^N \pi_i\pi_k\frac{(1-\pi_i)(1-\pi_k)}{\sum_{j=1}^N c_j}\left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k}\right)^2 \\
&= \frac{1}{2\sum_{j=1}^N c_j}\sum_{i=1}^N\sum_{k=1}^N \pi_i\pi_k(1-\pi_i)(1-\pi_k)\left(2\frac{t_i^2}{\pi_i^2} - 2\frac{t_i}{\pi_i}\frac{t_k}{\pi_k}\right) \\
&= \frac{1}{\sum_{j=1}^N c_j}\sum_{i=1}^N\sum_{k=1}^N \pi_i\pi_k(1-\pi_i)(1-\pi_k)\left(\frac{t_i^2}{\pi_i^2} - \frac{t_i}{\pi_i}\frac{t_k}{\pi_k}\right) \\
&= \sum_{i=1}^N \pi_i(1-\pi_i)\frac{t_i^2}{\pi_i^2} - \frac{1}{\sum_{j=1}^N c_j}\left(\sum_{i=1}^N \pi_i(1-\pi_i)\frac{t_i}{\pi_i}\right)^2 \\
&= \sum_{i=1}^N c_i\frac{t_i^2}{\pi_i^2} - \frac{1}{\sum_{j=1}^N c_j}\left(\sum_{i=1}^N c_i\frac{t_i}{\pi_i}\right)^2 \\
&= \sum_{i=1}^N c_i\left(\frac{t_i^2}{\pi_i^2} - A\right)^2.
\end{aligned}$$

**6.30**

(a) From (6.21),

$$\begin{aligned}
V(\hat{t}_{\mathrm{HT}}) &= \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N}(\pi_i\pi_k-\pi_{ik})\left(\frac{t_i}{\pi_i}-\frac{t_k}{\pi_k}\right)^2 \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N}(\pi_i\pi_k-\pi_{ik})\left(\frac{t_i}{\pi_i}-\frac{t}{n}+\frac{t}{n}-\frac{t_k}{\pi_k}\right)^2 \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N}(\pi_i\pi_k-\pi_{ik})\times \\
&\qquad \left\{\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2+\left(\frac{t_k}{\pi_k}-\frac{t}{n}\right)^2-2\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)\left(\frac{t_k}{\pi_k}-\frac{t}{n}\right)\right\} \\
&= \sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N}(\pi_i\pi_k-\pi_{ik})\left\{\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2-\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)\left(\frac{t_k}{\pi_k}-\frac{t}{n}\right)\right\}
\end{aligned}$$

From Theorem 6.1, we know that

$$\sum_{k=1}^{N}\pi_k = n$$

and

$$\sum_{\substack{k=1\\k\neq i}}^{N}\pi_{ik} = (n-1)\pi_i,$$

so

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N}(\pi_i\pi_k-\pi_{ik})\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2 &= \sum_{i=1}^{N}[\pi_i(n-\pi_i)-(n-1)\pi_i]\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2 \\
&= \sum_{i=1}^{N}\pi_i(1-\pi_i)\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2.
\end{aligned}$$

This gives the first two terms in (6.47); the third term is the cross-product term above.

(b)

For an SRS, $\pi_i = n/N$ and $\pi_{ik} = [n(n-1)]/[N(N-1)]$. The first term is

$$\sum_{i=1}^{N}\frac{n}{N}\left(\frac{Nt_i}{n}-\frac{t}{n}\right)^2 = \sum_{i=1}^{N}\frac{N}{n}(t_i-\bar{t}_U)^2 = N(N-1)\frac{S_t^2}{n}.$$

The second term is

$$\sum_{i=1}^{N}\left(\frac{n}{N}\right)^2\left(\frac{Nt_i}{n}-\frac{t}{n}\right)^2 = n(N-1)\frac{S_t^2}{n}.$$

(c) Substituting $\pi_i\pi_k(c_i+c_k)/2$ for $\pi_{ik}$, the third term in (6.47) is

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N}(\pi_{ik}-\pi_i\pi_k)\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)\left(\frac{t_k}{\pi_k}-\frac{t}{n}\right)\\
&= \sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N}\pi_i\pi_k\frac{c_i+c_k-2}{2}\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)\left(\frac{t_k}{\pi_k}-\frac{t}{n}\right)\\
&= \sum_{i=1}^{N}\sum_{k=1}^{N}\pi_i\pi_k\frac{c_i+c_k-2}{2}\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)\left(\frac{t_k}{\pi_k}-\frac{t}{n}\right)-\sum_{i=1}^{N}\pi_i^2(c_i-1)\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2\\
&= \sum_{i=1}^{N}\pi_i^2(1-c_i)\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2.
\end{aligned}$$

Then, from (6.47),

$$\begin{aligned}
V(\hat{t}_{\mathrm{HT}}) &= \sum_{i=1}^{N}\pi_i\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2-\sum_{i=1}^{N}\pi_i^2\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2\\
&\quad+\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N}(\pi_{ik}-\pi_i\pi_k)\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)\left(\frac{t_k}{\pi_k}-\frac{t}{n}\right)\\
&\approx \sum_{i=1}^{N}\pi_i\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2-\sum_{i=1}^{N}\pi_i^2\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2\\
&\quad+\sum_{i=1}^{N}\pi_i^2(1-c_i)\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2\\
&= \sum_{i=1}^{N}\pi_i(1-c_i\pi_i)\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2.
\end{aligned}$$

If $c_i=\frac{n-1}{n-\pi_i}$, then the variance approximation in (6.48) for an SRS is

$$\begin{aligned}
\sum_{i=1}^{N}\pi_i(1-c_i\pi_i)\left(\frac{t_i}{\pi_i}-\frac{t}{n}\right)^2 &= \sum_{i=1}^{N}\frac{N}{n}\left(1-\frac{n-1}{N-1}\right)(t_i-\bar{t}_U)^2\\
&= \frac{N(N-1)}{n}\left(1-\frac{n-1}{N-1}\right)S_t^2.
\end{aligned}$$

If

$$c_i = \frac{n-1}{\left(1 - 2\pi_i + \frac{1}{n}\sum_{k=1}^N \pi_k^2\right)}$$

then

$$\sum_{i=1}^N \pi_i(1-c_i\pi_i)\left(\frac{t_i}{\pi_i} - \frac{t}{n}\right)^2 = \frac{N(N-1)}{n}\left(1 - \frac{n(n-1)}{N-n}\right)S_t^2.$$

**6.31** We wish to minimize

$$\frac{1}{n}\sum_{i=1}^N \psi_i\left(\frac{t_i}{\psi_i} - t\right)^2 + \frac{1}{n}\sum_{i=1}^N \frac{M_i^2 S_i^2}{m_i\psi_i}$$

subject to the constraint that

$$C = E\left[\sum_{i\in\mathcal{S}} m_i\right] = E\left[\sum_{i=1}^N Q_i m_i\right] = n\sum_{i=1}^N \psi_i m_i.$$

Using Lagrange multipliers, let

$$g(m_1,\ldots,m_N,\lambda) = \sum_{i=1}^N \frac{M_i^2S_i^2}{m_i\psi_i} - \lambda\left(C - n\sum_{i=1}^N \psi_i m_i\right).$$

$$\begin{aligned}\frac{\partial g}{\partial m_k} &= -\frac{M_k^2S_k^2}{m_k^2\psi_k} + n\lambda\psi_k \\ \frac{\partial g}{\partial \lambda} &= n\sum_{i=1}^N \psi_i m_i - C.\end{aligned}$$

Setting the partial derivatives equal to zero gives

$$m_k = \frac{1}{\sqrt{n\lambda}}\frac{M_kS_k}{\psi_k}$$

and

$$\sqrt{\lambda} = \frac{\sqrt{n}}{C}\sum_{i=1}^N M_iS_i.$$

Thus, the optimal allocation has $m_i \propto M_iS_i/\psi_i$. For comparison, a self-weighting design would have $m_i \propto M_i/\psi_i$.

**6.32** Let $M_i$ be the number of residential numbers in psu $i$. When you dial a number based on the method,

$$\begin{aligned}&P(\text{reach working number in psu } i \text{ on an attempt}) \\ &= P(\text{select psu } i)P(\text{get working number} \mid \text{select psu } i) \\ &= \frac{1}{N}\frac{M_i}{100}.\end{aligned}$$

Also,

$$P(\text{reach no one on an attempt}) = \sum_{i=1}^{N} \frac{1}{N}\left(1 - \frac{M_i}{100}\right) = 1 - \frac{M_0}{100N}.$$

Then,

$$\begin{aligned}
&P(\text{select psu } i \text{ as first in sample}) \\
&= P(\text{select psu } i \text{ on first attempt}) \\
&\quad + P(\text{reach no one on first attempt, select psu } i \text{ on second attempt}) \\
&\quad + P(\text{reach no one on first and second attempts,} \\
&\qquad \text{select psu } i \text{ on third attempt}) \\
&\quad + \ldots \\
&= \frac{1}{N}\frac{M_i}{100} + \frac{1}{N}\frac{M_i}{100}\left(1 - \frac{M_0}{100N}\right) + \frac{1}{N}\frac{M_i}{100}\left(1 - \frac{M_0}{100N}\right)^2 + \ldots \\
&= \frac{1}{N}\frac{M_i}{100}\sum_{j=0}^{\infty}\left(1 - \frac{M_0}{100N}\right)^j \\
&= \frac{1}{N}\frac{M_i}{100}\frac{1}{1 - \left(1 - \frac{M_0}{100N}\right)} \\
&= \frac{M_i}{M_0}
\end{aligned}$$

# Chapter 7

# Complex Surveys

**7.6** Here is SAS code for solving this problem. Note that for the population, we have $\bar{y}_U = 17.73$,

$$S^2 = 38.1451726 = \frac{1}{1999}\left[\sum_{i=1}^{2000} y_i^2 - \frac{1}{2000}\left(\sum_{i=1}^{2000} y_i\right)^2\right] = \left(704958 - \frac{35460^2}{2000}\right)/1999,$$

$\hat{\theta}_{25} = 13.098684, \hat{\theta}_{50} = 16.302326, \hat{\theta}_{75} = 19.847458.$

```
data integerwt;
   infile integer delimiter="," firstobs=2;
   input stratum y;
   ysq = y*y;
run;

/* Calculate the population characteristics for comparison */

proc means data=integerwt mean var;
   var y;
run;

proc surveymeans data=integerwt mean sum percentile = (25 50 75);
   var y ysq;
   /* Without a weight statement, SAS assumes all weights are 1 */
run;

proc glm data=integerwt;
   class stratum;
   model y = stratum;
   means stratum;
run;
```

```
/* Before selecting the sample,
   you need to sort the data set by stratum */

proc sort data=integerwt;
   by stratum;

proc surveyselect data=integerwt method=srs sampsize = (50 50 20 25)
                  out = stratsamp seed = 38572 stats;
   strata stratum;
run;

proc print data = stratsamp;
run;

data strattot;
   input stratum   _total_;
   datalines;
1 200
2 800
3 400
4 600
;

proc surveymeans data=stratsamp total = strattot mean clm sum
                                percentile = (25 50 75);
   strata stratum;
   weight SamplingWeight;
   var y ysq;
run;

/* Create a pseudo-population using the weights */

data pseudopop;
   set stratsamp;
   retain stratum  y;
   do i = 1 to SamplingWeight;
       output ;
   end;

proc means data=pseudopop mean var;
   var y;
run;

proc surveymeans data=pseudopop mean sum percentile = (25 50 75);
```

```
   var y ysq;
run;
```

The estimates from the last two surveymeans statements are the same (not the standard errors, however).

**7.7** Let $y$ = number of species caught.

| $y$ | $\hat{f}(y)$ | $y$ | $\hat{f}(y)$ |
|---|---|---|---|
| 1 | .0328 | 10 | .2295 |
| 3 | .0328 | 11 | .0491 |
| 4 | .0820 | 12 | .0820 |
| 5 | .0328 | 13 | .0328 |
| 6 | .0656 | 14 | .0164 |
| 7 | .0656 | 16 | .0328 |
| 8 | .1803 | 17 | .0164 |
| 9 | .0328 | 18 | .0164 |

Here is SAS code for constructing this table:

```
data nybight;
   infile nybight delimiter=',' firstobs=2;
   input year  stratum catchnum catchwt numspp depth temp ;
   select (stratum);
      when (1,2) relwt=1;
  when (3,4,5,6) relwt=2;
   end;
   if year = 1974;

/*Construct empirical probability mass function and empirical cdf.*/

proc freq data=nybight;
    tables numspp / out = htpop_epmf outcum;
weight relwt;

/*SAS proc freq gives values in percents, so we divide each by 100*/

data htpop_epmf;
    set htpop_epmf;
epmf = percent/100;
ecdf =  cum_pct/100;

proc print data=htpop_epmf;
run;
```

**7.8** We first construct a new variable, *weight*, with the following values:

| Stratum | weight |
|---|---|
| large | $\frac{245}{23}\ \frac{M_i}{m_i}$ |
| sm/me | $\frac{66}{8}\ \frac{M_i}{m_i}$ |

Because there is nonresponse on the variable *hrwork*, for this exercise we take $m_i$ to be the number of respondents in that cluster. The weights for each teacher sampled in a school are given in the following table:

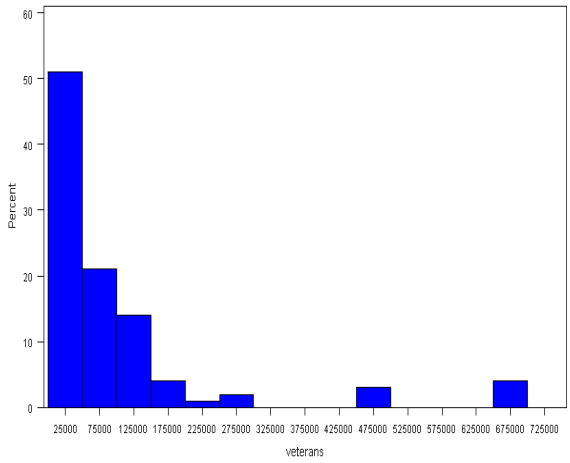
| dist | school | popteach | $m_i$ | weight |
|---|---|---|---|---|
| sm/me | 1 | 2 | 1 | 16.50000 |
| sm/me | 2 | 6 | 4 | 12.37500 |
| sm/me | 3 | 18 | 7 | 21.21429 |
| sm/me | 4 | 12 | 7 | 14.14286 |
| sm/me | 6 | 24 | 11 | 18.00000 |
| sm/me | 7 | 17 | 4 | 35.06250 |
| sm/me | 8 | 19 | 5 | 31.35000 |
| sm/me | 9 | 28 | 21 | 11.00000 |
| large | 11 | 33 | 10 | 35.15217 |
| large | 12 | 16 | 13 | 13.11037 |
| large | 13 | 22 | 3 | 78.11594 |
| large | 15 | 24 | 24 | 10.65217 |
| large | 16 | 27 | 24 | 11.98370 |
| large | 18 | 18 | 2 | 95.86957 |
| large | 19 | 16 | 3 | 56.81159 |
| large | 20 | 12 | 8 | 15.97826 |
| large | 21 | 19 | 5 | 40.47826 |
| large | 22 | 33 | 13 | 27.04013 |
| large | 23 | 31 | 16 | 20.63859 |
| large | 24 | 30 | 9 | 35.50725 |
| large | 25 | 23 | 8 | 30.62500 |
| large | 28 | 53 | 17 | 33.20972 |
| large | 29 | 50 | 8 | 66.57609 |
| large | 30 | 26 | 22 | 12.58893 |
| large | 31 | 25 | 18 | 14.79469 |
| large | 32 | 23 | 16 | 15.31250 |
| large | 33 | 21 | 5 | 44.73913 |
| large | 34 | 33 | 7 | 50.21739 |
| large | 36 | 25 | 4 | 66.57609 |
| large | 38 | 38 | 10 | 40.47826 |
| large | 41 | 30 | 2 | 159.78261 |

The epmf is given below, with $y$ =hrwork.

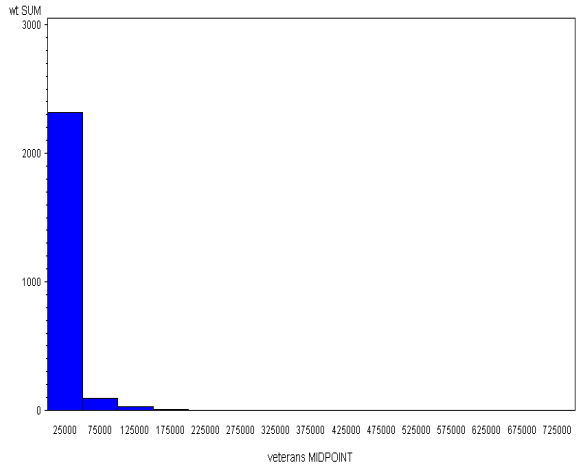
| $y$ | $\hat{f}(y)$ | $y$ | $\hat{f}(y)$ |
|---|---|---|---|
| 20.00 | 0.0040 | 34.55 | 0.0019 |
| 26.25 | 0.0274 | 34.60 | 0.0127 |
| 26.65 | 0.0367 | 35.00 | 0.1056 |
| 27.05 | 0.0225 | 35.40 | 0.0243 |
| 27.50 | 0.0192 | 35.85 | 0.0164 |
| 27.90 | 0.0125 | 36.20 | 0.0022 |
| 28.30 | 0.0050 | 36.25 | 0.0421 |
| 29.15 | 0.0177 | 36.65 | 0.0664 |
| 30.00 | 0.0375 | 37.05 | 0.0023 |
| 30.40 | 0.0359 | 37.10 | 0.0403 |
| 30.80 | 0.0031 | 37.50 | 0.1307 |
| 31.25 | 0.0662 | 37.90 | 0.0079 |
| 32.05 | 0.0022 | 37.95 | 0.0019 |
| 32.10 | 0.0031 | 38.35 | 0.0163 |
| 32.50 | 0.0370 | 38.75 | 0.0084 |
| 32.90 | 0.0347 | 39.15 | 0.0152 |
| 33.30 | 0.0031 | 40.00 | 0.0130 |
| 33.35 | 0.0152 | 40.85 | 0.0018 |
| 33.75 | 0.0404 | 41.65 | 0.0031 |
| 34.15 | 0.0622 | 52.50 | 0.0020 |

**7.10**

Without weights



With weights




Using the weights makes a huge difference, since the counties with large numbers of veterans also have small weights.

**7.13** The variable *agefirst* contains information on the age at first arrest. Missing values are coded as 99; for this exercise, we use the non-missing cases.

| Estimated Quantity | Without Weights | With Weights |
|---|---|---|
| Mean | 13.07 | 13.04 |
| Median | 13 | 13 |
| 25th Percentile | 12 | 12 |
| 75th Percentile | 15 | 15 |

Calculating these quantities in SAS is easy: simply include the weight variable in PROC UNIVARIATE.

The weights change the estimates very little, largely because the survey was designed to be self-weighting.

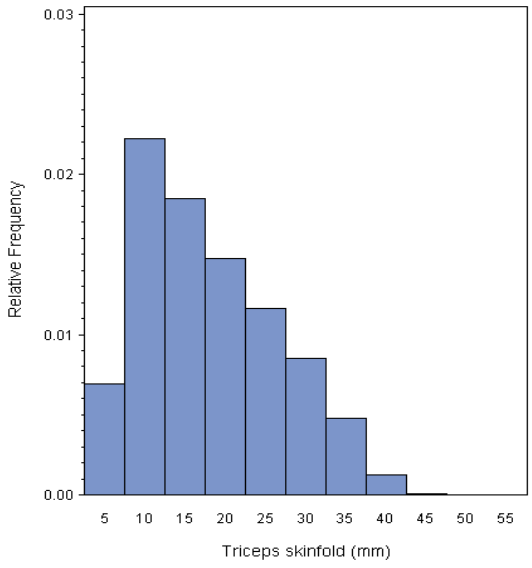
**7.14**

| Quantity | Variable | $\hat{p}$ |
|---|---|---|
| Age $\leq 14$ | age | .1233 |
| Violent offense | crimtype | .4433 |
| Both parents | livewith | .2974 |
| Male | sex | .9312 |
| Hispanic | ethnicity | .1888 |
| Single parent | livewith | .5411 |
| Illegal drugs | everdrug | .8282 |

**7.15** (a) We use the following SAS code to obtain $\hat{\bar{y}} = 18.03$, with 95% CI [17.48, 18.58].

```
data nhanes;
  infile nhanes delimiter=',' firstobs=2;
  input sdmvstra sdmvpsu wtmec2yr age ridageyr riagendr ridreth2
      dmdeduc indfminc bmxwt bmxbmi bmxtri
      bmxwaist bmxthicr  bmxarml;
  label age = "Age at Examination (years)"
        riagendr = "Gender"
        ridreth2 = "Race/Ethnicity"
        dmdeduc = "Education Level"
        indfminc = "Family income"
        bmxwt = "Weight (kg)"
        bmxbmi = "Body mass index"
        bmxtri = "Triceps skinfold (mm)"
        bmxwaist = "Waist circumference (cm)"
        bmxthicr = "Thigh circumference (cm)"
        bmxarml = "Upper arm length (cm)";
run;

proc surveymeans data=nhanes mean clm percentile = (0 25 50 75 100);
  stratum sdmvstra;
  cluster sdmvpsu;
  weight wtmec2yr;
  var bmxtri age;
run;
```

(b) The data appear skewed.




(c) The SAS code in part (a) also gives the following.

| Percentile | Value | Std Error |
|---|---|---|
| Minimum | 2.8 | |
| 25 | 10.98 | 0.177 |
| 50 | 16.35 | 0.324 |
| 75 | 23.95 | 0.425 |
| Maximum | 44.6 | |

Men:

| Percentile | Value |
|---|---|
| Minimum | 2.80 |
| 25 | 9.19 |
| 50 | 12.92 |
| 75 | 18.11 |
| Maximum | 42.4 |

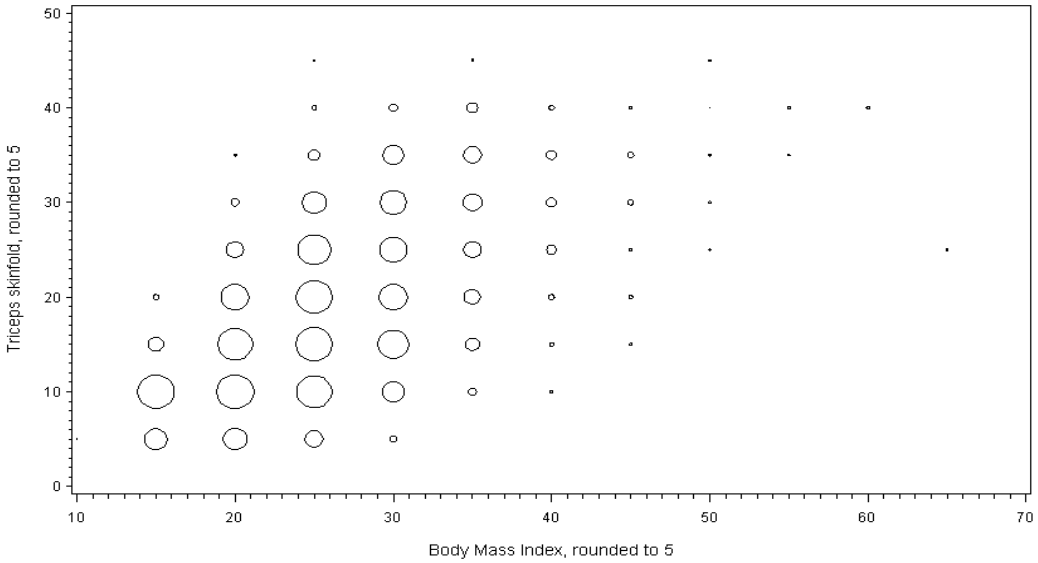
Women:

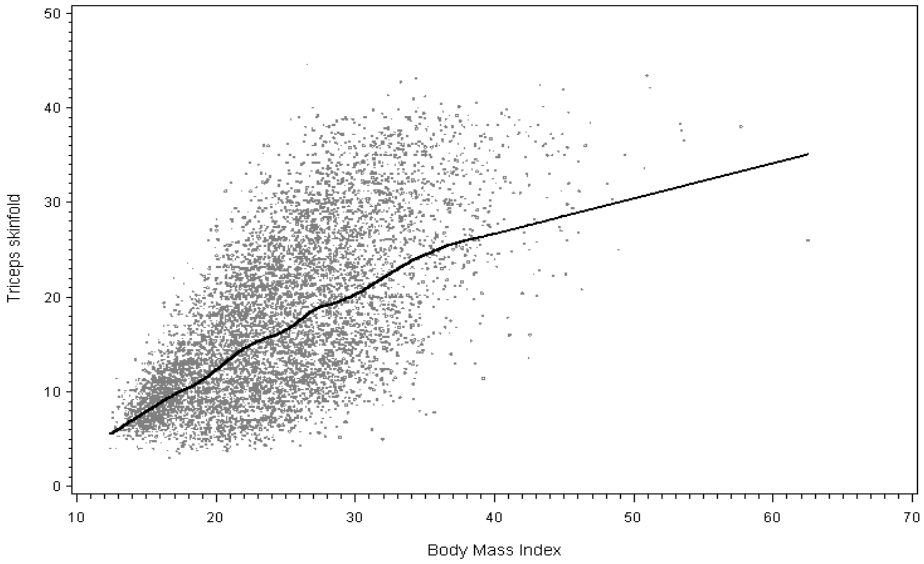
| Percentile | Value |
|---|---|
| Minimum | 4.00 |
| 25 | 14.92 |
| 50 | 21.94 |
| 75 | 28.36 |
| Maximum | 44.6 |

(d) Here is SAS code for constructing the plots:

```
data groupage;
   set nhanes;
   bmigroup = round(bmxbmi,5);
   trigroup = round(bmxtri,5);
run;

proc sort data=groupage;
   by bmigroup trigroup;

proc means data=groupage;
   by bmigroup trigroup;
   var  wtmec2yr;
   output out=circleage  sum=sumwts;

goptions reset=all;
goptions colors = (black);
axis3 label=('Body Mass Index, rounded to 5') order=(10 to 70 by 10);
axis4 label=(angle=90 'Triceps skinfold, rounded to 5')
             order=(0 to 55 by 10);

/* This gives the weighted circle plot */
```

```
proc gplot data=circleage;
   bubble trigroup * bmigroup= sumwts/
                    bsize=12 haxis = axis3 vaxis = axis4;
run;

/* The following draws the bubble plot with trend line */

ods graphics on;
proc loess data=nhanes;
   model bmxtri=bmxbmi / degree = 1 select=gcv;
   weight  wtmec2yr;
   ods output  OutputStatistics = bmxsmooth ;
run;
ods graphics off;

proc print data=bmxsmooth;
run;

proc sort data=bmxsmooth;
   by bmxbmi;

goptions reset=all;
goptions colors = (gray);
axis4 label=(angle=90 'Triceps skinfold') order =  (0 to 55 by 10);
axis3 label=('Body Mass Index') order=(10 to 70 by 10);
axis5  order=(0 to 55 by 5) major=none minor=none value=none;
symbol interpol=join width=2 color = black;

/* Display the trend line with the bubble plot */

data plotsmth;
   set bubbleage bmxsmooth; /* concatenates the data sets */
run;

proc gplot data=plotsmth;
   bubble bmxtri*bmxbmi = sumwts/
               bsize=10 haxis = axis3 vaxis = axis4;
   plot2 Pred*bmxbmi/haxis = axis3 vaxis = axis5;
run;
```




**7.17** We define new variables that take on the value 1 if the person has been a victim of at least one violent crime and 0 otherwise, and another variable for injury. The SAS code and output follows.

```
data ncvs;
   infile  ncvs delimiter = ",";
   input age married sex race hispanic hhinc away employ numinc
       violent injury medtreat medexp robbery assault
        pweight pstrat ppsu;
   if violent > 0 then isviol = 1;
   else isviol = 0;
   if  injury > 0 then isinjure = 1;
```

```
   else isinjure = 0;
run;

proc surveymeans data=ncvs;
   weight pweight;
   strata pstrat;
   cluster ppsu;
   var numinc  isviol isinjure;
run;

proc surveymeans data=ncvs;
   weight pweight;
   strata pstrat;
   cluster ppsu;
   var  medexp;
   domain isinjure;
run;
```

The SURVEYMEANS Procedure

Data Summary

| | |
|---|---|
| Number of Strata | 143 |
| Number of Clusters | 286 |
| Number of Observations | 79360 |
| Sum of Weights | 226204704 |

Statistics

| Variable | N | Mean | Std Error of Mean | 95% CL for Mean | |
|---|---|---|---|---|---|
| numinc | 79360 | 0.070071 | 0.002034 | 0.06605010 | 0.07409164 |
| isviol | 79360 | 0.013634 | 0.000665 | 0.01232006 | 0.01494718 |
| isinjure | 79360 | 0.003754 | 0.000316 | 0.00312960 | 0.00437754 |

Domain Analysis: isinjure

| isinjure | Variable | N | Mean | Std Error of Mean | 95% CL for Mean | |
|---|---|---|---|---|---|---|
| 0 | medexp | 79093 | 0 | 0 | 0.0000000 | 0.000000 |
| 1 | medexp | 267 | 101.6229 | 33.34777 | 35.7046182 | 167.541160 |

**7.18** Note that $\sum_{q_1 \le y \le q_2} yf(y)$ is the sum of the middle $N(1-2\alpha)$ observations in the population divided by $N$, and $\sum_{q_1 \le y \le q_2} f(y) = F(q_2) - F(q_1) \approx 1 - 2\alpha$. Consequently,

$$\bar{y}_{U\alpha} = \frac{\text{sum of middle } N(1-2\alpha) \text{ observations in the population}}{N(1-2\alpha)}.$$

To estimate the trimmed mean, substitute $\hat{f}$, $\hat{q}_1$, and $\hat{q}_2$ for $f$, $q_1$, and $q_2$.

**7.21** As stated in Section 7.1, the $y_i$'s are the measurements on observation units. If unit $i$ is in stratum $h$, then $w_i = N_h/n_h$. To express this formally, let

$$x_{hi} = \begin{cases} 1 & \text{if unit } i \text{ is in stratum } h \\ 0 & \text{otherwise.} \end{cases}$$

Then we can write

$$w_i = \sum_{h=1}^{H} \frac{N_h}{n_h} x_{hi}$$

and

$$\begin{aligned}
\textstyle\sum_y y\hat{f}(y) &= \frac{\sum_{i\in\mathcal{S}} y_i w_i}{\sum_{i\in\mathcal{S}} w_i} \\
&= \frac{\sum_{i\in\mathcal{S}} y_i \sum_{h=1}^{H} (N_h/n_h) x_{hi}}{\sum_{i\in\mathcal{S}} \sum_{h=1}^{H} (N_h/n_h) x_{hi}} \\
&= \frac{\sum_{h=1}^{H} N_h \sum_{i\in\mathcal{S}} (x_{hi} y_i/n_h)}{\sum_{h=1}^{H} N_h \sum_{i\in\mathcal{S}} (x_{hi}/n_h)} \\
&= \frac{\sum_{h=1}^{H} N_h \bar{y}_h}{\sum_{h=1}^{H} N_h} \\
&= \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h.
\end{aligned}$$

**7.22** For an SRS, $w_i = N/n$ for all $i$ and

$$\hat{f}(y) = \frac{\sum_{i\in\mathcal{S}: y_i = y} \frac{N}{n}}{\sum_{i\in\mathcal{S}} \frac{N}{n}}.$$

Thus,

$$\begin{aligned}\sum_y y^2 \hat{f}(y) &= \sum_{i\in\mathcal{S}} \frac{y_i^2}{n},\\ \sum_y y \hat{f}(y) &= \sum_{i\in\mathcal{S}} \frac{y_i}{n} = \bar{y},\end{aligned}$$

and

$$\begin{aligned}\hat{S}^2 &= \frac{N}{N-1}\left\{\sum_y y^2\hat{f}(y) - \left[\sum_y y\hat{f}(y)\right]^2\right\}\\ &= \frac{N}{N-1}\left\{\sum_{i\in\mathcal{S}}\frac{y_i^2}{n} - \bar{y}^2\right\}\\ &= \frac{N}{N-1}\sum_{i\in\mathcal{S}}\frac{(y_i-\bar{y})^2}{n}\\ &= \frac{N}{N-1}\frac{n-1}{n}s^2.\end{aligned}$$

If $n < N$, $\hat{S}^2$ is smaller than $s^2$ (although they will be close if $n$ is large).

**7.23** We need to show that the inclusion probability is the same for every unit in $\mathcal{S}_2$. Let $Z_i = 1$ if $i \in \mathcal{S}$ and 0 otherwise, and let $D_i = 1$ if $i \in \mathcal{S}_2$ and 0 otherwise. We have $P(Z_i = 1) = \pi_i$ and $P(D_i = 1 \mid Z_i = 1) \propto 1/\pi_i$.

$$\begin{aligned}P(i \in \mathcal{S}_2) &= P(Z_i = 1, D_i = 1)\\ &= P(D_i = 1 \mid Z_i = 1)P(Z_i = 1)\\ &\propto \frac{1}{\pi_i}\pi_i = 1.\end{aligned}$$

**7.24** A rare disease affects only a few children in the population. Even if all cases belong to the same cluster, a disease with estimated incidence of 2.1 per 1,000 is unlikely to affect all children in that cluster.

**7.25** (a) Inner-city areas are sampled at twice the rate of non-inner-city areas. Thus the selection probability for a household not in the inner city is one-half the selection probability for a household in the inner city. The relative weight for a non-inner-city household, then, is 2.

(b) Let $\pi$ represent the probability that a household in the inner city is selected. Then, for 1-person inner city households,

$$P \text{ (person selected } \mid \text{ household selected) } P \text{ (household selected)} = 1 \times \pi.$$

For $k$-person inner-city households,

$$P \text{ (person selected } \mid \text{ household selected) } P \text{ (household selected)} = \frac{1}{k}\pi.$$

Thus the relative weight for a person in an inner-city household is the number of adults in the household. The relative weight for a person in a non-inner-city household is $2 \times$ (number of adults in household).

The table of relative weights is:

| Number of adults | Inner city | Non-inner city |
|---|---|---|
| 1 | 1 | 2 |
| 2 | 2 | 4 |
| 3 | 3 | 6 |
| 4 | 4 | 8 |
| 5 | 5 | 10 |

# Chapter 8

# Nonresponse

**8.1** (a) Oversampling the low-income families is a form of substitution. One advantage of substitution is that the number of low-income families in the sample is larger. The main drawback, however, is that the low-income families that respond may differ from those that do not respond. For example, mothers who work outside the home may be less likely to breast feed *and* less likely to respond to the survey.

(b) The difference between percentage of mothers with one child indicates that the weighting does not completely adjust for the nonresponse.

(c) Weights were used to try to adjust for nonresponse in this survey. We can never know whether the adjustment is successful, however, unless we have some data from the nonrespondents. The response rate for the survey decreased from 54% in 1984 to 46% in 1989. It might have been better for the survey researchers to concentrate on increasing the response rate and obtaining accurate responses instead of tripling the sample size.

Because the survey was poststratified using ethnic background, age, and education, the weighted counts *must* agree with census figures for those variables. A possible additional variable to use for poststratification would be number of children.

**8.2** (a) The respondents report a total of

$$\sum y_i = (66)(32) + (58)(41) + (26)(54) = 5894$$

hours of TV, with

$$\sum y_i^2 = [65(15)^2 + 66(32)^2] + [57(19)^2 + 58(41)^2] + [25(25)^2 + 26(54)^2] = 291725.$$

Then, for the respondents,

$$\begin{aligned} \bar{y} &= \frac{5894}{150} = 39.3 \\ s^2 &= \frac{291725 - (150)(39.3)^2}{149} = 403.6 \end{aligned}$$

and

$$\mathrm{SE}(\bar{y}) = \sqrt{\left(1 - \frac{150}{2000}\right)\frac{403.6}{150}} = 1.58.$$

Note that this is technically a ratio estimate, since the number of respondents (here, 150) would vary if a different sample were taken. We are estimating the average hours of TV watched in the domain of respondents.

(b)

| GPA Group | Respondents | Non respondents | Total |
|---|---|---|---|
| 3.00–4.00 | 66 | 9 | 75 |
| 2.00–2.99 | 58 | 14 | 72 |
| Below 2.00 | 26 | 27 | 53 |
| Total | 150 | 50 | 200 |

$$\begin{aligned} X^2 &= \sum_{\text{cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{[66 - (.75)(75)]^2}{(.75)(75)} + \frac{[9 - (.25)(75)]^2}{(.25)(75)} + \cdots + \frac{[27 - (.25)(53)]^2}{(.25)(53)} \\ &= 1.69 + 5.07 + 0.30 + 0.89 + 4.76 + 14.27 \\ &= 26.97 \end{aligned}$$

Comparing the test statistic to a $\chi^2$ distribution with 2 df, the $p$-value is $1.4 \times 10^{-6}$. This is strong evidence against the null hypothesis that the three groups have the same response rates.

The hypothesis test indicates that the nonresponse is not MCAR, because response rates appear to be related to GPA. We do not know whether the nonresponse is MAR, or whether is it nonignorable.

(c)

$$\begin{aligned} \mathrm{SSB} &= \sum_{i=1}^{3}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2 = 9303.1 \\ \mathrm{MSW} &= s^2 = 403.6. \end{aligned}$$

The ANOVA table is as follows:

| Source | df | SS | MS | F | $p$-value |
|---|---|---|---|---|---|
| Between groups | 2 | 9303.1 | 4651.5 | 11.5 | 0.0002 |
| Within groups | 147 | 59323.0 | 403.6 | | |
| Total, about mean | 149 | 68626.1 | | | |

Both the nonresponse rate and the TV viewing appear to be related to GPA, so it would be a reasonable variable to consider for weighting class adjustment or poststratification.

(d) The initial weight for each person in the sample is 2000/200=10. After increasing the weights for the respondents in each class to adjust for the nonrespondents, the

weight for each respondent with GPA $\geq 3$ is

$$\text{initial weight} \times \frac{\text{sum of weights for sample}}{\text{sum of weights for respondents}} = 10 \times \frac{75(10)}{66(10)} = 11.36364.$$

| Sample Size | Number of Respondents ($n_R$) | Weight for each Respondent ($w$) | $\bar{y}$ | $wn_R\bar{y}$ | $wn_R$ |
|---|---|---|---|---|---|
| 75 | 66 | 11.36364 | 32 | 24000 | 750 |
| 72 | 58 | 12.41379 | 41 | 29520 | 720 |
| 53 | 26 | 20.38462 | 54 | 28620 | 530 |
| 200 | 150 | | | 82140 | 2000 |

Then $\hat{t}_{wc} = 82140$ and $\bar{y}_{wc} = 82140/2000 = 41.07$.

The weighting class adjustment leads to a higher estimate of average viewing time, because the GPA group with the highest TV viewing also has the most nonresponse.

(e) The poststratified weight for each respondent with GPA $\geq 3$ is

$$w_{\text{post}} = \text{initial weight} \times \frac{\text{population count}}{\text{sum of respondent weights}} = 10 \times \frac{700}{(10)(66)} = 10.60606.$$

Here, $n_R$ denotes number of respondents.

| $n_R$ | Population Count | $w_{\text{post}}$ | $\bar{y}$ | $w_{\text{post}}\bar{y}n_R$ | $w_{\text{post}}n_R$ |
|---|---|---|---|---|---|
| 66 | 700 | 10.60606 | 32 | 22400 | 700 |
| 58 | 800 | 13.79310 | 41 | 32800 | 800 |
| 26 | 500 | 19.2307 | 54 | 27000 | 500 |
| 150 | 2000 | | | 82200 | 2000 |

The last column is calculated to check the weights constructed—the sum of the poststratified weights in each poststratum should equal the population count for that poststratum.

$$\hat{t}_{\text{post}} = 82140$$

and

$$\hat{y}_{\text{post}} = \frac{82140}{2000} = 41.07.$$

**8.6** (a) For this exercise, we classify the missing data in the "Other/Unknown" category. Typically, raking would be used in situations in which the classification variables were known (and known to be accurate) for all respondents.

| | Population | Respondents | Response Rate (%) |
|---|---|---|---|
| Ph.D. | 10235 | 3036 | 30 |
| Master's | 7071 | 1640 | 23 |
| Other/Unknown | 1303 | 325 | 25 |
| Industry | 5397 | 1809 | 34 |
| Academia | 6327 | 2221 | 35 |
| Government | 2047 | 880 | 43 |
| Other/Unknown | 4838 | 91 | 19 |

These response rates are pretty dismal. The nonresponse does not appear to be MCAR, as it differs by degree and by type of employment. I doubt that it is MAR—I think that more information than is known from this survey would be needed to predict the nonresponse.

(b) The cell counts from the sample are:

| | Industry | Academia | Other | |
|---|---|---|---|---|
| PhD | 798 | 1787 | 451 | 3036 |
| non-PhD | 1011 | 434 | 520 | 1965 |
| | 1809 | 2221 | 971 | 5001 |

The initial sum of weights for each cell are:

| | Industry | Academia | Other | |
|---|---|---|---|---|
| PhD | 2969.4 | 6649.5 | 1678.2 | 11297.1 |
| non-PhD | 3762.0 | 1614.9 | 1934.9 | 7311.9 |
| | 6731.4 | 8264.5 | 3613.1 | 18609.0 |

After adjusting for the population row counts (10235 for Ph.D. and 8374 for non-Ph.D.) the new table is:

| | Industry | Academia | Other | |
|---|---|---|---|---|
| PhD | 2690.2 | 6024.4 | 1520.4 | 10235 |
| non-PhD | 4308.5 | 1849.5 | 2216.0 | 8374 |
| | 6998.7 | 7873.9 | 3736.4 | 18609 |

Raking to the population column totals (Industry, 5397; Academia, 6327; Other, 6885) gives:

| | Industry | Academia | Other | |
|---|---|---|---|---|
| PhD | 2074.6 | 4840.8 | 2801.6 | 9717.0 |
| non-PhD | 3322.4 | 1486.2 | 4083.4 | 8892.0 |
| | 5397.0 | 6327.0 | 6885.0 | 18609.0 |

As you can see, the previous two tables are still far apart. After iterating, the final table of the weight sums is:

| | Industry | Academia | Other | |
|---|---|---|---|---|
| PhD | 2239.2 | 4980.6 | 3015.2 | 10235.0 |
| non-PhD | 3157.8 | 1346.4 | 3869.8 | 8374.0 |
| | 5397.0 | 6327.0 | 6885.0 | 18609.0 |

The raking has dramatically increased the weights in the "Other" employment category.

To calculate the proportions using the raking weights, create a new variable *weight*. For respondents with PhD's who work in industry, weight = 2239.2/798 = 2.806.

For the question, "Should the ASA develop some sort of certification?" the estimated percentages are:

| | Without Weights | With Raking Weights |
|---|---|---|
| No response | 0.2 | 0.3 |
| Yes | 26.4 | 25.8 |
| Possibly | 22.3 | 22.3 |
| No opinion | 5.4 | 5.4 |
| Unlikely | 6.7 | 6.9 |
| No | 39.0 | 39.3 |

(c) I think such a conclusion is questionable because of the very high nonresponse rate. This survey is closer to a self-selected opinion poll than to a probability sample.

**8.7**

| Discipline | Response Rate (%) | Female Members (%) |
|---|---|---|
| Literature | 69.5 | 38 |
| Classics | 71.2 | 27 |
| Philosophy | 73.1 | 18 |
| History | 71.5 | 19 |
| Linguistics | 73.9 | 36 |
| Political Science | 69.0 | 13 |
| Sociology | 71.4 | 26 |

The model implicitly adopted in Example 4.3 was that nonrespondents within each stratum were similar to respondents in that stratum.

We can use a $\chi^2$ test to examine whether the nonresponse rate varies among strata. The observed counts are given in the following table, with expected counts in parentheses:

| | Respondent | | Nonrespondent | | |
|---|---|---|---|---|---|
| Literature | 636 | (651.6) | 279 | (263.4) | 915 |
| Classics | 451 | (450.8) | 182 | (182.2) | 633 |
| Philosophy | 481 | (468.6) | 177 | (189.4) | 658 |
| History | 611 | (608.9) | 244 | (246.1) | 855 |
| Linguistics | 493 | (475.0) | 174 | (192.0) | 667 |
| Political Science | 575 | (593.2) | 258 | (239.8) | 833 |
| Sociology | 588 | (586.8) | 236 | (237.2) | 824 |
| | 3835 | | 1550 | | 5385 |

The Pearson test statistic is

$$X^2 = \sum_{\text{cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = 6.8$$

Comparing the test statistic to a $\chi^2_6$ distribution, we calculate $p$-value 0.34. There is no evidence that the response rates differ among strata.

The estimated correlation coefficient of the response rate and the percent female members is 0.19. Performing a hypothesis test for association (Pearson correlation, Spearman correlation, or Kendall's $\tau$) gives $p$-value $> .10$. There is no evidence that the response rate is associated with the percentage of members who are female.

**8.12** (a) The overall response rate, using the file *teachmi.dat*, was 310/754=0.41.

(b) As with many nonresponse problems, it's easy to think of plausible reasons why the nonresponse bias might go either direction. The teachers who work many hours may be working so hard they are less likely to return the survey, or they may be more conscientious and thus more likely to return it.

(c) The means and variances from the file `teachnr.dat` (ignoring missing values) are

| | hrwork | size | preprmin | assist |
|---|---|---|---|---|
| responses | 26 | 25 | 26 | 26 |
| $\bar{y}$ | 36.46 | 24.92 | 160.19 | 152.31 |
| $s^2$ | 2.61 | 25.74 | 3436.96 | 49314.46 |
| $\hat{V}(\bar{y})$ | 0.10 | 1.03 | 132.19 | 1896.71 |

The corresponding estimates from *teachers.dat*, the original cluster sample, are:

| | hrwork | size | preprmin | assist |
|---|---|---|---|---|
| $\hat{\bar{y}}_r$ | 33.82 | 26.93 | 168.74 | 52.00 |
| $\hat{V}(\hat{\bar{y}}_r)$ | 0.50 | 0.57 | 70.57 | 228.96 |

**8.14** (a) We are more likely to delete an observation if the value of $x_i$ is small. Since $x_i$ and $y_i$ are positively correlated, we expect the mean of $y$ to be too big.

(b) The population mean of *acres92* is $\bar{y}_U = 308582$.

**8.16** We use the approximations from Chapter 3 to obtain:

$$\begin{aligned} E[\hat{\bar{y}}_R] &= E\left[\frac{\sum_{i=1}^N Z_i R_i w_i y_i}{\sum_{i=1}^N \phi_i}\left\{1 - \frac{\sum_{i=1}^N (Z_i R_i w_i - \phi_i)}{\sum_{i=1}^N Z_i R_i w_i}\right\}\right] \\ &\approx \frac{\sum_{i=1}^N \phi_i y_i}{\sum_{i=1}^N \phi_i} \\ &\approx \frac{1}{N\bar{\phi}_U}\sum_{i=1}^N \phi_i y_i. \end{aligned}$$

Thus the bias is

$$\begin{aligned} \text{Bias}\,[\hat{\bar{y}}_R] &\approx \frac{1}{N\bar{\phi}_U}\sum_{i=1}^N (\phi_i - \bar{\phi}_U) y_i \\ &= \frac{1}{N\bar{\phi}_U}\sum_{i=1}^N (\phi_i - \bar{\phi}_U)(y_i - \bar{y}_U) \\ &\approx \frac{1}{\bar{\phi}_U}\text{Cov}\,(\phi_i, y_i) \end{aligned}$$

**8.17** The argument is similar to the previous exercise. If the classes are sufficiently large, then $E[1/\tilde{\phi}_c] \approx 1/\bar{\phi}_c$.

**8.19**

$$\begin{aligned} & V(\hat{\bar{y}}_{wc}) \\ &= V\left[\frac{n_1}{n}\frac{1}{n_{1R}}\sum_{i=1}^N Z_i R_i x_i y_i + \frac{n_2}{n}\frac{1}{n_{2R}}\sum_{i=1}^N Z_i R_i (1-x_i) y_i\right] \\ &= E\left\{V\left[\frac{n_1}{n}\frac{1}{n_{1R}}\sum_{i=1}^N Z_i R_i x_i y_i + \frac{n_2}{n}\frac{1}{n_{2R}}\sum_{i=1}^N Z_i R_i (1-x_i) y_i \middle| Z_1, \ldots, Z_N\right]\right\} \\ &\quad + V\left\{E\left[\frac{n_1}{n}\frac{1}{n_{1R}}\sum_{i=1}^N Z_i R_i x_i y_i + \frac{n_2}{n}\frac{1}{n_{2R}}\sum_{i=1}^N Z_i R_i (1-x_i) y_i \middle| Z_1, \ldots, Z_N\right]\right\} \\ &= E\left\{V\left[\frac{n_1}{n}\frac{\sum_{i=1}^N Z_i R_i x_i y_i}{\sum_{i=1}^N Z_i R_i x_i} + \frac{n_2}{n}\frac{\sum_{i=1}^N Z_i R_i (1-x_i) y_i}{\sum_{i=1}^N Z_i R_i (1-x_i)} \middle| Z_1, \ldots, Z_N\right]\right\} \\ &\quad + V\left\{E\left[\frac{n_1}{n}\frac{\sum_{i=1}^N Z_i R_i x_i y_i}{\sum_{i=1}^N Z_i R_i x_i} + \frac{n_2}{n}\frac{\sum_{i=1}^N Z_i R_i (1-x_i) y_i}{\sum_{i=1}^N Z_i R_i (1-x_i)} \middle| Z_1, \ldots, Z_N\right]\right\}. \end{aligned}$$

We use the ratio approximations from Chapter 4 to find the approximate expected values and variances.

$$\begin{aligned}
&E\left[\frac{n_1}{n}\frac{\sum_{i=1}^N Z_i R_i x_i y_i}{\sum_{i=1}^N Z_i R_i x_i}\middle| Z_1,\ldots,Z_N\right] \\
&= \frac{n_1}{n}E\left[\frac{\sum_{i=1}^N Z_i R_i x_i y_i}{\phi_1\sum_{i=1}^N Z_i x_i}\left(1-\frac{\sum_{i=1}^N Z_i[R_i-\phi_1]x_i}{\sum_{i=1}^N Z_i R_i x_i}\right)\middle| Z_1,\ldots,Z_N\right] \\
&= \frac{1}{n\phi_1}E\left[\sum_{i=1}^N Z_i R_i x_i y_i - \sum_{i=1}^N Z_i R_i x_i y_i \frac{\sum_{i=1}^N Z_i[R_i-\phi_1]x_i}{\sum_{i=1}^N Z_i R_i x_i}\middle| Z_1,\ldots,Z_N\right] \\
&\approx \frac{1}{n}\sum_{i=1}^N Z_i x_i y_i - \frac{1}{(n\phi_1)^2}\sum_{i=1}^N Z_i V(R_i) x_i y_i \\
&\approx \frac{1}{n}\sum_{i=1}^N Z_i x_i y_i.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&V\left\{E\left[\frac{n_1}{n}\frac{\sum_{i=1}^N Z_i R_i x_i y_i}{\sum_{i=1}^N Z_i R_i x_i} + \frac{n_2}{n}\frac{\sum_{i=1}^N Z_i R_i (1-x_i) y_i}{\sum_{i=1}^N Z_i R_i (1-x_i)}\middle| Z_1,\ldots,Z_N\right]\right\} \\
&\approx V\left\{\frac{1}{n}\sum_{i=1}^N Z_i x_i y_i + \frac{1}{n}\sum_{i=1}^N Z_i (1-x_i) y_i\right\} \\
&= \left(1-\frac{n}{N}\right)\frac{S_y^2}{n},
\end{aligned}$$

the variance that would be obtained if there were no nonresponse. For the other term,

$$\begin{aligned}
&V\left[\frac{n_1}{n}\frac{\sum_{i=1}^N Z_i R_i x_i y_i}{\sum_{i=1}^N Z_i R_i x_i}\middle| Z_1,\ldots,Z_N\right] \\
&= V\left[\frac{1}{n\phi_1}\sum_{i=1}^N Z_i R_i x_i y_i\left(1-\frac{\sum_{i=1}^N Z_i[R_i-\phi_1]x_i}{\sum_{i=1}^N Z_i R_i x_i}\right)\middle| Z_1,\ldots,Z_N\right] \\
&\approx \frac{1}{(n\phi_1)^2}\sum_{i=1}^N Z_i V(R_i) x_i y_i^2 \\
&\approx \frac{\phi_1(1-\phi_1)}{(n\phi_1)^2}\sum_{i=1}^N Z_i x_i y_i^2.
\end{aligned}$$

Thus,

$$
\begin{aligned}
& E\left\{V\left[\frac{n_1}{n}\frac{\sum_{i=1}^N Z_i R_i x_i y_i}{\sum_{i=1}^N Z_i R_i x_i} + \frac{n_2}{n}\frac{\sum_{i=1}^N Z_i R_i (1-x_i) y_i}{\sum_{i=1}^N Z_i R_i (1-x_i)}\middle| Z_1, \ldots, Z_N\right]\right\} \\
\approx\ & E\left\{\frac{\phi_1(1-\phi_1)}{(n\phi_1)^2}\sum_{i=1}^N Z_i x_i y_i^2 + \frac{\phi_2(1-\phi_2)}{(n\phi_2)^2}\sum_{i=1}^N Z_i (1-x_i) y_i^2\right\} \\
=\ & \frac{\phi_1(1-\phi_1)}{n\phi_1^2}\sum_{i=1}^N x_i y_i^2 + \frac{\phi_2(1-\phi_2)}{n\phi_2^2}\sum_{i=1}^N (1-x_i) y_i^2.
\end{aligned}
$$

**8.20** (a) Respondents are divided into 5 classes on the basis of the number of nights the respondent was home during the 4 nights preceding the survey call.

The sampling weight $w_i$ for respondent $i$ is then multiplied by $5/(k_i + 1)$. The respondents with $k = 0$ were only home on one of the five nights and are assigned to represent their share of the population plus the share of four persons in the sample who were called on one of their "unavailable" nights. The respondents most likely to be home have $k = 4$; it is presumed that all persons in the sample who were home every night were reached, so their weights are unchanged.

(b) This method of weighting is based on the premise that the most accessible persons will tend to be overrepresented in the survey data. The method is easy to use, theoretically appealing, and can be used in conjunction with callbacks. But it still misses people who were not at home on any of the five nights, or who refused to participate in the survey. Since in many surveys done over the telephone, nonresponse is due in large part to refusals, the HPS method may not be helpful in dealing with all nonresponse. Values of $k$ may also be in error, because people may err when recalling how many evenings they were home.

# Chapter 9

# Variance Estimation in Complex Surveys

**9.1** All of the methods discussed in this chapter would be appropriate. Note that the replication methods might slightly overestimate the variance because sampling is done without replacement, but since the sampling fractions are fairly small we expect the overestimation to be small.

**9.2** We calculate $\bar{y} = 8.23333333$ and $s^2 = 15.978$, so $s^2/30 = 0.5326$.

For jackknife replicate $j$, the jackknife weight is $w_{j(j)} = 0$ for observation $j$ and $w_{i(j)} = (30/29)w_i = (30/29)(100/30) = 3.44828$ for $i \neq j$. Using the jackknife weights, we find $\bar{y}_{(1)} = 8.2413, \ldots, \bar{y}_{(30)} = 8.20690$, so, by (9.8),

$$\hat{V}_{\text{JK}}(\bar{y}) = \frac{29}{30}\sum_{j=1}^{30}[\bar{y}_{(j)} - \bar{y}]^2 = 0.5326054.$$

**9.3** Here is the empirical cdf $\hat{F}(y)$:

| Obs | y | COUNT | PERCENT | CUM_FREQ | CUM_PCT | epmf | ecdf |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3.3333 | 3.3333 | 3.333 | 3.333 | 0.03333 | 0.03333 |
| 2 | 3 | 10.0000 | 10.0000 | 13.333 | 13.333 | 0.10000 | 0.13333 |
| 3 | 4 | 3.3333 | 3.3333 | 16.667 | 16.667 | 0.03333 | 0.16667 |
| 4 | 5 | 6.6667 | 6.6667 | 23.333 | 23.333 | 0.06667 | 0.23333 |
| 5 | 6 | 16.6667 | 16.6667 | 40.000 | 40.000 | 0.16667 | 0.40000 |
| 6 | 7 | 10.0000 | 10.0000 | 50.000 | 50.000 | 0.10000 | 0.50000 |
| 7 | 8 | 10.0000 | 10.0000 | 60.000 | 60.000 | 0.10000 | 0.60000 |
| 8 | 9 | 6.6667 | 6.6667 | 66.667 | 66.667 | 0.06667 | 0.66667 |
| 9 | 10 | 10.0000 | 10.0000 | 76.667 | 76.667 | 0.10000 | 0.76667 |
| 10 | 12 | 6.6667 | 6.6667 | 83.333 | 83.333 | 0.06667 | 0.83333 |
| 11 | 14 | 6.6667 | 6.6667 | 90.000 | 90.000 | 0.06667 | 0.90000 |
| 12 | 15 | 6.6667 | 6.6667 | 96.667 | 96.667 | 0.06667 | 0.96667 |

```
 13   17      3.3333      3.3333     100.000     100.000     0.03333     1.00000
```

Note that $\hat{F}(7) = .5$, so the median is $\hat{\theta}_{0.5} = 7$. No interpolation is needed.

As in Example 9.12, $\hat{F}(\theta_{1/2})$ is the sample proportion of observations that take on value at most $\theta_{1/2}$, so

$$\hat{V}[\hat{F}(\hat{\theta}_{1/2})] = \left(1 - \frac{n}{N}\right)\frac{0.25862069}{n} = \left(1 - \frac{30}{100}\right)\frac{0.25862069}{30} = 0.006034483.$$

This is a small sample, so we use the $t_{29}$ critical value of 2.045 to calculate

$$2.045\sqrt{\hat{V}[\hat{F}(\hat{\theta}_{1/2})]} = 0.1588596.$$

The lower confidence bound is $\hat{F}^{-1}(.5 - 0.1588596) = \hat{F}^{-1}(0.3411404)$ and the upper confidence bound for the median is $\hat{F}^{-1}(.5 + 0.1588596) = \hat{F}^{-1}(0.6588596)$. Interpolating, we have that the lower confidence bound is

$$5 + \frac{0.34114 - 0.23333}{0.4 - 0.23333}(6 - 5) = 5.6$$

and the upper confidence bound is

$$8 + \frac{0.6588596 - 0.6}{0.666667 - 0.6}(9 - 8) = 8.8.$$

Thus an approximate 95% CI is [5.6, 8.8].

SAS code below gives approximately the same interval:

```
data srs30;
   input y @@;
   wt = 100/30;
   datalines;
8  5  2  6  6  3  8  6 10  7 15  9 15  3  5  6
7 10 14  3  4 17 10  6 14 12  7  8 12  9
;

/* We use two methods.  First, the "hand" calculations" */

/* Find the empirical cdf */

proc freq data=srs30;
    tables y / out = htpop_epmf outcum;
weight wt;
run;
```

```
data htpop_epmf;
    set htpop_epmf;
epmf = percent/100;
ecdf =  cum_pct/100;
run;

proc print data=htpop_epmf;
run;

/* Find the variance of \hat{F}(median) */

data calcvar;
    set srs30;
ui = 0;
if y le 7 then ui = 1;
ei = ui - .5;

proc univariate data=calcvar;
    var ei;
run;

/* Calculate the stratified variance for the total of variable ei */
proc surveymeans data=calcvar total = 100 sum stderr;
    weight wt;
var ei;
run;

/* Method 2: Use sas directly to find the CI */

proc surveymeans data=srs30 total=100
                 percentile=(25 50 75) nonsymcl;
   weight wt;
   var y;
run;
```

Quantiles

| Variable | Percentile | Estimate | Std Error | 95% Confidence Limits | |
|---|---|---|---|---|---|
| y | 25% Q1 | 5.100000 | 0.770164 | 2.65604792 | 5.8063712 |
| | 50% Median | 7.000000 | 0.791213 | 5.64673564 | 8.8831609 |
| | 75% Q3 | 9.833333 | 1.057332 | 7.16875624 | 11.4937313 |

**9.5**

| | (a) Age | (b) Violent | (c) Bothpar | (d) Male | (e) Hispanic | (f) Sinpar | (g) Drugs |
|---|---|---|---|---|---|---|---|
| $\hat{\theta}_1$ | 0.12447 | 0.52179 | 0.29016 | 0.90160 | 0.30106 | 0.55691 | 0.90072 |
| $\hat{\theta}_2$ | 0.09528 | 0.43358 | 0.31309 | 0.84929 | 0.20751 | 0.52381 | 0.84265 |
| $\hat{\theta}_3$ | 0.08202 | 0.36733 | 0.34417 | 0.99319 | 0.17876 | 0.51068 | 0.82960 |
| $\hat{\theta}_4$ | 0.21562 | 0.37370 | 0.25465 | 0.96096 | 0.08532 | 0.55352 | 0.80869 |
| $\hat{\theta}_5$ | 0.21660 | 0.42893 | 0.30181 | 0.91314 | 0.14912 | 0.54480 | 0.74491 |
| $\hat{\theta}_6$ | 0.07321 | 0.48006 | 0.30514 | 0.96786 | 0.15752 | 0.55350 | 0.82232 |
| $\hat{\theta}_7$ | 0.02402 | 0.51201 | 0.27299 | 0.96558 | 0.25170 | 0.54490 | 0.84977 |
| $\hat{\theta}$ | 0.12330 | 0.44325 | 0.29743 | 0.93119 | 0.18877 | 0.54108 | 0.82821 |
| $\tilde{\theta}$ | 0.11875 | 0.44534 | 0.29743 | 0.93594 | 0.19014 | 0.54116 | 0.82838 |
| $\hat{V}_1(\hat{\theta})$ | 0.00076 | 0.00055 | 0.00012 | 0.00036 | 0.00072 | 0.00004 | 0.00032 |
| $\hat{V}_2(\hat{\theta})$ | 0.00076 | 0.00055 | 0.00012 | 0.00036 | 0.00072 | 0.00004 | 0.00032 |

**9.6** From Exercise 3.4, $\hat{B} = 11.41946$, $\hat{\bar{y}}_r = t_x\hat{B} = 10.3\hat{B} = 117.6$, and $\mathrm{SE}\,(\hat{\bar{y}}_r) = 3.98$. Using the jackknife, we have $\hat{B}_{(\cdot)} = 11.41937$, $\hat{\bar{y}}_{r(\cdot)} = 117.6$, and $\mathrm{SE}\,(\hat{\bar{y}}_r) = 10.3\sqrt{0.1836} = 4.41$. The jackknife standard error is larger, partly because it does not include the fpc.

**9.7** We use

$$\hat{V}_{\mathrm{JK}}(\hat{\bar{y}}_r) = \frac{n-1}{n}\sum_{j=1}^{10}(\hat{\bar{y}}_{r(j)} - \hat{\bar{y}}_r)^2.$$

The $\hat{\bar{y}}_{r(j)}$'s for *returnf* and *hadmeas* are given in the following table:

| School, $j$ | returnf, $\hat{\bar{y}}_{r(j)}$ | hadmeas, $\hat{\bar{y}}_{r(j)}$ |
|---|---|---|
| 1 | 0.5822185 | 0.4253903 |
| 2 | 0.5860165 | 0.4647582 |
| 3 | 0.5504290 | 0.4109223 |
| 4 | 0.5768984 | 0.4214941 |
| 5 | 0.5950112 | 0.4275614 |
| 6 | 0.5829014 | 0.4615285 |
| 7 | 0.5726580 | 0.4379689 |
| 8 | 0.5785320 | 0.4313120 |
| 9 | 0.5650470 | 0.4951728 |
| 10 | 0.5986785 | 0.4304341 |

For returnf,

$$\hat{V}_{JK}(\hat{\bar{y}}_r) = \frac{9}{10}\sum_{j=1}^{10}(\hat{\bar{y}}_{r(j)} - 0.5789482)^2 = 0.00160$$

For hadmeas,

$$\hat{V}_{JK}(\hat{\bar{y}}_r) = \frac{9}{10}\sum_{j=1}^{10}(\hat{\bar{y}}_{r(j)} - 0.4402907)^2 = 0.00526$$

**9.8** We have $\hat{B}_{(\cdot)} = .9865651$ and $\hat{V}_{JK}(\hat{B}) = 3.707 \times 10^{-5}$. With the fpc, the linearization variance estimate is $\hat{V}_L(\hat{B}) = 3.071 \times 10^{-5}$; the linearization variance estimate if we ignore the fpc is $3.071 \times 10^{-5}/\sqrt{1 - \frac{300}{3078}} = 3.232 \times 10^{-5}$.

**9.9** The median weekday greens fee for nine holes is $\hat{\theta} = 12$. For the SRS of size 120,

$$V[\hat{F}(\theta_{0.5})] = \frac{(.5)(.5)}{120} = 0.0021.$$

An approximate 95% confidence interval for the median is therefore

$$[\hat{F}^{-1}(.5 - 1.96\sqrt{.0021}), \hat{F}^{-1}(.5 + 1.96\sqrt{.0021})] = [\hat{F}^{-1}(.4105), \hat{F}^{-1}(.5895)].$$

We have the following values for the empirical distribution function:

| $y$ | 10.25 | 10.8 | 11 | 11.5 | 12 |
|---|---|---|---|---|---|
| $\hat{F}(y)$ | .3917 | .4000 | .4167 | .4333 | .5167 |

| $y$ | 13 | 14 | 15 | 16 |
|---|---|---|---|---|
| $\hat{F}(y)$ | .5250 | .5417 | .5833 | .6000 |

Interpolating,

$$\hat{F}^{-1}(.4105) = 10.8 + \frac{.4105 - .4}{.4167 - .4}(11 - 10.8) = 10.9$$

and

$$\hat{F}^{-1}(.5895) = 15 + \frac{.5895 - .5833}{.6 - .5833}(16 - 15) = 15.4.$$

Thus, an approximate 95% confidence interval for the median is [10.9, 15.4].

Note: If we apply the bootstrap to these data, we get

$$\frac{1}{1000}\sum_{r=1}^{1000} \hat{\theta}_r^* = 12.86$$

with standard error 1.39. This leads to a 95% CI of [10.1, 15.6] for the median.

**9.13** (a) Since $h''(t) = -2t$, the remainder term is

$$\int_a^x (x - t)h''(t)dt = -2\int_a^x (x - t)dt = -2\left(x^2 - \frac{x^2}{2} - ax + \frac{a^2}{2}\right) = -(x - a)^2.$$

Thus,

$$h(\hat{p}) = h(p) + h'(p)(\hat{p} - p) - (\hat{p} - p)^2 = p(1 - p) + (1 - 2p)(\hat{p} - p) - (\hat{p} - p)^2.$$

(b) The remainder term is likely to be smaller than the other terms because it has $(\hat{p} - p)^2$ in it. This will be small if $\hat{p}$ is close to $p$.

(c) To find the exact variance, we need to find $V(\hat{p}-\hat{p}^2)$, which involves the fourth moments. For an SRSWR, $X = n\hat{p} \sim \text{Bin}(n,p)$, so we can find the moments using the moment generating function of the Binomial:

$$M_X(t) = (pe^t + q)^n$$

So,

$$E(X) = M'_X(t)\Big|_{t=0} = n(pe^t+q)^{n-1}pe^t\Big|_{t=0} = np$$

$$\begin{aligned} E(X^2) &= M''_X(t)\Big|_{t=0} \\ &= [n(n-1)(pe^t+q)^{n-2}(pe^t)^2 + n(pe^t+q)^{n-1}pe^t]\Big|_{t=0} \\ &= n(n-1)p^2 + np \\ &= n^2p^2 + np(1-p) \end{aligned}$$

$$E(X^3) = M'''_X(t)\Big|_{t=0} = np(1-3p+3np+2p^2-3np^2+n^2p^2)$$

$$E(X^4) = np(1-7p+7np+12p^2-18np^2+6n^2p^2-6p^3+11np^3-6n^2p^3+n^3p^3)$$

Then,

$$\begin{aligned} &V[\hat{p}(1-\hat{p})] \\ &= V(\hat{p}) + V(\hat{p}^2) - 2\text{Cov}\,(\hat{p},\hat{p}^2) \\ &= E[\hat{p}^2] - p^2 + E[\hat{p}^4] - [E(\hat{p}^2)]^2 - 2E\left[\hat{p}^3\right] + 2pE(\hat{p}^2) \\ &= \frac{p(1-p)}{n} \\ &\quad + \frac{p}{n^3}(1-7p+7np+12p^2-18np^2+6n^2p^2-6p^3+11np^3-6n^2p^3+n^3p^3) \\ &\quad - \left[p^2 + \frac{p(1-p)}{n}\right]^2 \\ &\quad -2\frac{p}{n^2}(1-3p+3np+2p^2-3np^2+n^2p^2) + 2p\left[p^2+\frac{p(1-p)}{n}\right] \\ &= \frac{p(1-p)}{n}(1-4p+4p^2) \\ &\quad + \frac{1}{n^2}(-2p+14p^2-22p^3+12p^4) + \frac{1}{n^3}(p-7p^2+12p^3-6p^4) \end{aligned}$$

Note that the first term is $(1-2p)^2V(\hat{p})/n$, and the other terms are (constant)$/n^2$ and (constant)$/n^3$. The remainder terms become small relative to the first term when $n$ is large. You can see why statisticians use the linearization method so frequently: even for this simple example, the exact calculations of the variance are nasty.

Note that with an SRS without replacement, the result is much more complicated. Results from the following paper may be used to find the moments.

Finucan, H. M., Galbraith, R. F., and Stone, M. (1974). Moments Without Tears in Simple Random Sampling from a Finite Population *Biometrika*, 61, 151–154.

**9.14** (a) Write $B_1 = h(t_{xy}, t_x, t_y, t_{x^2}, N)$, where

$$h(a, b, c, d, e) = \frac{a - bc/e}{d - b^2/e} = \frac{ea - bc}{ed - b^2}.$$

The partial derivatives, evaluated at the population quantities, are:

$$\begin{aligned}
\frac{\partial h}{\partial a} &= \frac{e}{ed - b^2} \\
\frac{\partial h}{\partial b} &= -\frac{c}{ed - b^2} + 2b\frac{ea - bc}{(ed - b^2)^2} \\
&= -\frac{c}{ed - b^2} + \frac{2bB_1}{ed - b^2} \\
&= -\frac{e}{ed - b^2}\left[\frac{c}{e} - \frac{b}{e}B_1 - \frac{b}{e}B_1\right] \\
&= -\frac{e}{ed - b^2}(B_0 - B_1\bar{x}_U) \\
\frac{\partial h}{\partial c} &= -\frac{b}{ed - b^2} = -\frac{e}{ed - b^2}\bar{x}_U \\
\frac{\partial h}{\partial d} &= -\frac{e}{ed - b^2}B_1 \\
\frac{\partial h}{\partial e} &= \frac{a}{ed - b^2} - \frac{d(ea - bc)}{(ed - b^2)^2} \\
&= \frac{a}{ed - b^2} - \frac{dB_1}{ed - b^2} \\
&= \frac{e}{ed - b^2}B_0\bar{x}_U.
\end{aligned}$$

The last equality follows from the normal equations. Then, by linearization,

$$\begin{aligned}
\hat{B}_1 - B_1 & \\
&\approx \frac{\partial h}{\partial a}(\hat{t}_{xy} - t_{xy}) + \frac{\partial h}{\partial b}(\hat{t}_x - t_x) + \frac{\partial h}{\partial c}(\hat{t}_y - t_y) \\
&\quad + \frac{\partial h}{\partial d}(\hat{t}_{x^2} - t_{x^2}) + \frac{\partial h}{\partial e}(\hat{N} - N) \\
&= \frac{N}{N t_{x^2} - (t_x)^2}\left[\hat{t}_{xy} - t_{xy} - (B_0 - B_1\bar{x}_U)(\hat{t}_x - t_x)\right. \\
&\quad \left. - \bar{x}_U(\hat{t}_y - t_y) - B_1(\hat{t}_{x^2} - t_{x^2}) + B_0\bar{x}_U(\hat{N} - N)\right] \\
&= \frac{N}{N t_{x^2} - (t_x)^2}\left[\sum_{i\in\mathcal{S}} w_i \left\{x_i y_i - (B_0 - B_1\bar{x}_U)x_i - \bar{x}_U y_i - B_1 x_i^2 + B_0\bar{x}_U\right\}\right] \\
&\quad - \frac{N}{N t_{x^2} - (t_x)^2}\left[t_{xy} - t_x(B_0 - B_1\bar{x}_U) - \bar{x}_U t_y - B_1 t_{x^2} + B_0 N\bar{x}_U\right] \\
&= \frac{N}{N t_{x^2} - (t_x)^2}\sum_{i\in\mathcal{S}} w_i\,(y_i - B_0 - B_1 x_i)\,(x_i - \bar{x}_U).
\end{aligned}$$

**9.15** (a) Write

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \frac{1}{N}\sum_{j=1}^{N} y_j\right)^2 = \frac{1}{t_3 - 1}\left(t_1 - \frac{t_2}{t_3}\right)$$

(b) Substituting, we have

$$\hat{S}^2 = \frac{1}{\hat{t}_3 - 1}\left(\hat{t}_1 - \frac{\hat{t}_2^2}{\hat{t}_3}\right)$$

(c) We need to find the partial derivatives:

$$\begin{aligned}
\frac{\partial h}{\partial t_1} &= \frac{1}{t_3 - 1} \\
\frac{\partial h}{\partial t_2} &= -2\frac{t_2}{t_3(t_3 - 1)} \\
\frac{\partial h}{\partial t_3} &= -\frac{1}{(t_3 - 1)^2}\left(t_1 - \frac{t_2}{t_3}\right) + \frac{1}{t_3 - 1}\frac{t_2}{t_3^2}
\end{aligned}$$

Then, by linearization,

$$\hat{S}^2 - S^2 \approx \frac{\partial h}{\partial t_1}(\hat{t}_1 - t_1) + \frac{\partial h}{\partial t_2}(\hat{t}_2 - t_2) + \frac{\partial h}{\partial t_3}(\hat{t}_3 - t_3)$$

Let

$$\begin{aligned} q_i &= \frac{\partial h}{\partial t_1} y_i^2 + \frac{\partial h}{\partial t_2} y_i + \frac{\partial h}{\partial t_3} \\ &= \frac{1}{t_3 - 1} y_i^2 - 2 \frac{t_2}{t_3(t_3-1)} y_i - \frac{1}{(t_3-1)^2} \left( t_1 - \frac{t_2}{t_3} \right) + \frac{1}{t_3 - 1} \frac{t_2}{t_3^2} \\ &= \frac{1}{t_3 - 1} \left( y_i^2 - 2 \frac{t_2}{t_3} y_i - \frac{1}{(t_3-1)} \left( t_1 - \frac{t_2}{t_3} \right) + \frac{t_2}{t_3^2} \right) \end{aligned}$$

**9.16** (a) Write $R = h(t_1, \ldots, t_6)$, where

$$h(a, b, c, d, e, f) = \frac{d - ab/f}{\sqrt{(c - a^2/f)(e - b^2/f)}} = \frac{fd - ab}{\sqrt{(fc - a^2)(fe - b^2)}}.$$

The partial derivatives, evaluated at the population quantities, are:

$$\begin{aligned} \frac{\partial h}{\partial a} &= \frac{1}{\sqrt{(fc - a^2)(fe - b^2)}} \left( -b + \frac{a(fd - ab)}{fc - a^2} \right) \\ &= \frac{-t_y}{N(N-1)S_x S_y} + \frac{t_x R}{N(N-1)S_x^2} \\ \frac{\partial h}{\partial b} &= \frac{1}{\sqrt{(fc - a^2)(fe - b^2)}} \left( -a + \frac{b(fd - ab)}{fe - b^2} \right) \\ &= \frac{-t_y}{N(N-1)S_x S_y} + \frac{t_y R}{N(N-1)S_y^2} \\ \frac{\partial h}{\partial c} &= -\frac{1}{2\sqrt{(fc - a^2)(fe - b^2)}} \left( \frac{f(fd - ab)}{fc - a^2} \right) \\ &= -\frac{R}{2} \frac{1}{(N-1)S_x^2} \\ \frac{\partial h}{\partial d} &= \frac{f}{\sqrt{(fc - a^2)(fe - b^2)}} \\ \frac{\partial h}{\partial e} &= -\frac{1}{2\sqrt{(fc - a^2)(fe - b^2)}} \left( \frac{f(fd - ab)}{fe - b^2} \right) \\ &= -\frac{R}{2} \frac{1}{(N-1)S_y^2} \\ \frac{\partial h}{\partial f} &= \frac{d}{\sqrt{(fc - a^2)(fe - b^2)}} - \frac{fd - ab}{2\sqrt{(fc - a^2)(fe - b^2)}} \left( \frac{e}{fe - b^2} + \frac{c}{fc - a^2} \right) \\ &= \frac{t_{xy}}{N(N-1)S_x S_y} - \frac{R}{2} \left( \frac{t_{y^2}}{N(N-1)S_y^2} + \frac{t_{x^2}}{N(N-1)S_x^2} \right) \end{aligned}$$

Then, by linearization,

$$\begin{aligned}
\hat{R} - R \\
&\approx \frac{\partial h}{\partial a}(\hat{t}_x - t_x) + \frac{\partial h}{\partial b}(\hat{t}_y - t_y) + \frac{\partial h}{\partial c}(\hat{t}_{x^2} - t_{x^2}) \\
&\quad + \frac{\partial h}{\partial d}(\hat{t}_{xy} - t_{xy}) + \frac{\partial h}{\partial e}(\hat{t}_{y^2} - t_{y^2}) + \frac{\partial h}{\partial f}(\hat{N} - N) \\
&= \frac{1}{N(N-1)S_xS_y}\left[\left(-t_y + \frac{t_xRS_y}{S_x}\right)(\hat{t}_x - t_x) + (-t_x + \frac{t_yRS_x}{S_y})(\hat{t}_y - t_y)\right] \\
&\quad - \frac{R}{2}\frac{1}{(N-1)S_x^2}(\hat{t}_{x^2} - t_{x^2}) + \frac{1}{(N-1)S_xS_y}(\hat{t}_{xy} - t_{xy}) \\
&\quad - \frac{R}{2}\frac{1}{(N-1)S_y^2}(\hat{t}_{y^2} - t_{y^2}) \\
&\quad + \left[\frac{t_{xy}}{N(N-1)S_xS_y} - \frac{R}{2}\left(\frac{t_{y^2}}{N(N-1)S_y^2} + \frac{t_{x^2}}{N(N-1)S_x^2}\right)\right](\hat{N} - N) \\
&= \frac{1}{N(N-1)S_xS_y}\left[\left(-t_y + \frac{t_xRS_y}{S_x}\right)(\hat{t}_x - t_x) + \left(-t_x + \frac{t_yRS_x}{S_y}\right)(\hat{t}_y - t_y)\right. \\
&\quad - \frac{NRS_y}{2S_x}(\hat{t}_{x^2} - t_{x^2}) + N(\hat{t}_{xy} - t_{xy}) - \frac{NRS_x}{2S_y}(\hat{t}_{y^2} - t_{y^2}) \\
&\quad \left. + \left\{t_{xy} - \frac{R}{2}\left(\frac{t_{y^2}S_x}{S_y} + \frac{t_{x^2}S_y}{S_x}\right)\right\}(\hat{N} - N)\right]
\end{aligned}$$

This is somewhat easier to do in matrix terms. Let

$$\begin{aligned}
\boldsymbol{\delta} &= \left[\left(-\bar{y}_U + \frac{\bar{x}_URS_y}{S_x}\right), \left(-\bar{x}_U + \frac{\bar{y}_URS_x}{S_y}\right), -\frac{RS_y}{2S_x}, -\frac{RS_x}{2S_y}, 1,\right. \\
&\quad \left.\frac{t_{xy}}{N} - \frac{R}{2N}\left(\frac{t_{y^2}S_x}{S_y} + \frac{t_{x^2}S_y}{S_x}\right)\right]^T,
\end{aligned}$$

then

$$\text{Cov}\,(\hat{R}) \approx \frac{1}{[(N-1)S_xS_y]^2}\boldsymbol{\delta}^T\text{Cov}\,(\hat{\mathbf{t}})\boldsymbol{\delta}.$$

**9.17** Write the function as $h(a_1, \ldots, a_L, b_1, \ldots, b_L)$. Then

$$\left.\frac{\partial h}{\partial a_l}\right|_{t_1,\ldots,t_L,N_1,\ldots,N_L} = 1$$

and

$$\left.\frac{\partial h}{\partial b_l}\right|_{t_1,\ldots,t_L,N_1,\ldots,N_L} = -\frac{t_l}{N_l}.$$

Consequently,

$$h(\hat{t}_1, \ldots, \hat{t}_L, \hat{N}_1, \ldots, \hat{N}_L) \approx t + \sum_{l=1}^{L}(\hat{t}_l - t_l) - \sum_{l=1}^{L}\frac{t_l}{N_l}(\hat{N}_l - N_l)$$

and

$$V(\hat{t}_{\text{post}}) \approx V\left[\sum_{l=1}^{L}\left(\hat{t}_l - \frac{t_l}{N_l}\hat{N}_l\right)\right].$$

**9.18** From (9.5),

$$\hat{V}_2(\hat{\theta}) = \frac{1}{R}\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}_r - \hat{\theta})^2.$$

Without loss of generality, let $\bar{y}_U = 0$. We know that $\bar{y} = \sum_{r=1}^{R}\bar{y}_r/R$.

Suppose the random groups are independent. Then $\bar{y}_1, \ldots, \bar{y}_R$ are independent and identically distributed random variables with

$$E[\bar{y}_r] = 0,$$

$$V[\bar{y}_r] = E[\bar{y}_r^2] = \frac{S^2}{m} = \kappa_2(\bar{y}_1),$$

$$E[\bar{y}_r^4] = \kappa_4(\bar{y}_1).$$

We have

$$\begin{aligned}
E\left[\frac{1}{R(R-1)}\sum_{r=1}^{R}(\bar{y}_r - \bar{y})^2\right] &= \frac{1}{R(R-1)}\sum_{r=1}^{R}E\left[\bar{y}_r^2 - (\bar{y})^2\right] \\
&= \frac{1}{R(R-1)}\sum_{r=1}^{R}[V(\bar{y}_r) - V(\bar{y})] \\
&= \frac{1}{R(R-1)}\sum_{r=1}^{R}\left[\frac{S^2}{m} - \frac{S^2}{n}\right] \\
&= \frac{1}{R(R-1)}\sum_{r=1}^{R}\left[R\frac{S^2}{n} - \frac{S^2}{n}\right] \\
&= \frac{S^2}{n}.
\end{aligned}$$

Also,

$$\begin{aligned}
& E\left[\left\{\sum_{r=1}^{R}(\bar{y}_r-\bar{y})^2\right\}^2\right] \\
&= E\left[\left\{\sum_{r=1}^{R}\bar{y}_r^2-R\bar{y}^2\right\}^2\right] \\
&= E\left[\sum_{r=1}^{R}\sum_{s=1}^{R}\bar{y}_r^2\bar{y}_s^2-2R\bar{y}^2\sum_{r=1}^{R}\bar{y}_r^2+R^2\bar{y}^4\right] \\
&= E\left[\sum_{r=1}^{R}\sum_{s=1}^{R}\left(\bar{y}_r^2\bar{y}_s^2-\frac{2}{R}\bar{y}_r^2\bar{y}_s^2\right)+\frac{1}{R^2}\sum_{j}\sum_{k}\sum_{r}\sum_{s}\bar{y}_j\bar{y}_k\bar{y}_r\bar{y}_s\right] \\
&= E\left[\left(1-\frac{2}{R}+\frac{1}{R^2}\right)\sum_{r=1}^{R}\bar{y}_r^4+\left(1-\frac{2}{R}+\frac{3}{R^2}\right)\sum_{r=1}^{R}\sum_{s\neq r}^{R}\bar{y}_r^2\bar{y}_s^2\right] \\
&= \left(1-\frac{2}{R}+\frac{1}{R^2}\right)R\kappa_4(\bar{y}_1)+\left(1-\frac{2}{R}+\frac{3}{R^2}\right)R(R-1)\kappa_2^2(\bar{y}_1)
\end{aligned}$$

Consequently,

$$\begin{aligned}
& E\left[\hat{V}_2^2(\hat{\theta})\right] \\
&= \frac{1}{R^2(R-1)^2}\left[\left(1-\frac{2}{R}+\frac{1}{R^2}\right)R\kappa_4(\bar{y}_1)+\left(1-\frac{2}{R}+\frac{3}{R^2}\right)R(R-1)\kappa_2^2(\bar{y}_1)\right] \\
&= \frac{1}{R^3}\kappa_4(\bar{y}_1)+\frac{1}{R^3(R-1)}(R^2-2R+3)\kappa_2^2(\bar{y}_1)
\end{aligned}$$

and

$$\begin{aligned}
V\left[\frac{1}{R(R-1)}\sum_{r=1}^{R}(\bar{y}_r-\bar{y})^2\right] &= \frac{1}{R^3}\kappa_4(\bar{y}_1)+\frac{R^2-2R+3}{R^3(R-1)}\kappa_2^2(\bar{y}_1)-\left(\frac{S^2}{n}\right)^2 \\
&= \frac{1}{R^3}\kappa_4(\bar{y}_1)+\frac{R^2-2R+3}{R^3(R-1)}\left(\frac{S^2}{m}\right)^2-\left(\frac{S^2}{Rm}\right)^2
\end{aligned}$$

$$\begin{aligned}
\mathrm{CV}^2\left[\frac{1}{R(R-1)}\sum_{r=1}^{R}(\bar{y}_r-\bar{y})^2\right] &= \frac{\frac{1}{R^3}\kappa_4(\bar{y}_1)+\frac{R^2-2R+3}{R^3(R-1)}\left(\frac{S^2}{m}\right)^2-\left(\frac{S^2}{Rm}\right)^2}{\left(\frac{S^2}{Rm}\right)^2} \\
&= \frac{1}{R}\left[\frac{\kappa_4(\bar{y}_1)m^2}{S^4}-\frac{R-3}{R-1}\right].
\end{aligned}$$

We now need to find $\kappa_4(\bar{y}_1) = E[\bar{y}_r^4]$ to finish the problem. A complete argument giving the fourth moment for an SRSWR is given by

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). *Sample Survey Methods and Theory, Volume 2.* New York: Wiley, pp. 99-100.

They note that

$$\begin{aligned}\bar{y}_r^4 &= \frac{1}{m^4}\left[\sum_{i\in\mathcal{S}_r} y_i^4 + 4\sum_{i\neq j} y_i^3 y_j + 3\sum_{i\neq j} y_i^2 y_j^2 \right. \\ &\qquad \left. +6\sum_{i\neq j\neq k} y_i^2 y_j y_k + \sum_{i\neq j\neq k\neq l} y_i y_j y_k y_l\right]\end{aligned}$$

so that

$$\kappa_4(\bar{y}_1) = E[\bar{y}_r^4] = \frac{1}{m^3(N-1)}\sum_{i=1}^{N}(y_i - \bar{y}_U)^4 + 3\frac{m-1}{m^3}S^4.$$

This results in

$$\begin{aligned}\mathrm{CV}^2\left[\frac{1}{R(R-1)}\sum_{r=1}^{R}(\bar{y}_r - \bar{y})^2\right] &= \frac{1}{R}\left[\frac{\kappa_4(\bar{y}_1)m^2}{S^4} - \frac{R-3}{R-1}\right] \\ &= \frac{1}{R}\left[\frac{\kappa}{m} + 3\frac{m-1}{m^3} - \frac{R-3}{R-1}\right].\end{aligned}$$

The number of groups, $R$, has more impact on the CV than the group size $m$ : the random group estimator of the variance is unstable if $R$ is small.

**9.19** First note that

$$\begin{aligned}\bar{y}_{\mathrm{str}}(\alpha_r) - \bar{y}_{\mathrm{str}} &= \sum_{h=1}^{H}\frac{N_h}{N}y_h(\alpha_r) - \sum_{h=1}^{H}\frac{N_h}{N}\frac{y_{h1}+y_{h2}}{2} \\ &= \sum_{h=1}^{H}\frac{N_h}{N}\left(\frac{\alpha_{rh}+1}{2}y_{h1} - \frac{\alpha_{rh}-1}{2}y_{h2}\right) - \sum_{h=1}^{H}\frac{N_h}{N}\frac{y_{h1}+y_{h2}}{2} \\ &= \sum_{h=1}^{H}\frac{N_h}{N}\alpha_{rh}\frac{y_{h1}-y_{h2}}{2}.\end{aligned}$$

Then

$$\begin{aligned}
\hat{V}_{\text{BRR}}(\bar{y}_{\text{str}}) &= \frac{1}{R}\sum_{r=1}^{R}[\bar{y}_{\text{str}}(\alpha_r) - \bar{y}_{\text{str}}]^2 \\
&= \frac{1}{R}\sum_{r=1}^{R}\left[\sum_{h=1}^{H}\frac{N_h}{N}\alpha_{rh}\frac{y_{h1}-y_{h2}}{2}\right]^2 \\
&= \frac{1}{R}\sum_{r=1}^{R}\sum_{h=1}^{H}\sum_{\ell=1}^{H}\frac{N_h}{N}\alpha_{rh}\frac{y_{h1}-y_{h2}}{2}\frac{N_\ell}{N}\alpha_{r\ell}\frac{y_{\ell 1}-y_{\ell 2}}{2} \\
&= \frac{1}{R}\sum_{r=1}^{R}\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\alpha_{rh}^2\frac{(y_{h1}-y_{h2})^2}{4} \\
&\quad + \frac{1}{R}\sum_{h=1}^{H}\sum_{\ell=1,\ell\neq h}^{H}\frac{N_h}{N}\frac{y_{h1}-y_{h2}}{2}\frac{N_\ell}{N}\frac{y_{\ell_1}-y_{\ell_2}}{2}\sum_{r=1}^{R}\alpha_{rh}\alpha_{r\ell} \\
&= \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\frac{(y_{h1}-y_{h2})^2}{4} \\
&= \hat{V}_{\text{str}}(\bar{y}_{\text{str}}).
\end{aligned}$$

The last step holds because $\sum_{r=1}^{R}\alpha_{rh}\alpha_{r\ell} = 0$ for $\ell \neq h$.

**9.20** As noted in the text,

$$\hat{V}_{\text{str}}(\bar{y}_{\text{str}}) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\frac{(y_{h1}-y_{h2})^2}{4}.$$

Also,

$$\hat{\theta}(\alpha_r) = \bar{y}_{\text{str}}(\alpha_r) = \sum_{h=1}^{H}\frac{\alpha_{rh}}{2}\frac{N_h}{N}(y_{h1}-y_{h2}) + \bar{y}_{\text{str}}$$

so

$$\hat{\theta}(\alpha_r) - \hat{\theta}(-\alpha_r) = \sum_{h=1}^{H}\alpha_{rh}\frac{N_h}{N}(y_{h1}-y_{h2})$$

and, using the property $\sum_{r=1}^{R}\alpha_{rh}\alpha_{rk} = 0$ for $k \neq h$,

$$\begin{aligned}
\frac{1}{4R}\sum_{r=1}^{R}[\hat{\theta}(\alpha_r) - \hat{\theta}(-\alpha_r)]^2 &= \frac{1}{4R}\sum_{r=1}^{R}\sum_{h=1}^{H}\sum_{k=1}^{H}\alpha_{rh}\alpha_{rk}\frac{N_h}{N}\frac{N_k}{N}(y_{h1}-y_{h2})(y_{k1}-y_{k2}) \\
&= \frac{1}{4R}\sum_{r=1}^{R}\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2(y_{h1}-y_{h2})^2 \\
&= \frac{1}{4}\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2(y_{h1}-y_{h2})^2 = \hat{V}_{\text{str}}(\bar{y}_{\text{str}}).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\frac{1}{2R}\sum_{r=1}^{R}\{[\hat{\theta}(\alpha_r)-\hat{\theta}]^2+[\hat{\theta}(-\alpha_r)-\hat{\theta}]^2\} \\
&= \frac{1}{2R}\sum_{r=1}^{R}\left\{\left[\sum_{h=1}^{H}\frac{\alpha_{rh}}{2}\frac{N_h}{N}(y_{h1}-y_{h2})\right]^2+\left[\sum_{h=1}^{H}\frac{-\alpha_{rh}}{2}\frac{N_h}{N}(y_{h1}-y_{h2})\right]^2\right\} \\
&= \frac{1}{2R}\sum_{r=1}^{R}\sum_{h=1}^{H}\frac{\alpha_{rh}^2}{2}\left(\frac{N_h}{N}\right)^2(y_{h1}-y_{h2})^2 \\
&= \frac{1}{4}\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2(y_{h1}-y_{h2})^2.
\end{aligned}$$

**9.21** Note that

$$\begin{aligned}
\hat{t}(\alpha_r) &= \sum_{h=1}^{H}N_h\left[\frac{\alpha_{rh}}{2}(y_{h1}-y_{h2})+\bar{y}_h\right] \\
&= \sum_{h=1}^{H}\frac{N_h a_{rh}}{2}(y_{h1}-y_{h2})+\hat{t}
\end{aligned}$$

and

$$\begin{aligned}
[\hat{t}(\alpha_r)]^2 &= \sum_{h=1}^{H}\sum_{k=1}^{H}\frac{N_h N_k \alpha_{rh}\alpha_{rk}}{4}(y_{h1}-y_{h2})(y_{k1}-y_{k2}) \\
&\quad +2\hat{t}\sum_{h=1}^{H}\frac{N_h\alpha_{rh}}{2}(y_{h1}-y_{h2})+\hat{t}^2.
\end{aligned}$$

Thus,

$$\hat{t}(\alpha_r)-\hat{t}(-\alpha_r)=\sum_{h=1}^{H}N_h\alpha_{rh}(y_{h1}-y_{h2}),$$

$$[\hat{t}(\alpha_r)]^2-[\hat{t}(-\alpha_r)]^2=2\hat{t}\sum_{h=1}^{H}N_h\alpha_{rh}(y_{h1}-y_{h2}),$$

and

$$\hat{\theta}(\alpha_r)-\hat{\theta}(-\alpha_r)=(2a\hat{t}+b)\sum_{h=1}^{H}N_h\alpha_{rh}(y_{h1}-y_{h2}).$$

Consequently, using the balanced property $\sum_{r=1}^{R} \alpha_{rh}\alpha_{rk} = 0$ for $k \neq h$, we have

$$\begin{aligned}
&\frac{1}{4R}\sum_{r=1}^{R}[\hat{\theta}(\alpha_r) - \hat{\theta}(-\alpha_r)]^2 \\
&= \frac{1}{4R}(2a\hat{t}+b)^2 \sum_{r=1}^{R}\sum_{h=1}^{H}\sum_{k=1}^{H} N_h N_k \alpha_{rh}\alpha_{rk}(y_{h1}-y_{h2})(y_{k1}-y_{k2}) \\
&= \frac{1}{4}(2a\hat{t}+b)^2 \sum_{h=1}^{H} N_h^2 (y_{h1}-y_{h2})^2.
\end{aligned}$$

Using linearization,

$$h(\hat{t}) \approx h(t) + (2at+b)(\hat{t}-t),$$

so

$$V_L(\hat{\theta}) = (2at-b)^2 V(\hat{t})$$

and

$$\hat{V}_L(\hat{\theta}) = (2a\hat{t}-b)^2 \frac{1}{4}\sum_{h=1}^{H} N_h^2 (y_{h1}-y_{h2})^2,$$

which is the same as $\frac{1}{4R}\sum_{r=1}^{R}[\hat{\theta}(\alpha_r) - \hat{\theta}(-\alpha_r)]^2$.

**9.23** We can write

$$\hat{t}_{\text{post}} = g(\mathbf{w}, \mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_L) = \sum_{l=1}^{L} \frac{N_l \sum_{j\in\mathcal{S}} w_j x_{lj} y_j}{\sum_{j\in\mathcal{S}} w_j x_{lj}}.$$

Then,

$$\begin{aligned}
z_i &= \frac{\partial g(\mathbf{w}, \mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_L)}{\partial w_i} \\
&= \sum_{l=1}^{L}\left\{ \frac{N_l x_{li} y_i}{\sum_{j\in\mathcal{S}} w_j x_{lj}} - \frac{N_l x_{li}\sum_{j\in\mathcal{S}} w_j x_{lj} y_j}{\left(\sum_{j\in\mathcal{S}} w_j x_{lj}\right)^2} \right\} \\
&= \sum_{l=1}^{L}\left\{ \frac{N_l x_{li} y_i}{\hat{N}_l} - \frac{N_l x_{li}\hat{t}_{yl}}{\hat{N}_l^2} \right\} \\
&= \sum_{l=1}^{L} x_{li}\frac{N_l}{\hat{N}_l}\left(y_i - \frac{\hat{t}_{yl}}{\hat{N}_l}\right).
\end{aligned}$$

Thus,

$$\hat{V}(\hat{t}_{\text{post}}) \;=\; \hat{V}\left(\sum_{i\in\mathcal{S}} w_i z_i\right).$$

Note that this variance estimator differs from the one in Exercise 9.17, although they are asymptotically equivalent.

**9.24** From Chapter 5,

$$\begin{aligned} V(\hat{t}) &\approx N^2\frac{M\text{MSB}}{n} \\ &= \frac{N^2M}{n}\frac{NM-1}{M(N-1)}S^2[1+(M-1)ICC] \\ &\approx \frac{NM}{n}\frac{NM}{M}p(1-p)[1+(M-1)ICC] \end{aligned}$$

Consequently, the relative variance $v$ can be written as $\beta_0 + \beta_1/t$, where $\beta_0 = -\frac{1}{nM}[1+(M-1)ICC] + \frac{1}{nt}N[1+(M-1)ICC]$.

**9.25** (a) From (9.2),

$$\begin{aligned} V[\hat{B}] &\approx E\left[\left\{-\frac{t_y}{t_x^2}(\hat{t}_x - t_x) + \frac{1}{t_x}(\hat{t}_y - t_y)\right\}^2\right] \\ &= \frac{t_y^2}{t_x^2}E\left[\left\{-\frac{1}{t_x}(\hat{t}_x - t_x) + \frac{1}{t_y}(\hat{t}_y - t_y)\right\}^2\right] \\ &= \frac{t_y^2}{t_x^2}\left[\frac{V(\hat{t}_x)}{t_x^2} + \frac{V(\hat{t}_y}{t_y^2} - 2\text{Cov}\left(\frac{\hat{t}_x}{t_x}, \frac{\hat{t}_y}{t_y}\right)\right] \\ &= \frac{t_y^2}{t_x^2}\left[\frac{V(\hat{t}_x)}{t_x^2} + \frac{V(\hat{t}_y}{t_y^2} - 2\frac{B}{t_x t_y}V\left(\hat{t}_x\right)\right] \\ &= \frac{t_y^2}{t_x^2}\left[\frac{V(\hat{t}_y}{t_y^2} - \frac{V(\hat{t}_x)}{t_x^2}\right] \end{aligned}$$

Using the fitted model from (9.13),

$$\frac{\hat{V}(\hat{t}_x)}{\hat{t}_x^2} = a + b/\hat{t}_x$$

and

$$\frac{\hat{V}(\hat{t}_y)}{\hat{t}_y^2} = a + b/\hat{t}_y$$

Consequently, substituting estimators for the population quantitites,

$$\hat{V}[\hat{B}] \;=\; \hat{B}^2\left[a + \frac{b}{\hat{t}_y} - a - \frac{b}{\hat{t}_x}\right],$$

which gives the result.

(b) When $B$ is a proportion,

$$\hat{V}[\hat{B}] = \hat{B}^2 \left[ \frac{b}{\hat{t}_y} - \frac{b}{\hat{t}_x} \right] = \hat{B}^2 \left[ \frac{b}{\hat{t}_x \hat{B}} - \frac{b}{\hat{t}_x} \right] = \frac{b\hat{B}(1-\hat{B})}{\hat{t}_x}.$$

# Chapter 10

# Categorical Data Analysis in Complex Surveys

**10.1** Many data sets used for chi-square tests in introductory statistics books use dependent data. See Alf and Lohr (2007) for a review of how books ignore clustering in the data.

**10.3** (a) Observed and expected (in parentheses) proportions are given in the following table:

| | | Abuse | |
|---|---|---|---|
| | | No | Yes |
| Symptom | No | .7542 (.7109) | .1017 (.1451) |
| | Yes | .0763 (.1196) | .0678 (.0244) |

(b)

$$\begin{aligned} X^2 &= 118\left[\frac{(.7542-.7109)^2}{.7109}+\cdots+\frac{(.0678-.0244)^2}{.0244}\right] \\ &= 12.8 \\ G^2 &= 2(118)\left[.7542\ln\left(\frac{.7542}{.7109}\right)+\cdots+.0678\ln\left(\frac{.0678}{.0244}\right)\right] \\ &= 10.3. \end{aligned}$$

Both $p$-values are less than .002.

Because the expected count in the Yes-Yes cell is small, we also perform Fisher's exact test, which gives $p$-value .0016.

**10.4** (a) This is a test of independence. A sample of students is taken, and each student classified based on instructors and grade.

(b) $X^2 = 34.8$. Comparing this to a $\chi^2_3$ distribution, we see that the p-value is

less than 0.0001. A similar conclusion follows from the likelihood ratio test, with $G^2 = 34.5$.

(c) Students are probably not independent–most likely, a cluster sample of students was taken, with the Math II classes as the clusters. The $p$-values in part (b) are thus lower than they should be.

**10.5** The following table gives the value of $\hat{\theta}$ for the 7 random groups:

| Random Group | $\hat{\theta}$ |
|---|---|
| 1 | 0.0132 |
| 2 | 0.0147 |
| 3 | 0.0252 |
| 4 | -0.0224 |
| 5 | 0.0073 |
| 6 | -0.0057 |
| 7 | 0.0135 |
| Average | 0.0065 |
| std. dev. | 0.0158 |

Using the random group method, the standard error of $\hat{\theta}$ is $0.0158/\sqrt{7} = 0{:}0060$, so the test statistic is

$$\frac{\hat{\theta}^2}{V(\hat{\theta})} = 0.79.$$

Since our estimate of the variance from the random group method has only 6 df, we compare the test statistic to an $F(1, 6)$ distribution rather than to a $\chi^2_1$ distribution, obtaining a $p$-value of 0.4.

**10.6** (a) The contingency table (for complete data) is as follows:

| | Break again? | | |
|---|---|---|---|
| | No | Yes | |
| Faculty | 65 | 167 | 232 |
| Classified staff | 55 | 459 | 514 |
| Administrative staff | 11 | 75 | 86 |
| Academic professional | 9 | 58 | 67 |
| | 140 | 759 | 899 |

$X^2_p = 37.3$; comparing to a $\chi^2_3$ distribution gives $\rho$-value $< .0001$. We can use the $\chi^2$ test for homogeneity because we assume product-multinomial sampling. (*Class* is the stratification variable.)

(b) Using the weights (with the respondents who answer both questions), we estimate the probabilities as

| | | Work | | |
|---|---|---|---|---|
| | | No | Yes | |
| Breakaga | No | 0.0832 | 0.0859 | 0.1691 |
| | Yes | 0.6496 | 0.1813 | 0.8309 |
| | | 0.7328 | 0.2672 | 1.0000 |

To estimate the proportion in the Yes-Yes cell, I used:

$$\hat{p}_{yy} = \frac{\text{sum of weights of persons answering yes to both questions}}{\text{sum of weights of respondents to both questions}}.$$

Other answers are possible, depending on how you want to treat the nonresponse.

(c) The odds ratio, calculated using the table in part (b), is

$$\frac{0.0832/0.0859}{0.6496/0.1813} = 0.27065.$$

(Or, you could get $1/.27065 = 3.695$.)

The estimated proportions ignoring the weights are

| | | Work | | |
|---|---|---|---|---|
| | | No | Yes | |
| breakaga | No | 0.0850 | 0.0671 | 0.1521 |
| | Yes | 0.6969 | 0.1510 | 0.8479 |
| | | 0.7819 | 0.2181 | 1.0000 |

Without weights the odds ratio is

$$\frac{0.0850/0.0671}{0.6969/0.1510} = 0.27448$$

(or, $1/.27448 = 3.643$).

Weights appear to make little difference in the value of the odds ratio.

(d) $\hat{\theta} = (.0832)(.1813) - (.6496)(.0859) = -0.04068$.

(e) Using linearization, define

$$q_i = \hat{p}_{22}y_{11i} + \hat{p}_{11}y_{12i} - \hat{p}_{12}y_{21i} - \hat{p}_{21}y_{22i}$$

where $y_{jki}$ is an indicator variable for membership in class $(j, k)$. We then estimate $V(\bar{q}_{\text{str}})$ using the usual methods for stratified samples. Using the summary statistics,

| Stratum | $N_h$ | $n_h$ | $\bar{q}_h$ | $s_h^2$ | $\frac{N_h}{N}\left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}$ |
|---|---|---|---|---|---|
| Faculty | 1374 | 228 | $-.117$ | 0.0792 | $4.04 \times 10^{-5}$ |
| C.S. | 1960 | 514 | $-.059$ | 0.0111 | $4.52 \times 10^{-6}$ |
| A.S. | 252 | 86 | $-.061$ | 0.0207 | $7.42 \times 10^{-7}$ |
| A.P. | 95 | 66 | $-.076$ | 0.0349 | $1.08 \times 10^{-7}$ |
| Total | 3681 | 894 | | | $4.58 \times 10^{-5}$ |

Thus $\hat{V}_L(\hat{\theta}) = 4.58 \times 10^{-5}$ and

$$X^2_W = \frac{\hat{\theta}^2}{\hat{V}_L(\hat{\theta})} = \frac{0.00165}{4.58 \times 10^{-5}} = 36.2.$$

We reject the null hypothesis with $p$-value $< 0.0001$.

**10.7** Answers will vary, depending on how the categories for zprep are formed.

**10.8** (a) Under the null hypothesis of independence the expected proportions are:

| | | Fitness Level | | |
|---|---|---|---|---|
| | | Recommended | Minimum Acceptable | Unacceptable |
| Smoking Status | Current | .241 | .140 | .159 |
| | Occasional | .020 | .011 | .013 |
| | Never | .186 | .108 | .123 |

Using (10.2) and (10.3),

$$\begin{aligned} X^2 &= n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(\hat{p}_{ij} - \hat{p}_i + \hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}} = 18.25 \\ G^2 &= 2n \sum_{i=1}^{r} \sum_{j=1}^{c} \hat{p}_{ij} \ln\left(\frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}}\right) = 18.25 \end{aligned}$$

Comparing each statistic to a $\chi^2_4$ distribution gives $p$-value= .001.

(b) Using (10.9), $E[X^2] \approx E[G^2] \approx 6.84$

(c) $X^2_F = G^2_F = \dfrac{4X^2}{6.84} = 10.7$, with $p$-value = .03 (comparing to a $\chi^2_4$ distribution).

**10.9** (a) Under the null hypothesis of independence, the expected proportions are:

| | Males | Females |
|---|---|---|
| Decision-making managers | 0.076 | 0.065 |
| Advisor-managers | 0.018 | 0.016 |
| Supervisors | 0.064 | 0.054 |
| Semi-autonomous workers | 0.103 | 0.087 |
| Workers | 0.279 | 0.238 |

Using (10.2) and (10.3),

$$\begin{aligned} X^2 &= n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}} = 55.1 \\ G^2 &= 2n \sum_{i=1}^{r} \sum_{j=1}^{c} \hat{p}_{ij} \ln\left(\frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}}\right) = 56.6 \end{aligned}$$

Comparing each statistic to a $\chi^2$ distribution with $(2-1)(5-1) = 4$ *df* gives "$p$-values" that are less than $1 \times 10^{-9}$.

(b) Using (10.9), we have

$$E[X^2] \approx E[G^2] \approx 4.45$$

(c) $df$ = number of psu's − number of strata = 34

(d)

$$\begin{aligned} X_F^2 &= \frac{4X^2}{4.45} = .899X^2 = 49.5 \\ G_F^2 &= \frac{4G^2}{4.45} = 50.8 \end{aligned}$$

The $p$-values for these statistics are still small, less than $1.0 \times 10^{-9}$.

(e) The $p$-value for $X_s^2$ is $2.6 \times 10^{-8}$, still very small.

**10.11** Here is SAS code and output:

```
options ovp nocenter ls=85;
filename nhanes  'C:\nhanes.csv';

data nhanes;
  infile nhanes delimiter=',' firstobs=2;
  input sdmvstra sdmvpsu wtmec2yr age ridageyr riagendr ridreth2
      dmdeduc indfminc bmxwt bmxbmi bmxtri
      bmxwaist bmxthicr  bmxarml;
  bmiclass = .;
  if 0 > bmxbmi and bmxbmi  < 25 then bmiclass = 1;
  else if bmxbmi >= 25 and bmxbmi <  30 then bmiclass = 2;
  else if bmxbmi >= 30 then bmiclass = 3;
  if age < 30 then ageclass = 1;
  else if age >= 30 then ageclass = 2;
  label age = "Age at Examination (years)"
        riagendr = "Gender"
        ridreth2 = "Race/Ethnicity"
        dmdeduc = "Education Level"
        indfminc = "Family income"
        bmxwt = "Weight (kg)"
        bmxbmi = "Body mass index"
        bmxtri = "Triceps skinfold (mm)"
        bmxwaist = "Waist circumference (cm)"
        bmxthicr = "Thigh circumference (cm)"
        bmxarml = "Upper arm length (cm)";
run;

proc surveyfreq data=nhanes ;
  stratum sdmvstra;
  cluster sdmvpsu;
```

```
  weight wtmec2yr;
  tables bmiclass*ageclass/chisq deff;
run;
```

Table of bmiclass by ageclass

| bmiclass | ageclass | Frequency | Weighted Frequency | Std Dev of Wgt Freq | Percent | Std Err of Percent | Design Effect |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 881 | 9566761 | 716532 | 6.0302 | 0.3131 | 0.8572 |
| | 2 | 75 | 2686500 | 495324 | 1.6934 | 0.2434 | 1.7639 |
| | Total | 956 | 12253261 | 1083245 | 7.7236 | 0.3951 | 1.0855 |
| 2 | 1 | 788 | 19494074 | 1615408 | 12.2878 | 0.6689 | 2.0573 |
| | 2 | 1324 | 57089320 | 4300350 | 35.9853 | 1.2334 | 3.2727 |
| | Total | 2112 | 76583394 | 5452762 | 48.2731 | 1.2969 | 3.3381 |
| 3 | 1 | 627 | 15269676 | 1538135 | 9.6250 | 0.6261 | 2.2338 |
| | 2 | 1262 | 54539886 | 4631512 | 34.3783 | 1.2349 | 3.3499 |
| | Total | 1889 | 69809562 | 5816243 | 44.0033 | 1.4066 | 3.9794 |
| Total | 1 | 2296 | 44330511 | 3336974 | 27.9430 | 1.0011 | 2.4666 |
| | 2 | 2661 | 114315706 | 8538003 | 72.0570 | 1.0011 | 2.4666 |
| | Total | 4957 | 158646217 | 11335366 | 100.000 | | |

Frequency Missing = 4686

Rao-Scott Chi-Square Test

| | |
|---|---|
| Pearson Chi-Square | 525.1560 |
| Design Correction | 1.6164 |
| Rao-Scott Chi-Square | 324.8848 |
| DF | 2 |
| Pr > ChiSq | <.0001 |
| F Value | 162.4424 |
| Num DF | 2 |
| Den DF | 30 |
| Pr > F | <.0001 |

There is strong evidence of an association.

**10.12** Here is SAS code and output:

```
data ncvs;
   infile  ncvs delimiter = "," firstobs=2;
   input age married sex race hispanic hhinc away employ numinc
```

```
       violent injury medtreat medexp robbery assault
        pweight pstrat ppsu;
   if violent > 0 then isviol = 1;
   else if violent = 0 then isviol = 0;
run;

proc surveyfreq data=ncvs;
   stratum pstrat ;
   cluster ppsu;
   weight pweight;
   tables  isviol*sex/chisq;
run;
```

Table of isviol by sex

| isviol | sex | Frequency | Weighted Frequency | Std Dev of Wgt Freq | Percent | Std Err of Percent |
|---|---|---|---|---|---|---|
| 0 | 0 | 36120 | 107752330 | 1969996 | 47.6830 | 0.1620 |
| | 1 | 42161 | 115149882 | 1906207 | 50.9566 | 0.1638 |
| | Total | 78281 | 222902212 | 3813397 | 98.6396 | 0.0663 |
| 1 | 0 | 558 | 1746669 | 101028 | 0.7729 | 0.0445 |
| | 1 | 441 | 1327436 | 81463 | 0.5874 | 0.0341 |
| | Total | 999 | 3074105 | 155677 | 1.3604 | 0.0663 |
| Total | 0 | 36678 | 109498999 | 1985935 | 48.4560 | 0.1597 |
| | 1 | 42602 | 116477318 | 1930795 | 51.5440 | 0.1597 |
| | Total | 79280 | 225976317 | 3853581 | 100.000 | |

Frequency Missing = 80

Rao-Scott Chi-Square Test

| | |
|---|---|
| Pearson Chi-Square | 30.6160 |
| Design Correction | 1.0466 |
| Rao-Scott Chi-Square | 29.2529 |
| DF | 1 |
| Pr > ChiSq | <.0001 |
| F Value | 29.2529 |
| Num DF | 1 |
| Den DF | 143 |
| Pr > F | <.0001 |

There is strong evidence that males are more likely to be victims of violent crime

than females.

**10.13** This test statistic does not in general give correct $p$-values for data from a complex survey. It ensures that the sum of the "observed" counts is $n$ but does not adjust for stratification or clustering.

To see this, note that for the data in Example 10.4, the proposed test statistic is the same as $X^2$ because all weights are equal. But in that example $X^2/2$, not $X^2$, has a null $\chi^2_1$ distribution because of the clustering.

**10.14** (a) For the Wald test,

$$\theta = p_{11}p_{22} - p_{12}p_{21}$$

and

$$\hat{\theta} = \hat{p}_{11}\hat{p}_{22} - \hat{p}_{12}\hat{p}_{21}.$$

Then, using Taylor linearization,

$$\hat{\theta} \approx \theta + p_{22}(\hat{p}_{11} - p_{11}) + p_{11}(\hat{p}_{22} - p_{22}) - p_{12}(\hat{p}_{21} - p_{21}) - p_{21}(\hat{p}_{12} - p_{12})$$

and

$$\begin{aligned} V_L(\hat{\theta}) &= V[p_{22}\hat{p}_{11} + p_{11}\hat{p}_{22} - p_{12}\hat{p}_{21} - p_{21}\hat{p}_{12}] \\ &= p_{22}^2 V(\hat{p}_{11}) + p_{11}^2 V(\hat{p}_{22}) + p_{12}^2 V(\hat{p}_{21}) + p_{21}^2 V(\hat{p}_{12}) \\ &\quad +2p_{11}p_{22}\ \mathrm{Cov}\,(\hat{p}_{11}, \hat{p}_{22}) - 2p_{22}p_{12}\ \mathrm{Cov}\,(\hat{p}_{11}, \hat{p}_{21}) \\ &\quad -2p_{22}p_{21}\ \mathrm{Cov}\,(\hat{p}_{11}, \hat{p}_{12}) - 2p_{11}p_{12}\ \mathrm{Cov}\,(\hat{p}_{22}, \hat{p}_{21}) \\ &\quad -2p_{11}p_{21}\ \mathrm{Cov}\,(\hat{p}_{22}, \hat{p}_{12}) + 2p_{12}p_{21}\ \mathrm{Cov}\,(\hat{p}_{21}, \hat{p}_{12}). \end{aligned}$$

To estimate $V_L(\hat{\theta})$, define

$$y_{jki} = \begin{cases} 1 & \text{if unit } i \text{ in cell } (j,k) \\ 0 & \text{otherwise} \end{cases}$$

for $j, k \in \{1, 2\}$ and let

$$q_i = \hat{p}_{22}y_{11i} + \hat{p}_{11}y_{12i} - \hat{p}_{12}y_{21i} - \hat{p}_{21}y_{22i}.$$

Then

$$\hat{V}_L(\hat{\theta}) = V(\hat{\bar{q}}).$$

(b) For multinomial sampling,

$$\begin{aligned} V_L(\hat{\theta}) &= p_{22}^2\frac{p_{11}(1-p_{11})}{n} + \cdots + p_{21}^2\frac{p_{12}(1-p_{12})}{n} \\ &\quad -2\frac{p_{11}^2p_{22}^2}{n} + 2\frac{p_{11}p_{22}p_{12}p_{21}}{n} + \cdots - 2\frac{p_{12}^2p_{21}^2}{n} \\ &= \frac{1}{n}\{-4p_{11}^2p_{22}^2 - 4p_{12}^2p_{21}^2 + 8p_{11}p_{22}p_{12}p_{21} \\ &\quad +p_{11}p_{22}(p_{11}+p_{22}) + p_{12}p_{21}(p_{12}+p_{21})\}. \end{aligned}$$

Under $H_0 : p_{11}p_{22} = p_{12}p_{21}$,

$$V_L(\hat{\theta}) = \frac{1}{n}p_{11}p_{22} = \frac{1}{n}p_{12}p_{21} = \frac{1}{n}p_{1+}p_{+1}p_{2+}p_{+2}$$

and

$$\frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}} = \frac{p_{11}+p_{22}}{p_{11}p_{22}} + \frac{p_{21}+p_{12}}{p_{12}p_{21}} = \frac{1}{p_{11}p_{22}}.$$

Thus, if $H_0$ is true,

$$\begin{aligned}\frac{1}{V_L(\hat{\theta})} &= n\left(\frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}\right) \\ &= n\left(\frac{1}{p_{1+}p_{+1}} + \frac{1}{p_{1+}p_{+2}} + \frac{1}{p_{2+}p_{+1}} + \frac{1}{p_{2+}p_{+2}}\right).\end{aligned}$$

Also note that, for any $j, k \in \{1, 2\}$,

$$\hat{\theta}^2 = (\hat{p}_{11}\hat{p}_{22} - \hat{p}_{12}\hat{p}_{21})^2 = (\hat{p}_{jk} - \hat{p}_{j+}\hat{p}_{+k})^2.$$

Thus, estimating $V_L(\hat{\theta})$ under $H_0$,

$$X^2_W = n\sum_{j=1}^{2}\sum_{k=1}^{2}\frac{(\hat{p}_{jk} - \hat{p}_{j+}\hat{p}_{+k})^2}{\hat{p}_{j+}\hat{p}_{+k}} = X^2_p.$$

**10.15** (a) We can rewrite

$$\theta = \log(p_{11}) + \log(p_{22}) - \log(p_{12}) - \log(p_{21}).$$

Then, using Taylor linearization,

$$\hat{\theta} \approx \theta + \frac{\hat{p}_{11} - p_{11}}{p_{11}} + \frac{\hat{p}_{22} - p_{22}}{p_{22}} - \frac{\hat{p}_{12} - p_{12}}{p_{12}} - \frac{\hat{p}_{21} - p_{21}}{p_{21}}$$

and

$$\begin{aligned}V_L(\hat{\theta}) &= V\left(\frac{\hat{p}_{11}}{p_{11}} + \frac{\hat{p}_{22}}{p_{22}} - \frac{\hat{p}_{12}}{p_{12}} - \frac{\hat{p}_{21}}{p_{21}}\right) \\ &= \sum_{j=1}^{2}\sum_{k=1}^{2}\frac{1}{p^2_{jk}}V(\hat{p}_{jk}) \\ &\quad + \frac{2}{p_{11}p_{22}}\,\mathrm{Cov}\,(\hat{p}_{11}, \hat{p}_{22}) - \frac{2}{p_{11}p_{12}}\,\mathrm{Cov}\,(\hat{p}_{11}, \hat{p}_{12}) \\ &\quad - \frac{2}{p_{11}p_{21}}\,\mathrm{Cov}\,(\hat{p}_{11}, \hat{p}_{21}) - \frac{2}{p_{12}p_{22}}\,\mathrm{Cov}\,(\hat{p}_{22}, \hat{p}_{12}) \\ &\quad - \frac{2}{p_{22}p_{21}}\,\mathrm{Cov}\,(\hat{p}_{22}, \hat{p}_{21}) + \frac{2}{p_{12}p_{21}}\,\mathrm{Cov}\,(\hat{p}_{12}, \hat{p}_{21}).\end{aligned}$$

(b) Under multinomial sampling,

$$\begin{aligned}V_L(\hat{\theta}) &= \sum_{j=1}^{2}\sum_{k=1}^{2}\frac{p_{jk}(1-p_{jk})}{np^2_{jk}} + \frac{4}{n} \\ &= \frac{1}{n}\left(\frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}\right)\end{aligned}$$

and

$$\begin{aligned}\hat{V}_L(\hat{\theta}) &= \frac{1}{n}\left(\frac{1}{\hat{p}_{11}}+\frac{1}{\hat{p}_{12}}+\frac{1}{\hat{p}_{21}}+\frac{1}{\hat{p}_{22}}\right)\\ &= \frac{1}{x_{11}}+\frac{1}{x_{12}}+\frac{1}{x_{21}}+\frac{1}{x_{22}}.\end{aligned}$$

This is the estimated variance given in Section 10.1.1.

**10.17** In a multinomial sample, all design effects are 1. From (10.9), under $H_0$

$$\begin{aligned}E[X^2] &= \sum_{i=1}^{r}\sum_{j=1}^{c}(1-p_{ij})-\sum_{i=1}^{r}(1-p_{i+})-\sum_{j=1}^{c}(1-p_{+j})\\ &= rc-1-(r-1)-(c-1)\\ &= (r-1)(c-1).\end{aligned}$$

**10.18** (a) Write

$$\mathbf{Y}^T\mathbf{C}\mathbf{Y} = \mathbf{Y}^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{C}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}\mathbf{Y}.$$

Since $\mathbf{C}$ is symmetric and positive definite, so is $\boldsymbol{\Sigma}^{1/2}\mathbf{C}\boldsymbol{\Sigma}^{1/2}$ and we can write $\boldsymbol{\Sigma}^{1/2}\mathbf{C}\boldsymbol{\Sigma}^{1/2} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T$ for an orthogonal matrix $\mathbf{P}$ and diagonal matrix $\boldsymbol{\Lambda}$, where each diagonal entry of $\boldsymbol{\Lambda}$ is positive. Let $\mathbf{U} = \mathbf{P}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{Y}$; then

$$\mathbf{Y}^T\mathbf{C}\mathbf{Y} = \mathbf{U}^T\boldsymbol{\Lambda}\mathbf{U} = \sum_{i=1}^{k}\lambda_i U_i^2.$$

Since $\mathbf{Y} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $\mathbf{U} \sim N(\mathbf{0}, \mathbf{P}^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2}\mathbf{P}) = N(\mathbf{0}, \mathbf{I})$, so $W_i = U_i^2 \sim \chi^2_1$ and the $W_i$'s are independent.

(b) Using a central limit theorem for survey sampling, we know that $\mathbf{V}(\hat{\boldsymbol{\theta}})^{-1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ has asymptotic $N(\mathbf{0}, \mathbf{I})$ distribution under $H_0 : \boldsymbol{\theta} = \mathbf{0}$. Using part (a), then,

$$\hat{\boldsymbol{\theta}}^T\mathbf{A}^{-1}\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^T\mathbf{V}(\hat{\boldsymbol{\theta}})^{-1/2}\mathbf{V}(\hat{\boldsymbol{\theta}})^{1/2}\mathbf{A}^{-1}\mathbf{V}(\hat{\boldsymbol{\theta}})^{1/2}\mathbf{V}(\hat{\boldsymbol{\theta}})^{-1/2}\hat{\boldsymbol{\theta}}$$

has the same asymptotic distribution as $\sum\lambda_i W_i$, where the $\lambda_i$'s are the eigenvalues of

$$\mathbf{V}(\hat{\boldsymbol{\theta}})^{1/2}\mathbf{A}^{-1}\mathbf{V}(\hat{\boldsymbol{\theta}})^{1/2}.$$

(c)

$$E[\hat{\boldsymbol{\theta}}^T\mathbf{A}^{-1}\hat{\boldsymbol{\theta}}] \approx \sum\lambda_i,$$

$$V[\hat{\boldsymbol{\theta}}^T\mathbf{A}^{-1}\hat{\boldsymbol{\theta}}] \approx 2\sum\lambda_i,$$

**10.19** This sample is self-weighting, so the estimated cell probabilities are

| | $S$ | $N$ | |
|---|---|---|---|
| Male | .3028 | .1056 | .4084 |
| Female | .2254 | .3662 | .5916 |
| | .5282 | .4718 | 1.0000 |

The variances under the (incorrect) assumption of multinomial sampling are:

| | $S$ | $N$ | |
|---|---|---|---|
| Male | .001487 | .000665 | .001701 |
| Female | .001229 | .001634 | .001701 |
| | .001755 | .001755 | |

We use the information in Table 10.1 to find the estimated variances for each cell and margin using the cluster sample. For the schizophrenic males, define

$$t_{\text{SMi}} = \text{ number of schizophrenic males in psu } i,$$

and similarly for the other cells. Then we use equations (5.4)–(5.6) to estimate the mean and variance for each cell. We have the following frequency data:

| Freq. | SM | SF | NM | NF | M | F | S | N |
|---|---|---|---|---|---|---|---|---|
| 0 | 41 | 45 | 58 | 34 | 30 | 17 | 24 | 28 |
| 1 | 17 | 20 | 11 | 22 | 24 | 24 | 19 | 19 |
| 2 | 13 | 6 | 2 | 15 | 17 | 30 | 28 | 24 |
| $\hat{\bar{y}}$ | .3028 | .2254 | .1056 | .3662 | .4085 | .5915 | .5282 | .4718 |
| $\hat{V}(\hat{\bar{y}})$ | .0022 | .0015 | .0008 | .0022 | .0022 | .0022 | .0026 | .0026 |

Thus the estimated variances using the clustering are

| | $S$ | $N$ | |
|---|---|---|---|
| Male | .002161 | .000796 | .002244 |
| Female | .001488 | .002209 | .002244 |
| | .002604 | .002604 | |

and the estimated design effects are

| | $S$ | $N$ | |
|---|---|---|---|
| Male | 1.453 | 1.197 | 1.319 |
| Female | 1.210 | 1.352 | 1.319 |
| | 1.484 | 1.484 | |

Using equation (10.9), $E[X^2]$ is estimated by

$$\sum_{i=1}^{2}\sum_{j=1}^{2}(1-\hat{p}_{ij})d_{ij} - \sum_{i=1}^{2}(1-\hat{p}_{i+})d_i^R - \sum_{j=1}^{2}(1-\hat{p}_{+j})d_j^c = 1.075.$$

Then

$$X_F^2 = \frac{X^2}{1.07} = \frac{17.89}{1.07} = 16.7$$

with $p$-value $< 0.0001$.

**10.20**

Both statistics are very large. We obtain the Rao-Scott $\chi^2$ statistic is 2721, while the Wald test statistic is 4838. There is strong evidence that the variables are associated.

# Chapter 11

# Regression with Complex Survey Data

**11.3** The average score for the students planning a trip is $\bar{y}_1 = 77.158730$ and the average score for the students not planning a trip is $\bar{y}_2 = 61.887218$. Using SAS PROC SURVEYREG, we get $\bar{y}_1 - \bar{y}_2 = 15.27$ with 95% CI [7.6247634, 22.9182608]. Since 0 is not in the CI, there is evidence that the domain means differ.

**11.4** (a) From SAS, the fitted regression line for the truncated data set is

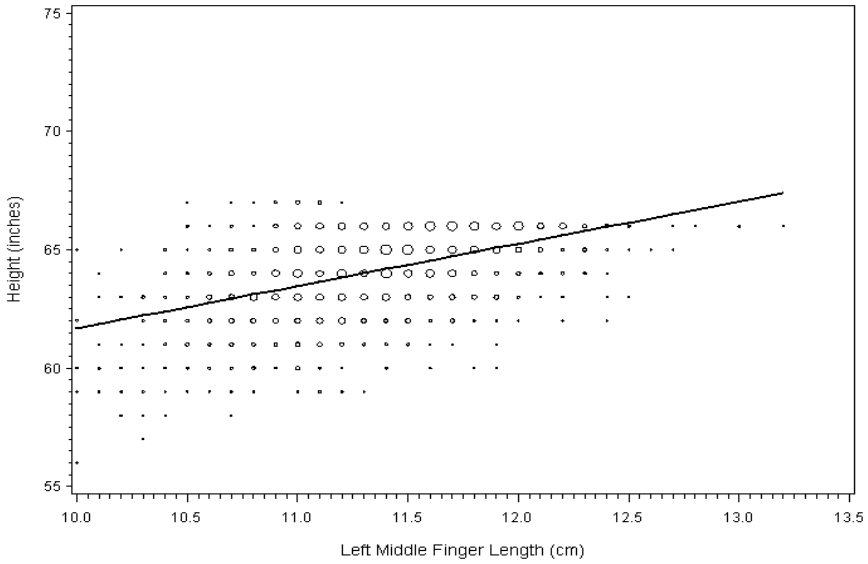
```
data anthrop;
   infile anthrop firstobs=2 delimiter=",";
   input finger height ;
   one = 1;
run;

proc sort data=anthrop;
   by  height;

data anthrop1; /* Keep the lowest 2000 values in the data set */
   set anthrop;
   if _N_ <= 2000;
run;

proc reg data = anthrop1;
   model height = finger;
   output out=regpred pred=predicted residual=resid;
run;

goptions reset=all;
goptions colors = (black);
```

```
axis1 label=('Left Middle Finger Length (cm)')
               order =  (10 to 13.5 by .5);
axis2 label=(angle=90 'Height (inches)') order=(55 to 75 by 5);
axis3  order=(55 to 75 by 5) major=none minor=none value=none;
symbol interpol=join width=2 color = black;

proc sort data=regpred;
   by finger height;
run;

proc means data=regpred noprint;
   by finger height;
   var  one predicted resid;
   output out=circlepred  sum=sumn sumpred sumresid
                          mean = meanone meanpred meanresid;
run;

proc gplot data=circlepred;
   bubble height*finger=sumn/haxis=axis1 vaxis=axis2;
   plot2 meanpred*finger/haxis=axis1 vaxis =axis3 ;
run;
```

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 43.78957 | 0.80553 | 54.36 | <.0001 |
| finger | 1 | 1.78861 | 0.07094 | 25.21 | <.0001 |

The line is much flatter than one the on Figure 11.4.



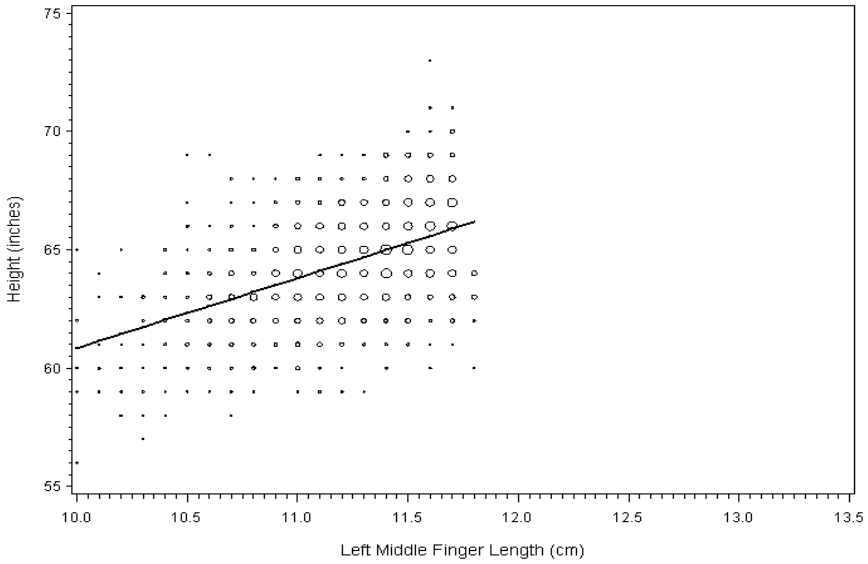

(b) We use exactly the same code as before, except now we sort the data by finger instead of by height.

```
proc sort data=anthrop;
   by  finger;

data anthrop2;
   set anthrop;
   if _N_ <= 2000;
run;

proc reg data = anthrop2;
   model height = finger;
   output out=regpred pred=predicted residual=resid;
run;
```

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 31.19794 | 1.33050 | 23.45 | <.0001 |
| finger | 1 | 2.96393 | 0.11821 | 25.07 | <.0001 |

These values of the slope and intercept are quite close to the values given in Figure 11.4. The standard errors are larger, however, reflecting the lesser number of data points and the reduced spread of the $x$'s.




(c) Regression uses conditional expectation, conditional on $x$. Thus, if the model holds for all the observations, you should get unbiased estimates of the parameters if you take a subset of the data using $x$ to define the mechanism.

**11.5** We obtain $\hat{B}_0 = 14.2725$ and $\hat{B}_1 = 0.08138$. Using equation (11.8), with $q_i = (y_i - \hat{B}_0 - \hat{B}_1 x_i)(x_i - \hat{\bar{x}})$ and $\hat{\bar{x}} = \sum w_i x_i / \sum w_i = 180.541$, we have

$$\begin{aligned} \hat{V}_L(\hat{B}_1) &= \hat{V}\left(\sum w_i q_i\right) \Big/ \left[\sum w_i x_i^2 - \frac{(\sum w_i x_i)^2}{\sum w_i}\right]^2 \\ &= \hat{V}\left(\frac{\sum w_i q_i}{\sum w_i x_i^2 - (\sum w_i x_i)^2 / \sum w_i}\right) \\ &= 0.000261 \\ SE_L(\hat{B}_1) &= .016. \end{aligned}$$

Here is output from SAS:

```
data nybight;
   infile nybight delimiter="," firstobs=2;
   input year  stratum  catchnum catchwt  numspp  depth  temp ;
   if stratum = 1 or stratum = 2 then relwt = 1;
   else if (stratum ge 3 and stratum le 6) then relwt = 2;
   if year = 1974;
run;

proc surveyreg data=nybight;
   weight relwt;
   stratum stratum;
   model catchwt = catchnum;
run;
```

Estimated Regression Coefficients

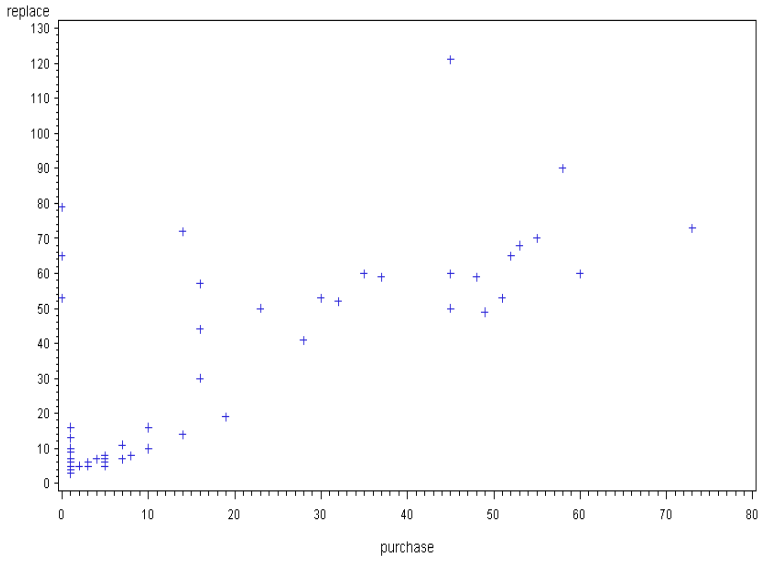
| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 14.2725002 | 2.13426397 | 6.69 | <.0001 |
| catchnum | 0.0813836 | 0.01612252 | 5.05 | <.0001 |

**11.7** Using the weights as in Exercise 11.4, the estimated regression coefficients are $\hat{B}_0 = 7.569$ and $\hat{B}_1 = 0.0778$. From equation (11.8), $\hat{V}_L(\hat{B}_1) = 0.068$, (alternatively, $\hat{V}_{JK}(\hat{B}_1) = 0.070$). The slope is not significantly different from 0. Here is SAS output:

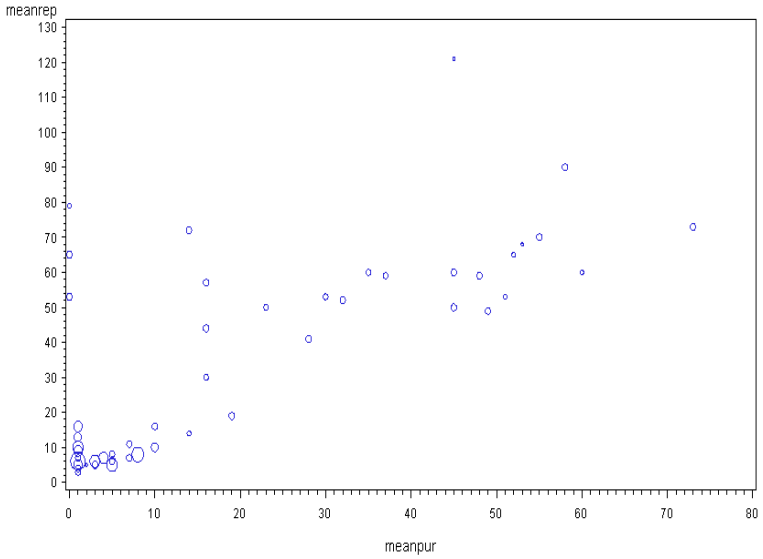
Estimated Regression Coefficients

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| Intercept | 7.56852711 | 4.31739366 | 1.75 | 0.0879 | -1.1793434 | 16.3163976 |
| temp | 0.07780553 | 0.26477065 | 0.29 | 0.7705 | -0.4586708 | 0.6142818 |

**11.10** (a)




(b)




(c) Here is output from SAS.

Estimated Regression Coefficients

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| Intercept | 9.61708567 | 3.56639563 | 2.70 | 0.0208 | 1.76750181 | 17.4666695 |
| purchase | 1.08119004 | 0.12136848 | 8.91 | <.0001 | 0.81405982 | 1.3483203 |

NOTE: The denominator degrees of freedom for the t tests is 11.

We use (number of psus) $-$ (number of strata) $= 12 - 1 = 11$ df.

**11.14** Here is code and output from SAS:

```
data nhanes;
  infile nhanes delimiter=',' firstobs=2;
  input sdmvstra sdmvpsu wtmec2yr age ridageyr riagendr ridreth2
      dmdeduc indfminc bmxwt bmxbmi bmxtri
      bmxwaist bmxthicr  bmxarml;
  if riagendr = 1 then x = 0; /* x=0 is male*/
  if riagendr = 2 then x = 1; /* x=1 is female */
  if age ge 15 then over15 = 1;
  else if age lt 15 then over15 = 0;
  else over15=.;
  one = 1;
  label age = "Age at Examination (years)"
```

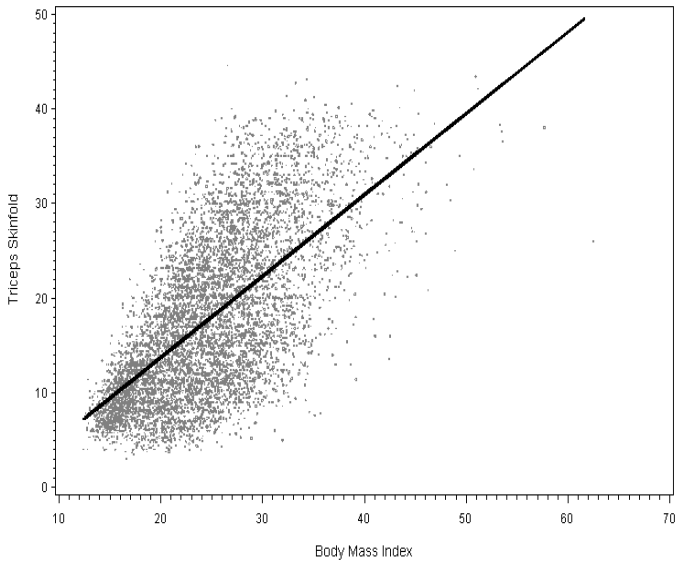
```
        agesq = "Age^2"
        riagendr = "Gender"
        ridreth2 = "Race/Ethnicity"
        dmdeduc = "Education Level"
        indfminc = "Family income"
        bmxwt = "Weight (kg)"
        bmxbmi = "Body mass index"
        bmxtri = "Triceps skinfold (mm)"
        bmxwaist = "Waist circumference (cm)"
        bmxthicr = "Thigh circumference (cm)"
        bmxarml = "Upper arm length (cm)";
run;

proc surveyreg data=nhanes;
  stratum sdmvstra;
  cluster sdmvpsu;
  weight wtmec2yr;
  model bmxtri=bmxbmi /clparm;
  output out=quad pred=quadpred residual=quadres;
  ods output ParameterEstimates = quadcoefs;
run;

/* Do a weighted bubble plot of data with the regression line. */

goptions reset=all;
goptions colors = (gray);
axis4 label=(angle=90 'Triceps Skinfold') order=(0 to 50 by 10);
axis3 label=('Body Mass Index') order =  (10 to 70 by 10);
axis5  order=(0 to 50 by 10) major=none minor=none value=none;
symbol interpol=join width=2 color = black;

proc sort data=quad;
   by  bmxtri bmxbmi;

proc means data=quad noprint;
   by bmxtri bmxbmi;
   var  wtmec2yr quadpred quadres;
   output out=quadplot  sum=sumwts sumquad sumres
                        mean=meanwt meanquad meanres;
run;

proc gplot data=quadplot;
   bubble bmxtri*bmxbmi= sumwts/bsize=10 haxis = axis3 vaxis=axis4;
   plot2 meanquad*bmxbmi/ vaxis=axis5;
run;
```

```
/* Plot residuals vs predicted values */

goptions reset=all;
goptions colors = (gray);
axis3 label=('Predicted Values') order =  (15 to 30 by 5);
axis4 label=(angle=90 'Residuals') order=(-20 to 40 by 10);
axis5  order=(10 to 70 by 10) major=none minor=none value=none;

proc gplot data=quadplot;
   bubble meanres*meanquad= sumwts;
run;
```

Estimated Regression Coefficients

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| Intercept | -3.4248758 | 0.47196343 | -7.26 | <.0001 | -4.4308420 | -2.4189096 |
| bmxbmi | 0.8581404 | 0.02265388 | 37.88 | <.0001 | 0.8098548 | 0.9064260 |

NOTE: The denominator degrees of freedom for the t tests is 15.




$R^2 = 0.38$. Note the pattern in the residuals vs. predicted values plot. You may want to use a model with log transformations instead.

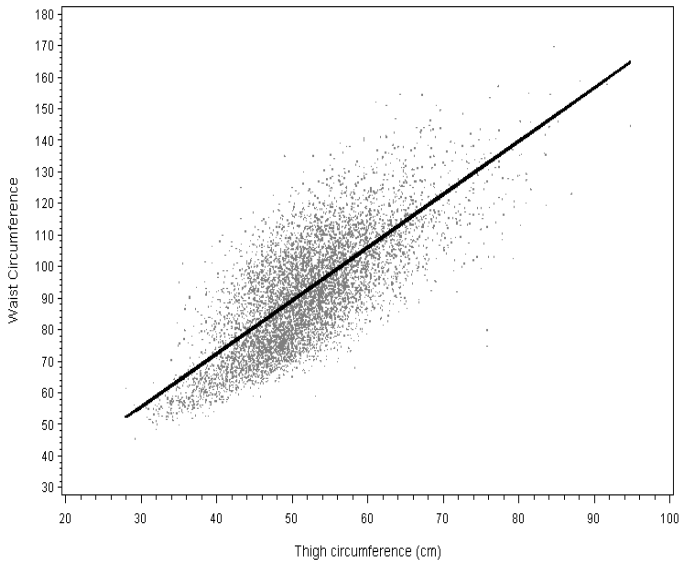
**11.15**

Estimated Regression Coefficients

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| Intercept | 5.02260509 | 1.38143897 | 3.64 | 0.0024 | 2.07813762 | 7.96707255 |
| bmxthicr | 1.68332900 | 0.02540473 | 66.26 | <.0001 | 1.62918010 | 1.73747790 |

NOTE: The denominator degrees of freedom for the t tests is 15.

$R^2 = 0.57$.




### 11.16

```
data ncvs;
   infile  ncvs delimiter = ",";
   input age married sex race hispanic hhinc away employ numinc
       violent injury medtreat medexp robbery assault
        pweight pstrat ppsu;
   agesq = age*age;
   if violent ge 1 then isviol = 1;
   else if violent = 0 then isviol = 0;
run;

proc surveylogistic data=ncvs;
  stratum pstrat;
  cluster ppsu;
  weight pweight;
  model isviol (event='1')= age sex ;
run;
```

```
proc surveylogistic data=ncvs;
  stratum pstrat;
  cluster ppsu;
  weight pweight;
  model isviol (event='1')= age agesq sex ;
run;
```

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -2.6743 | 0.0891 | 900.7557 | <.0001 |
| age | 1 | -0.0418 | 0.00211 | 392.9618 | <.0001 |
| sex | 1 | -0.2925 | 0.0637 | 21.0673 | <.0001 |

From this model, younger people and males are more likely to have at least one violent victimization. A quadratic term in age is not significant.

**11.17** From (11.4),

$$\begin{aligned} B_1 &= \frac{\sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)\Big/N}{\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2 /N} \\ &= \frac{N_1\bar{y}_{1U} - N_1\bar{y}_U}{N_1 - N_1^2/N} \\ &= \frac{\bar{y}_{1U} - (N_1\bar{y}_{1U} + N_2\bar{y}_{2U})/N}{1 - N_1/N} \\ &= \bar{y}_{1U} - \bar{y}_{2U}. \end{aligned}$$

From (11.5),

$$\begin{aligned} B_0 &= \frac{t_y - B_1 t_x}{N} \\ &= \bar{y}_U - B_1\bar{x}_U \\ &= \frac{N_1\bar{y}_{1U} + N_2\bar{y}_{2U}}{N} - \frac{N_1}{N}(\bar{y}_{1U} - \bar{y}_{2U}) \\ &= \bar{y}_{2U}. \end{aligned}$$

**11.18** (a)

(b) $\hat{\beta}_0 = -4.096$; $\hat{\beta}_1 = 6.049$

(c) We estimate the model-based variance using the regression software: $\hat{V}_M(\hat{\beta}_1) =$

0.541.

$$\begin{aligned}\hat{V}_L(\hat{\beta}_1) &= \frac{n\sum(x_i-\bar{x})^2(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2}{(n-1)[\sum(x_i-\bar{x})^2]^2} \\ &= 0.685\end{aligned}$$

$\hat{V}_L$ is larger, as we would expect since the plot exhibits unequal variances.

**11.19** From (11.10), for straight-line regression,

$$\hat{\mathbf{B}} = \left(\sum_{i\in\mathcal{S}} w_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_{i\in\mathcal{S}} w_i \mathbf{x}_i y_i$$

with $\mathbf{x}_i = [1\ x_i]^T$. Here,

$$\sum_{i\in\mathcal{S}} w_i \mathbf{x}_i \mathbf{x}_i^T = \begin{bmatrix} \sum_{i\in\mathcal{S}} w_i & \sum_{i\in\mathcal{S}} w_i x_i \\ \sum_{i\in\mathcal{S}} w_i x_i & \sum_{i\in\mathcal{S}} w_i x_i^2 \end{bmatrix}$$

and

$$\sum_{i\in\mathcal{S}} w_i \mathbf{x}_i y_i = \begin{bmatrix} \sum_{i\in\mathcal{S}} w_i y_i \\ \sum_{i\in\mathcal{S}} w_i x_i y_i \end{bmatrix},$$

so

$$\begin{aligned}\hat{\mathbf{B}} &= \frac{1}{\left(\sum_{i\in\mathcal{S}} w_i\right)\left(\sum_{i\in\mathcal{S}} w_i x_i^2\right) - \left(\sum_{i\in\mathcal{S}} w_i x_i\right)^2} \\ &\quad \begin{bmatrix} \sum_{i\in\mathcal{S}} w_i x_i^2 & -\sum_{i\in\mathcal{S}} w_i x_i \\ -\sum_{i\in\mathcal{S}} w_i x_i & \sum_{i\in\mathcal{S}} w_i \end{bmatrix} \begin{bmatrix} \sum_{i\in\mathcal{S}} w_i y_i \\ \sum_{i\in\mathcal{S}} w_i x_i y_i \end{bmatrix}\end{aligned}$$

Thus,

$$\hat{B}_1 = \frac{\sum_{i\in\mathcal{S}} w_i x_i y_i - \left(\sum_{i\in\mathcal{S}} w_i x_i\right)\left(\sum_{i\in\mathcal{S}} w_i y_i\right) \Big/ \left(\sum_{i\in\mathcal{S}} w_i\right)}{\sum_{i\in\mathcal{S}} w_i x_i^2 - \left(\sum_{i\in\mathcal{S}} w_i x_i\right)^2 \Big/ \left(\sum_{i\in\mathcal{S}} w_i\right)}$$

and

$$\begin{aligned}\hat{B}_0 &= \frac{\left(\sum_{i\in\mathcal{S}} w_i x_i^2\right)\left(\sum_{i\in\mathcal{S}} w_i y_i\right) - \left(\sum_{i\in\mathcal{S}} w_i x_i\right)\left(\sum_{i\in\mathcal{S}} w_i x_i y_i\right)}{\left(\sum_{i\in\mathcal{S}} w_i\right)\left(\sum_{i\in\mathcal{S}} w_i x_i^2\right) - \left(\sum_{i\in\mathcal{S}} w_i x_i\right)^2} \\ &= \left(\sum_{i\in\mathcal{S}} w_i\right)^{-1}\left[\sum_{i\in\mathcal{S}} w_i y_i - \hat{B}_1 \sum_{i\in\mathcal{S}} w_i x_i\right].\end{aligned}$$

**11.20** In matrix terms,

$$\hat{\mathbf{B}} = \left( \sum_{j \in \mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \sum_{j \in \mathcal{S}} w_j \mathbf{x}_j y_j.$$

Then, using the matrix identity in the hint,

$$\begin{aligned}
\mathbf{z}_i &= \frac{\partial \hat{\mathbf{B}}}{\partial w_i} \\
&= \left[ \frac{\partial}{\partial w_i} \left( \sum_{j \in \mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \right] \sum_{j \in \mathcal{S}} w_j \mathbf{x}_j y_j \\
&\qquad + \left( \sum_{j \in \mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \left[ \frac{\partial}{\partial w_i} \sum_{j \in \mathcal{S}} w_j \mathbf{x}_j y_j \right] \\
&= -\left( \sum_{j \in \mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \mathbf{x}_i \mathbf{x}_i^T \left( \sum_{j \in \mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \sum_{j \in \mathcal{S}} w_j \mathbf{x}_j y_j \\
&\qquad + \left( \sum_{j \in \mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \mathbf{x}_i y_i \\
&= \left( \sum_{j \in \mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \left( -\mathbf{x}_i \mathbf{x}_i^T \hat{\mathbf{B}} + \mathbf{x}_i y_i \right) \\
&= \left( \sum_{j \in \mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \mathbf{x}_i \left( y_i - \mathbf{x}_i^T \hat{\mathbf{B}} \right).
\end{aligned}$$

Then the estimated variance is

$$\begin{aligned}
\hat{V}(\hat{\mathbf{B}}) &= \hat{V} \left( \sum_{i \in \mathcal{S}} w_i \mathbf{z}_i \right) \\
&= \left( \sum_{j \in \mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \hat{V} \left( \sum_{i \in \mathcal{S}} w_i \mathbf{q}_i \right) \left( \sum_{j \in \mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1}.
\end{aligned}$$

**11.21** First express the estimator as a function of the weights:

$$\hat{t}_{y\text{GREG}} = \hat{t}_y + (\mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_\mathbf{x})^T \hat{\mathbf{B}},$$

with

$$\hat{\mathbf{B}} = \left( \sum_{i \in \mathcal{S}} w_i \frac{1}{\sigma_i^2} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i \in \mathcal{S}} w_i \frac{1}{\sigma_i^2} \mathbf{x}_i y_i.$$

Thus,

$$\hat{t}_{y\text{GREG}} = \sum_{i\in\mathcal{S}} w_i y_i + \left(\mathbf{t}_{\mathbf{x}} - \sum_{i\in\mathcal{S}} w_i \mathbf{x}_i\right)^T \hat{\mathbf{B}}$$

Using the same argument as in Exercise 11.20,

$$\begin{aligned}
\frac{\partial \hat{\mathbf{B}}}{\partial w_i} &= \left[\frac{\partial}{\partial w_i}\left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2}\mathbf{x}_j\mathbf{x}_j^T\right)^{-1}\right]\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2}\mathbf{x}_j y_j \\
&\quad + \left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2}\mathbf{x}_j\mathbf{x}_j^T\right)^{-1}\left[\frac{\partial}{\partial w_i}\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2}\mathbf{x}_j y_j\right] \\
&= -\left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2}\mathbf{x}_j\mathbf{x}_j^T\right)^{-1}\frac{1}{\sigma_i^2}\mathbf{x}_i\mathbf{x}_i^T\left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2}\mathbf{x}_j\mathbf{x}_j^T\right)^{-1}\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2}\mathbf{x}_j y_j \\
&\quad + \left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2}\mathbf{x}_j\mathbf{x}_j^T\right)^{-1}\frac{1}{\sigma_i^2}\mathbf{x}_i y_i \\
&= \left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2}\mathbf{x}_j\mathbf{x}_j^T\right)^{-1}\left(-\frac{1}{\sigma_i^2}\mathbf{x}_i\mathbf{x}_i^T\hat{\mathbf{B}} + \frac{1}{\sigma_i^2}\mathbf{x}_i y_i\right) \\
&= \left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2}\mathbf{x}_j\mathbf{x}_j^T\right)^{-1}\frac{1}{\sigma_i^2}\mathbf{x}_i\left(y_i - \mathbf{x}_i^T\hat{\mathbf{B}}\right)
\end{aligned}$$

$$\begin{aligned}
z_i &= \frac{\partial \hat{t}_{y\text{GREG}}}{\partial w_i} \\
&= y_i - \mathbf{x}_i^T\hat{\mathbf{B}} + \left(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}}\right)^T\frac{\partial \hat{\mathbf{B}}}{\partial w_i} \\
&= y_i - \mathbf{x}_i^T\hat{\mathbf{B}} + \left(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}}\right)^T\left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2}\mathbf{x}_j\mathbf{x}_j^T\right)^{-1}\frac{1}{\sigma_i^2}\mathbf{x}_i\left(y_i - \mathbf{x}_i^T\hat{\mathbf{B}}\right) \\
&= g_i(y_i - \mathbf{x}_i^T\hat{\mathbf{B}}).
\end{aligned}$$

**11.26** The OLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Thus,

$$\begin{aligned}
\text{Cov}_M(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{Cov}\,(\mathbf{Y})[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}.
\end{aligned}$$

For a straight-line regression model,

$$\begin{aligned}\mathbf{X}^T\mathbf{X} &= \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \\ (\mathbf{X}^T\mathbf{X})^{-1} &= \frac{1}{\sum(x_i-\bar{x})^2}\begin{bmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}\end{aligned}$$

and

$$\mathbf{X}^T\mathbf{\Sigma}\mathbf{X} = \begin{bmatrix} \sum \sigma_i^2 & \sum x_i\sigma_i^2 \\ \sum x_i\sigma_i^2 & \sum x_i^2\sigma_i^2 \end{bmatrix}.$$

Thus, in straight-line regression,

$$\mathrm{Cov}_M(\hat{\boldsymbol{\beta}}) = \frac{1}{[\sum(x_i-\bar{x})^2]^2}\begin{bmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}\begin{bmatrix} \sum \sigma_i^2 & \sum x_i\sigma_i^2 \\ \sum x_i\sigma_i^2 & \sum x_i^2\sigma_i^2 \end{bmatrix}\begin{bmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

$$\begin{aligned}V_M(\hat{\beta}_1) &= \left(\bar{x}^2\sum\sigma_i^2 - 2\bar{x}\sum x_i\sigma_i^2 + \sum x_i^2\sigma_i^2\right) / \left[\sum(x_i-\bar{x})^2\right]^2 \\ &= \frac{\sum(x_i-\bar{x})^2\sigma_i^2}{[\sum(x_i-\bar{x})^2]^2}.\end{aligned}$$

To see the relation to Section 11.2.1, let

$$Q_i = Y_i - \beta_0 - \beta_1 x_i.$$

Then $V_M(Q_i) = \sigma_i^2$; since observations are independent under the model,

$$V_M\left[\sum_i (x_i-\bar{x})Q_i\right] = \sum_i (x_i-\bar{x})^2\sigma_i^2.$$

If we take $w_i = 1$ for all $i$, then Equation (11.8) provides an estimate of $V_M(\hat{\beta}_1)$.

**11.29** For this model (see Exercise 11.19)

$$\mathbf{X}_\mathcal{S}^T\mathbf{W}_\mathcal{S}\mathbf{\Sigma}_\mathcal{S}^{-1}\mathbf{X}_\mathcal{S} = \frac{1}{\sigma^2}\begin{bmatrix} \sum_{i\in\mathcal{S}} w_i & \sum_{i\in\mathcal{S}} w_i x_i \\ \sum_{i\in\mathcal{S}} w_i x_i & \sum_{i\in\mathcal{S}} w_i x_i^2 \end{bmatrix}$$

and

$$\mathbf{X}_\mathcal{S}^T\mathbf{W}_\mathcal{S}\mathbf{\Sigma}_\mathcal{S}^{-1}\mathbf{y}_\mathcal{S} = \frac{1}{\sigma^2}\begin{bmatrix} \sum_{i\in\mathcal{S}} w_i y_i \\ \sum_{i\in\mathcal{S}} w_i x_i y_i \end{bmatrix} = \frac{1}{\sigma^2}\begin{bmatrix} \hat{t}_y \\ \hat{t}_{xy} \end{bmatrix}.$$

Using (11.20),

$$\hat{\mathbf{B}} = \begin{bmatrix} \hat{N} & \hat{t}_x \\ \hat{t}_x & \hat{t}_{x^2} \end{bmatrix}^{-1}\begin{bmatrix} \hat{t}_y \\ \hat{t}_{xy} \end{bmatrix} = \frac{1}{\hat{N}\hat{t}_{x^2} - (\hat{t}_x)^2}\begin{bmatrix} \hat{t}_{x^2}\hat{t}_y - \hat{t}_x\hat{t}_{xy} \\ -\hat{t}_x\hat{t}_y + \hat{N}\hat{t}_{xy} \end{bmatrix}$$

and

$$\begin{aligned}\hat{t}_{y\text{GREG}} &= \hat{t}_y + [N - \hat{N}, \quad t_x - \hat{t}_x]\hat{\mathbf{B}} \\ &= \hat{t}_y + \frac{1}{\hat{N}\hat{t}_{x^2} - (\hat{t}_x)^2}[(N - \hat{N})(\hat{t}_{x^2}\hat{t}_y - \hat{t}_x\hat{t}_{xy}) + (t_x - \hat{t}_x)(-\hat{t}_x\hat{t}_y + \hat{N}\hat{t}_{xy})].\end{aligned}$$

If $y = x$,

$$\begin{aligned}\hat{t}_{x\text{GREG}} &= \hat{t}_x + \frac{1}{\hat{N}\hat{t}_{x^2} - (\hat{t}_x)^2}[0 + (t_x - \hat{t}_x)(-\hat{t}_x^2 + \hat{N}\hat{t}_{x^2})] \\ &= t_x.\end{aligned}$$

# Chapter 12

# Two-Phase Sampling

**12.1** We use (12.4) to estimate the total, and (12.7) to estimate its variance. We obtain

$$\hat{t}_{\text{str}}^{(2)} = \frac{100{,}000}{1000} \sum_{\text{cells}} \frac{n_h}{m_h} r_h = 48310.$$

From (12.7), we estimate $s_h^{2(2)} = \hat{p}_h^{(2)}(1 - \hat{p}_h^{(2)})/(m_h - 1)$ and obtain

$$\begin{aligned} \hat{V}\left(\hat{t}_{\text{str}}^{(2)}\right) &= N(N-1) \sum_{h=1}^{H} \left( \frac{n_h - 1}{n - 1} - \frac{m_h - 1}{N - 1} \right) \frac{n_h}{n} \frac{s_h^{2(2)}}{m_h} \\ &\quad + \frac{N^2}{n-1} \left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} \frac{n_h}{n} (\bar{y}_h^{(2)} - \hat{\bar{y}}_{\text{str}}^{(2)})^2 \\ &= 100000(99999)0.000912108 + \frac{100000^2}{999} 0.109282588 \\ &= 10214904. \end{aligned}$$

Thus $\text{SE}(\hat{t}_{\text{str}}^{(2)}) = 3196$.

Note that we can do a rough check using SAS PROC SURVEYMEANS, which will capture the variability due to the phase II sample.

```
data exer1201;
   input strat nh mh diabcount ;
   nondiab = mh - diabcount ;
   datalines;
1 241 96 86
2 113 45 17
3 174 35 29
4 472 47 8
;
```

```
data exer1201;
   set exer1201;
   do i = 1 to diabcount;
      sampwt = nh/mh*(100000/1000);
  diab = 1;
      output;
   end;
   do i = 1 to nondiab;
      sampwt = nh/mh*(100000/1000);
  diab = 0;
      output;
   end;

proc surveymeans data=exer1201 mean clm sum clsum;
    stratum strat;
   weight sampwt;
   var diab;
run;
```

This code gives $\hat{t} = 48310$ with SE 3059.0785. We can then add the second term in (12.7) to obtain

$$\hat{V}\left(\hat{t}_{\text{str}}^{(2)}\right) = 3059.0785^2 + \frac{100000^2}{999} 0.109282588 = 10451881.$$

This is a little bit larger than the estimate obtained above because we did not incorporate fpcs.

**12.3** Using the population and sample sizes of $N = 2130$, $n^{(1)} = 201$, and $n^{(2)} = 12$, the phase I weight is $w_i^{(1)} = 2130/201 = 10.597$ for every phase I unit. For the units in phase II, the phase II weight is $w_i^{(2)} = 201/12 = 16.75$. We have the following information and summary statistics from shorebirds.dat:

$$\hat{t}_x^{(1)} = \sum_{i \in \mathcal{S}^{(1)}} w_i^{(1)} x_i = 44284.93.$$

$$\hat{t}_x^{(2)} = \sum_{i \in \mathcal{S}^{(2)}} w_i^{(1)} w_i^{(2)} x_i = 34790,$$

and $\hat{t}_y^{(2)} = 43842.5$. Using (12.9),

$$\hat{t}_{yr}^{(2)} = \hat{t}_x^{(1)} \frac{\hat{t}_y^{(2)}}{\hat{t}_x^{(2)}} = 44284.93 \frac{43842.5}{34790} = 55808.$$

We estimate the variance using (12.11): we have $s_y^2 = 115.3561$, $s_e^2 = 7.453911$, and

$$\begin{aligned}\hat{V}(\hat{t}_{yr}^{(2)}) &= N^2\left(1-\frac{n^{(1)}}{N}\right)\frac{s_y^2}{n^{(1)}} + N^2\left(1-\frac{n^{(2)}}{n^{(1)}}\right)\frac{s_e^2}{n^{(2)}} \\ &= (2130)^2\left(1-\frac{201}{2130}\right)\frac{115.3561}{201} + (2130)^2\left(1-\frac{12}{201}\right)\frac{7.453911}{12} \\ &= 2358067 + 2649890 = 5007958,\end{aligned}$$

so the standard error is 2238.

Note that the paper by Bart and Earnst has some inconsistencies so it is possible that the data set constructed for this problem does not reflect the distribution of shorebirds in the region.

**12.4** (a) The phase 2 weight is 1049/60 =17.48 for stratum 1, 237/48 = 4.9375 for stratum 2, and 272/142 =1.915 for stratum 3.

(b) We use (12.5) to estimate

$$\hat{\bar{y}}_{\text{str}}^{(2)} = \sum_{h=1}^{H}\frac{n_h}{n}\bar{y}_h^{(2)} = 0.3030 + 0.1078 + 0.1426 = 0.5534.$$

Then, using (12.8) (which we may use since the fpc is negligible),

$$\begin{aligned}\hat{V}(\hat{\bar{y}}_{\text{str}}^{(2)}) &\approx \sum_{h=1}^{H}\frac{n_h-1}{n-1}\frac{n_h}{n}\frac{s_h^{2(2)}}{m_h} + \frac{1}{n-1}\sum_{h=1}^{H}\frac{n_h}{n}(\bar{y}_h^{(2)} - \hat{\bar{y}}_{\text{str}}^{(2)})^2 \\ &= 0.00186941 + 9.924\times 10^{-05} + 3.201\times 10^{-05} \\ &\quad + \frac{1}{1558}(0.007192 + 0.003654 + 0.012126) \\ &= 0.002015\end{aligned}$$

so the standard error is 0.045. Note that the second term adds little to the variability since the phase I sample size is large.

**12.5** (a) We use the final weights for the phase 2 sample to calculate the proportions in the table, and use (12.8) to find the standard error for each (given in parentheses behind the proportion).

| | | Case? | | |
|---|---|---|---|---|
| Proportion (SE) | | No | Yes | |
| Gender | Male | 0.2265 (0.0397) | 0.1496 (0.0312) | 0.3761 (0.0444) |
| | Female | 0.2164 (0.0399) | 0.4075 (0.0426) | 0.6239 (0.0444) |
| | | 0.4430 (0.0449) | 0.5570 (0.0449) | |

(b) We can calculate the Rao-Scott correction for a test statistic based on a sample of size $n = 1558$. Then, (10.2) gives

$$X^2 = n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}} = (1558)0.062035726 = 96.65$$

To find the design effect, we divide the estimated variance from part (a) by the variance that would have been obtained if an SRS of size 1558 had been selected, namely $\hat{p}(1-\hat{p})/1558$. We obtain the following table:

| Design effect | | Case? No | Case? Yes | |
|---|---|---|---|---|
| Gender | Male | 13.987 | 11.929 | 13.072 |
| | Female | 14.661 | 11.708 | 13.072 |
| | | 12.715 | 12.715 | |

Using (10.9),

$$E[X^2] \approx \sum_{i=1}^{r}\sum_{j=1}^{c}(1-p_{ij})d_{ij} - \sum_{i=1}^{r}(1-p_{i+})d_i^R - \sum_{j=1}^{c}(1-p_{+j})d_j^C = 13.601.$$

Then, from Section 10.2.2, $X^2/E[X^2]$ approximately follows a $\chi^2_1$ distribution if the null hypothesis is true. We calculate $X^2/E[X^2] = 96.65/13.601 = 7.1$, so the $p$-value is approximately 0.008. There is evidence against the null hypothesis of independence.

Note that we could equally well have calculated $X^2$ as $m\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\hat{p}_{ij}-\hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}}$ and calculated the variance under an SRS as $\hat{p}(1-\hat{p})/m$; this gives the same result.

Since the phase I sample size is relatively large, and the second term in (12.8) is small relative to the first term, we can use SAS PROC SURVEYFREQ to obtain an approximate check on our results.

```
data exer1205;
   input strat nh mh gender $ case count ;
   datalines;
1 1049 60 m 0 16
1 1049 60 m 1 8
1 1049 60 f 0 17
1 1049 60 f 1 19
2 237 48 m 0 9
2 237 48 m 1 8
2 237 48 f 0 5
2 237 48 f 1 26
```

```
3 272 142 m 0 15
3 272 142 m 1 28
3 272 142 f 0 8
3 272 142 f 1 91
;

data exer1205;
   set exer1205;
   do i = 1 to count;
      sampwt = nh/mh;
      output;
   end;

proc surveyfreq data=exer1205;
   stratum strat;
   weight sampwt;
   tables gender*case / chisq deff;
run;
```

This code treats the phase I sample as a population, so it underestimates the variance slightly. But since in this case $n$ is large, the results are very close. SAS calculates the Rao-Scott chi-square statistic as 7.01, and $p$-value as 0.008.

**12.9** We estimate $W_h$ by $n_h/n$, and estimate $S_h^2$ by $s_h^{2(2)}$. Then, using (12.17), we have

| Stratum | $\hat{W}_h$ | $\hat{S}_h^2$ | $\hat{W}_h\hat{S}_h^2$ | $\nu_n$ |
|---|---|---|---|---|
| Yes | 0.3658 | 0.1995 | 0.0730 | 0.40 |
| No | 0.3895 | 0.1313 | 0.0511 | 0.32 |
| Not available | 0.2447 | 0.2437 | 0.0596 | 0.44 |
| Total | 1.0000 | | 0.1837 | |

We estimate $S^2$ using

$$(n-1)\hat{S}^2 = \sum_{h=1}^{H}(n_h-1)\hat{S}_h^2 + \sum_{h=1}^{H} n_h(\hat{p}_h-\hat{p})^2,$$

which gives $\hat{S}^2 = 0.2468$.

This allocation takes many more observations in the "Yes" and "No" strata than did the allocation that was used. Proportional allocation would have $\nu_1 = \nu_2 = \nu_3$.

**12.10** From property 5 in Section A.4,

$$V(\hat{t}_y^{(2)}) = V(\hat{t}_y^{(1)}) + E(V[\hat{t}_y^{(2)} \mid \mathbf{Z}]).$$

Because the phase I sample is an SRS,

$$V(\hat{t}_y^{(1)}) = N^2 \left(1 - \frac{n^{(1)}}{N}\right) \frac{S_y^2}{n^{(1)}}.$$

Because the phase II sample is also an SRS,

$$P(D_i = 1 \mid Z_i = 1) = n^{(2)} \Big/ n^{(1)},$$

$$P(D_i D_j = 1 \mid Z_i Z_j = 1) = \frac{n^{(2)}(n^{(2)} - 1)}{n^{(1)}(n^{(1)} - 1)} \text{ for } j \neq i,$$

$$w_i^{(1)} = \frac{N}{n^{(1)}},$$

and

$$w_i^{(2)} = \frac{n^{(1)}}{n^{(2)}} Z_i.$$

In addition,

$$P(Z_i = 1) = \frac{n^{(1)}}{N}$$

and

$$P(Z_i Z_j = 1) = \frac{n^{(1)}(n^{(1)} - 1)}{N(N-1)}.$$

Thus, using (12.1) to write $\hat{t}_y^{(2)}$,

$$\begin{aligned}
&V[\hat{t}_y^{(2)} \mid \mathbf{Z}] \\
&= V\left[\sum_{i=1}^{N} Z_i D_i \frac{N}{n^{(1)}} \frac{n^{(1)}}{n^{(2)}} y_i \mid \mathbf{Z}\right] \\
&= \left(\frac{N}{n^{(2)}}\right)^2 E\left[\sum_{i=1}^{N}\sum_{k=1}^{N} Z_i Z_k D_i D_k y_i y_k \mid \mathbf{Z}\right] - [\hat{t}_y^{(1)}]^2 \\
&= \left(\frac{N}{n^{(2)}}\right)^2 \sum_{i=1}^{N} Z_i y_i^2 \frac{n^{(2)}}{n^{(1)}} + \left(\frac{N}{n^{(2)}}\right)^2 \sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} Z_i Z_j y_i y_j \frac{n^{(2)}(n^{(2)} - 1)}{n^{(1)}(n^{(1)} - 1)} - [\hat{t}_y^{(1)}]^2
\end{aligned}$$

$$
\begin{aligned}
E(V[\hat{t}_y^{(2)} \mid \mathbf{Z}]) &= \left(\frac{N}{n^{(2)}}\right)^2 \sum_{i=1}^{N} \frac{n^{(1)}}{N} y_i^2 \frac{n^{(2)}}{n^{(1)}} \\
&\quad + \left(\frac{N}{n^{(2)}}\right)^2 \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \frac{n^{(1)}(n^{(1)}-1)}{N(N-1)} y_i y_j \frac{n^{(2)}(n^{(2)}-1)}{n^{(1)}(n^{(1)}-1)} \\
&\quad - V[\hat{t}_y^{(1)}] - (E[\hat{t}_y^{(1)}])^2 \\
&= \frac{N}{n^{(2)}} \sum_{i=1}^{N} y_i^2 + \frac{N(n^{(2)}-1)}{n^{(2)}(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} y_i y_j \\
&\quad - V[\hat{t}y^{(1)}] - t_y^2 \\
&= N^2 \left(1 - \frac{n^{(2)}}{N}\right) \frac{S_y^2}{n^{(2)}} - V[\hat{t}_y^{(1)}].
\end{aligned}
$$

Thus,

$$
V[\hat{t}_y^{(2)}] = N^2 \left(1 - \frac{n^{(2)}}{N}\right) \frac{S_y^2}{n^{(2)}}.
$$

**12.11** Conditioning on the phase I units,

$$
\begin{aligned}
E[\hat{V}(\hat{t}_{\text{str}}^{(2)}) | \mathbf{Z}] &= N(N-1) \sum_{h=1}^{H} \left(\frac{n_h - 1}{n-1} - \frac{m_h - 1}{N-1}\right) \frac{n_h}{n} E\left[\frac{s_h^{2(2)}}{m_h} \Big| \mathbf{Z}\right] \\
&\quad + \frac{N^2}{n-1} \left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} E\left[\frac{n_h}{n} (\bar{y}_h^{(2)} - \hat{\bar{y}}_{\text{str}}^{(2)})^2 \Big| \mathbf{Z}\right].
\end{aligned}
$$

Now

$$
E\left[\frac{s_h^{2(2)}}{m_h} \Big| \mathbf{Z}\right] = \frac{s_h^{2(1)}}{m_h}
$$

and

$$
\begin{aligned}
&\sum_{h=1}^{H} E\left[\frac{n_h}{n} (\bar{y}_h^{(2)} - \hat{\bar{y}}_{\text{str}}^{(2)})^2 \Big| \mathbf{Z}\right] \\
&= E\left[\sum_{h=1}^{H} \frac{n_h}{n} (\bar{y}_h^{(2)})^2 - (\hat{\bar{y}}_{\text{str}}^{(2)})^2 \Big| \mathbf{Z}\right] \\
&= \sum_{h=1}^{H} \frac{n_h}{n} \left[\frac{s_h^{2(1)}}{m_h} \left(1 - \frac{m_h}{n_h}\right) + (\bar{y}_h^{(1)})^2\right] \\
&\quad - \sum_{h=1}^{H} \left(\frac{n_h}{n}\right)^2 \left(1 - \frac{m_h}{n_h}\right) \frac{s_h^{2(1)}}{m_h} - \left[\sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h^{(1)}\right]^2 \\
&= \sum_{h=1}^{H} \frac{n_h}{n} \left(1 - \frac{n_h}{n}\right) \left(1 - \frac{m_h}{n_h}\right) \frac{s_h^{2(1)}}{m_h} + \frac{1}{n} \left[(n-1) s_y^{2(1)} - \sum_{h=1}^{H} (n_h - 1) s_h^{2(1)}\right];
\end{aligned}
$$

the last equality follows by applying the sums of squares identity

$$(n-1)s_y^{2(1)} = \sum_{h=1}^{H}(n_h-1)s_h^{2(1)} + \sum_{h=1}^{H} n_h\left(\bar{y}_h^{(1)} - \sum_{k=1}^{H}\frac{n_k}{n}\bar{y}_k^{(1)}\right)^2$$

to the phase I sample.

Plugging in to the first equation, we have

$$\begin{aligned} E[\hat{V}(\hat{t}_{\text{str}}^{(2)})|\mathbf{Z}] &= \frac{N^2}{n}\left(1-\frac{n}{N}\right)s_y^{2(1)} \\ &\quad +N^2\sum_{h=1}^{H}\frac{s_h^{2(1)}}{m_h}\left\{\frac{N-1}{N}\left(\frac{n_h-1}{n-1}-\frac{m_h-1}{N-1}\right)\frac{n_h}{n}\right. \\ &\quad \left.+\frac{1}{n-1}\left(1-\frac{n}{N}\right)\left[\frac{n_h}{n}\left(1-\frac{n_h}{n}\right)\left(1-\frac{m_h}{n_h}\right)-\frac{m_h(n_h-1)}{n}\right]\right\} \\ &= \frac{N^2}{n}\left(1-\frac{n}{N}\right)s_y^{2(1)} + N^2\sum_{h=1}^{H}\frac{s_h^{2(1)}}{m_h}\left(\frac{n_h}{n}\right)^2\left(1-\frac{m_h}{n_h}\right) \end{aligned}$$

(The last equality follows after a lot of algebra.) Since $E[s_y^{2(1)}] = S_y^2$, the unbiasedness is shown.

**12.12** (a) Equation (A.9) implies these results.

(b) From the solution to Exercise 12.10,

$$\begin{aligned} V(\hat{t}_{yr}^{(2)}) &= V[\hat{t}_y^{(1)}] + E[V(\hat{t}_d^{(2)} \mid \mathbf{Z})], \\ V[\hat{t}_y^{(1)}] &= N^2\left(1-\frac{n^{(1)}}{N}\right)\frac{S_y^2}{n^{(1)}}, \end{aligned}$$

and

$$\begin{aligned} E[V(\hat{t}_d^{(2)} \mid \mathbf{Z})] &= N^2\left(1-\frac{n^{(2)}}{N}\right)\frac{S_d^2}{n^{(2)}} - V[\hat{t}_d^{(1)}] \\ &= N^2\left(1-\frac{n^{(2)}}{N}\right)\frac{S_d^2}{n^{(2)}} - N^2\left(1-\frac{n^{(1)}}{N}\right)\frac{S_d^2}{n^{(1)}} \\ &= NS_d^2\left[\frac{N-n^{(2)}}{n^{(2)}} - \frac{N-n^{(1)}}{n^{(1)}}\right] \\ &= N^2\left(1-\frac{n^{(2)}}{n^{(1)}}\right)\frac{S_d^2}{n^{(2)}}. \end{aligned}$$

(c) Follows because $s_y^2$ and $s_e^2$ estimate $S_y^2$ and $S_d^2$, respectively.

**12.13** Using the hint,

$$\begin{aligned} S_y^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{y}_U)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N}(y_i-Bx_i+Bx_i-B\bar{x}_U)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left[(y_i-Bx_i)^2+B^2(x_i-\bar{x}_U)^2+2(y_i-Bx_i)B(x_i-\bar{x}_U)\right] \\ &= S_d^2+B^2S_x^2+2BS_{xd}. \end{aligned}$$

Then,

$$\begin{aligned} V(\hat{t}_{yr}^{(2)}) &\approx N^2\left(1-\frac{n^{(1)}}{N}\right)\frac{S_y^2}{n^{(1)}}+N^2\left(1-\frac{n^{(2)}}{n^{(1)}}\right)\frac{S_d^2}{n^{(2)}} \\ &= N^2\left(1-\frac{n^{(1)}}{N}\right)\frac{2BS_{xd}+B^2S_x^2+S_d^2}{n^{(1)}}+N^2\left(1-\frac{n^{(2)}}{n^{(1)}}\right)\frac{S_d^2}{n^{(2)}} \\ &= N^2\left(1-\frac{n^{(1)}}{N}\right)\frac{2BS_{xd}+B^2S_x^2}{n^{(1)}}+N^2\left(1-\frac{n^{(2)}}{N}\right)\frac{S_d^2}{n^{(2)}}. \end{aligned}$$

**12.14** (a)

$$\begin{aligned} z_i^{(1)} &= \frac{\partial\hat{t}_{yr}^{(2)}}{\partial w_i^{(1)}} \\ &= x_i\frac{\hat{t}_y^{(2)}}{\hat{t}_x^{(2)}} \end{aligned}$$

and

$$\begin{aligned} z_i^{(2)} &= \frac{\partial\hat{t}_{yr}^{(2)}}{\partial w_i} \\ &= \hat{t}_x^{(1)}\left[\frac{y_i}{\hat{t}_x^{(2)}}-\frac{x_i}{\hat{t}_x^{(2)}}\frac{\hat{t}_y^{(2)}}{\hat{t}_x^{(2)}}\right] \\ &= \frac{\hat{t}_x^{(1)}}{\hat{t}_x^{(2)}}\left[y_i-x_i\hat{t}_x^{(2)}\hat{B}^{(2)}\right] \end{aligned}$$

Thus,

$$\begin{aligned} \hat{V}_{\mathrm{DR}}(\hat{t}_{yr}^{(2)}) &= \hat{V}\left(\sum_{i\in\mathcal{S}^{(1)}}w_i^{(1)}z_i^{(1)}+\sum_{i\in\mathcal{S}^{(2)}}w_iz_i^{(2)}\right) \\ &= \hat{V}\left(\sum_{i\in\mathcal{S}^{(1)}}w_i^{(1)}x_i\hat{B}^{(2)}+\sum_{i\in\mathcal{S}^{(2)}}w_i\frac{\hat{t}_x^{(1)}}{\hat{t}_x^{(2)}}\left[y_i-x_i\hat{t}_x^{(2)}\hat{B}^{(2)}\right]\right) \end{aligned}$$

**12.16** Note that

$$\begin{aligned}
E[\hat{V}_{HT}(\hat{t}_y^{(1)}) \mid \mathbf{Z}] &= E\left[\sum_{i\in\mathcal{S}^{(2)}}\sum_{k\in\mathcal{S}^{(2)}} \frac{\pi_{ik}^{(1)} - \pi_i^{(1)}\pi_k^{(1)}}{\pi_{ik}^{(1)}\pi_{ik}^{(2)}} \frac{y_i}{\pi_i^{(1)}} \frac{y_k}{\pi_k^{(1)}} \mid \mathbf{Z}\right] \\
&= E\left[\sum_{i\in\mathcal{S}^{(1)}}\sum_{k\in\mathcal{S}^{(1)}} D_i D_k \frac{\pi_{ik}^{(1)} - \pi_i^{(1)}\pi_k^{(1)}}{\pi_{ik}^{(1)}\pi_{ik}^{(2)}} \frac{y_i}{\pi_i^{(1)}} \frac{y_k}{\pi_k^{(1)}} \mid \mathbf{Z}\right] \\
&= E\left[\sum_{i\in\mathcal{S}^{(1)}}\sum_{k\in\mathcal{S}^{(1)}} \frac{\pi_{ik}^{(1)} - \pi_i^{(1)}\pi_k^{(1)}}{\pi_{ik}^{(1)}} \frac{y_i}{\pi_i^{(1)}} \frac{y_k}{\pi_k^{(1)}} \mid \mathbf{Z}\right] \\
&= \hat{V}_{HT}^{(1)}(\hat{t}_y^{(1)})
\end{aligned}$$

**12.17** (a) Since $m_h = \nu_h n_h$ and $\nu_h$ is known,

$$\begin{aligned}
E[m_h] &= \nu_h E[n_h] \\
&= \nu_h E\left[\sum_{i=1}^{N} Z_i x_{ih}\right] \\
&= \nu_h \sum_{i=1}^{N} \frac{n}{N} x_{ih} \\
&= n\nu_h W_h.
\end{aligned}$$

Thus,

$$E[C] = cn + n\sum_{h=1}^{H} c_h \nu_h W_h.$$

(b) Using the constraint, set

$$n = \frac{E[C]}{c + \sum_{h=1}^{H} c_h \nu_h W_h}.$$

Then

$$\begin{aligned}
V(\hat{\bar{y}}_{\text{str}}^{(2)}) &= S^2 \left[\frac{c + \sum_{h=1}^{H} c_h \nu_h W_h}{E[C]} - \frac{1}{N}\right] \\
&\quad + \frac{c + \sum_{h=1}^{H} c_h \nu_h W_h}{E[C]} \sum_{h=1}^{H} W_h S_h^2 \left(\frac{1}{\nu_h} - 1\right)
\end{aligned}$$

and

$$\frac{\partial V(\hat{\bar{y}}_{\text{str}}^{(2)})}{\partial \nu_k} = \frac{S^2}{E[C]} c_k W_k + \frac{1}{E[C]} \left[ c_k W_k \sum_{h=1}^{H} W_h S_h^2 \left( \frac{1}{\nu_h} - 1 \right) - \frac{W_k S_k^2}{\nu_k^2} \left( c + \sum_{h=1}^{H} c_h \nu_h W_h \right) \right].$$

Setting the derivatives equal to 0, we have

$$S^2 c_k \nu_k W_k + c_k \nu_k W_k \sum_{h=1}^{H} W_h S_h^2 \left( \frac{1}{\nu_h} - 1 \right) - \frac{W_k S_k^2}{\nu_k} \left( c + \sum_{h=1}^{H} c_h \nu_h W_h \right) = 0$$

for $k = 1, \ldots, H$. Thus

$$\begin{aligned} 0 &= \sum_{k=1}^{H} c_k \nu_k W_k \left[ S^2 + \sum_{h=1}^{H} W_h S_h^2 \left( \frac{1}{\nu_h} - 1 \right) \right] - \sum_{k=1}^{H} \frac{W_k S_k^2}{\nu_k} \left( c + \sum_{h=1}^{H} c_h \nu_h W_h \right) \\ &= \sum_{k=1}^{H} c_k \nu_k W_k \left[ S^2 - \sum_{h=1}^{H} W_h S_h^2 \right] - c \sum_{k=1}^{H} \frac{W_k S_k^2}{\nu_k} \end{aligned}$$

and

$$\sum_{h=1}^{H} \frac{W_h S_h^2}{\nu_h} = \frac{1}{c} \left( S^2 - \sum_{h=1}^{H} W_h S_h^2 \right) \sum_{k=1}^{H} c_k \nu_k W_k.$$

Substituting into $(*)$, we have

$$\begin{aligned} 0 &= c_k \nu_k W_k \left[ S^2 - \sum_{h=1}^{H} W_h S_h^2 + \frac{1}{c} \left( S^2 - \sum_{h=1}^{H} W_h S_h^2 \right) \sum_{h=1}^{H} c_h \nu_h W_h \right] \\ &\quad - \frac{W_k S_k^2}{\nu_k} \left( c + \sum_{h=1}^{H} c_h \nu_h W_h \right) \\ &= c_k \nu_k W_k \left( S^2 - \sum_{h=1}^{H} W_h S_h^2 \right) \frac{1}{c} \left( c + \sum_{h=1}^{H} c_h \nu_h W_h \right) \\ &\quad - \frac{W_k S_k^2}{\nu_k} \left( c + \sum_{h=1}^{H} c_h \nu_h W_h \right), \end{aligned}$$

which implies that

$$\frac{c_k \nu_k^2}{c} \left( S^2 - \sum_{h=1}^{H} W_h S_h^2 \right) = S_k^2,$$

and, consequently,

$$\nu_k = \sqrt{\frac{c S_k^2}{c_k (S^2 - \sum_{h=1}^{H} W_h S_h^2)}}.$$

(c) To meet the expected cost constraint with the optimal allocation, set

$$n = \frac{E[C]}{c + \sum_{h=1}^{H} c_h \nu_h^* W_h},$$

with

$$\nu_h^* = \sqrt{\frac{cS_h^2}{c_h(S^2 - \sum_{j=1}^{H} W_j S_j^2)}}.$$

**12.18** Let $A = S_y^2 - \sum_{h=1}^{H} W_h S_h^2$. Then, from, (12.17),

$$\nu_{h,opt} = \sqrt{\frac{c^{(1)} S_h^2}{c_h \left( S^2 - \sum_{j=1}^{H} W_j S_j^2 \right)}} = \sqrt{\frac{c^{(1)} S_h^2}{c_h A}}$$

and

$$n_{opt}^{(1)} = \frac{C^*}{c^{(1)} + \sum_{h=1}^{H} c_h W_h \nu_{h,opt}} = \frac{C^*}{c^{(1)} + \sum_{h=1}^{H} W_h S_h \sqrt{c_h} \sqrt{\frac{c^{(1)}}{A}}}.$$

Then,

$$\begin{aligned}
V_{opt}(\hat{\bar{y}}^{(2)}_{\text{str}}) &= \frac{S_y^2}{n^{(1)}_{opt}} - \frac{S_y^2}{N} + \frac{1}{n^{(1)}_{opt}} \sum_{h=1}^{H} W_h S_h^2 \left( \frac{1}{\nu_{h,opt}} - 1 \right) \\
&= \frac{S_y^2}{n^{(1)}_{opt}} - \frac{S_y^2}{N} + \frac{1}{n^{(1)}_{opt}} \left( \sum_{h=1}^{H} W_h S_h \sqrt{c_h} \sqrt{\frac{A}{c^{(1)}}} - \sum_{h=1}^{H} W_h S_h^2 \right) \\
&= \frac{S_y^2}{n^{(1)}_{opt}} - \frac{S_y^2}{N} + \frac{1}{n^{(1)}_{opt}} \left( \sum_{h=1}^{H} W_h S_h \sqrt{c_h} \sqrt{\frac{A}{c^{(1)}}} + A - S_y^2 \right) \\
&= \frac{1}{n^{(1)}_{opt}} \left( \sum_{h=1}^{H} W_h S_h \sqrt{c_h} \sqrt{\frac{A}{c^{(1)}}} + A \right) - \frac{S_y^2}{N} \\
&= \frac{1}{C^*} \left( c^{(1)} + \sum_{h=1}^{H} W_h S_h \sqrt{c_h} \sqrt{\frac{c^{(1)}}{A}} \right) \left( \sum_{h=1}^{H} W_h S_h \sqrt{c_h} \sqrt{\frac{A}{c^{(1)}}} + A \right) - \frac{S_y^2}{N} \\
&= \frac{1}{C^*} \left[ \sqrt{c^{(1)} A} \sum_{h=1}^{H} W_h S_h \sqrt{c_h} + \left( \sum_{h=1}^{H} W_h S_h \sqrt{c_h} \right)^2 \right. \\
&\qquad \left. + A c^{(1)} + \sqrt{c^{(1)} A} \sum_{h=1}^{H} W_h S_h \sqrt{c_h} \right] - \frac{S_y^2}{N} \\
&= \frac{1}{C^*} \left[ \sum_{h=1}^{H} W_h S_h \sqrt{c_h} + \sqrt{c^{(1)}} \sqrt{S^2 - \sum_{h=1}^{H} W_h S_h^2} \right]^2 - \frac{S_y^2}{N}.
\end{aligned}$$

**12.19** The easiest way to solve this optimization problem is to use Lagrange multipliers. Using the variance in (12.10), the function we wish to minimize is

$$g(n^{(1)}, n^{(2)}, \lambda) = \left( \frac{1}{n^{(1)}} - \frac{1}{N} \right) S_y^2 + \left( \frac{1}{n^{(2)}} - \frac{1}{n^{(1)}} \right) S_d^2 - \lambda \left[ C - c^{(1)} n^{(1)} - c^{(2)} n^{(2)} \right].$$

Setting the partial derivatives with respect to $n^{(1)}$, $n^{(2)}$, and $\lambda$ equal to 0, we have

$$\frac{\partial g}{\partial n^{(1)}} = -\frac{S_y^2}{\left[ n^{(1)} \right]^2} + \frac{S_d^2}{\left[ n^{(1)} \right]^2} + \lambda c^{(1)} = 0,$$

$$\frac{\partial g}{\partial n^{(2)}} = -\frac{S_d^2}{\left[ n^{(2)} \right]^2} + \lambda c^{(2)} = 0,$$

and

$$\frac{\partial g}{\partial \lambda} = -\left[ C - c^{(1)} n^{(1)} - c^{(2)} n^{(2)} \right] = 0.$$

Consequently, using the first two equations, we have

$$\left[ n^{(1)} \right]^2 = \frac{S_y^2 - S_d^2}{\lambda c^{(1)}}$$

and

$$\left[n^{(2)}\right]^2 = \frac{S_d^2}{\lambda c^{(2)}}.$$

Taking the ratios gives

$$\left(\frac{n^{(2)}}{n^{(1)}}\right)^2 = \frac{c^{(1)} S_d^2}{c^{(2)}(S_y^2 - S_d^2)},$$

which proves the result.

**12.20** (a) These results follow directly from the contingency table. For example,

$$\frac{N_1}{N} p_1 = \frac{C_{21}}{N} = \frac{C_{21}}{C_{2+}} \frac{C_{2+}}{N} = \left(1 - \frac{C_{22}}{C_{2+}}\right) \frac{C_{2+}}{N} = (1 - S_2)p.$$

$$\frac{N_2}{N} p_2 = \frac{C_{22}}{N} = \frac{C_{22}}{C_{2+}} \frac{C_{2+}}{N} = S_2 p.$$

The other results are shown similarly.

(b) From (12.19),

$$\frac{V_{opt}(\hat{p}_{\text{str}}^{(2)})}{V_{SRS}(\hat{p})} \approx \left[\sum_{h=1}^{2} W_h \frac{S_h}{S_y} + \sqrt{\frac{c^{(1)}}{c^{(2)}}} \sqrt{\frac{S_y^2 - \sum_{h=1}^{2} W_h S_h^2}{S_y^2}}\right]^2.$$

Using the results from part (a),

$$\begin{aligned}
\sum_{h=1}^{2} W_h \frac{S_h}{S_y} &= \sqrt{\left(\frac{N_1}{N}\right)^2 \frac{p_1(1-p_1)}{p(1-p)}} + \sqrt{\left(\frac{N_2}{N}\right)^2 \frac{p_2(1-p_2)}{p(1-p)}} \\
&= \sqrt{\frac{(1-S_2)p(1-p)S_1}{p(1-p)}} + \sqrt{\frac{pS_2(1-p)(1-S_1)}{p(1-p)}} \\
&= \sqrt{(1-S_2)S_1} + \sqrt{S_2(1-S_1)}.
\end{aligned}$$

For the second term, note that

$$\sum_{i=1}^{N} (x_i - \bar{x}_U)(y_i - \bar{y}_U) = \sum_{i=1}^{N} x_i(y_i - p) = C_{22} - NpW_2,$$

$$\sum_{i=1}^{N} (x_i - \bar{x}_U)^2 = \sum_{i=1}^{N} (x_i - W_2)^2 = \sum_{i=1}^{N} x_i - NW_2^2 = NW_1W_2,$$

and

$$\sum_{i=1}^{N} (y_i - \bar{y}_U)^2 = \sum_{i=1}^{N} (y_i - p)^2 = \sum_{i=1}^{N} y_i - Np^2 = Np(1-p),$$

Consequently,

$$S_y R = \frac{C_{22} - NpW_2}{N\sqrt{W_1W_2}} = \frac{p(S_2 - W_2)}{\sqrt{W_1W_2}}.$$

$$\begin{aligned} S_y^2 - \sum_{h=1}^{2} W_h S_h^2 &= p(1-p) - W_1 p_1(1-p_1) - W_2 p_2(1-p_2) \\ &= W_1 p_1^2 + W_2 p_2^2 - p^2 \\ &= \frac{(1-S_2)^2 p^2}{W_1} + \frac{S_2^2 p^2}{W_2} - p^2 \\ &= \frac{p^2}{W_1 W_2}\left[W_2(1-S^2)^2 + W_1 S_2^2 - W_1 W_2\right] \\ &= \frac{p^2}{W_1 W_2}\left[W_2^2 - 2W_1 S_2 + S_2^2\right] \\ &= \frac{p^2}{W_1 W_2}\left[S_2 - W_2\right]^2 \\ &= S_y^2 R^2. \end{aligned}$$

(c) The following calculations were done in Excel.

| | Cost Ratio | | | | |
|---|---|---|---|---|---|
| $S_1$ | 0.0001 | 0.01 | 0.1 | 0.5 | 1 |
| 0.5 | 1.00 | 1.02 | 1.06 | 1.15 | 1.21 |
| 0.6 | 0.97 | 1.02 | 1.15 | 1.42 | 1.64 |
| 0.7 | 0.85 | 0.93 | 1.15 | 1.61 | 2.01 |
| 0.8 | 0.65 | 0.76 | 1.04 | 1.68 | 2.25 |
| 0.9 | 0.37 | 0.48 | 0.78 | 1.53 | 2.25 |
| 0.95 | 0.20 | 0.28 | 0.54 | 1.23 | 1.92 |

**12.21** Suppose that $\mathcal{S}$ is a subset of $m$ units from $U$, and suppose $\mathcal{S}$ has $m_h$ units from stratum $h$, for $h = 1, \ldots, H$. Let $Z_i = 1$ if unit $i$ is selected to be in the final sample and 0 otherwise; similarly, let $F_i = 1$ if unit $i$ is selected to be in the stratified sample and 0 otherwise, and let $D_i = 1$ if unit $i$ is selected to be in the subsample chosen from the stratified sample and 0 otherwise. Then the probability that $\mathcal{S}$ is chosen to be the sample is

$$P(\mathcal{S}) = P(Z_i = 1, i \in \mathcal{S}, \text{ and } Z_i = 0, i \notin \mathcal{S})$$

We can write $P(\mathcal{S})$ as

$$P(\mathcal{S}) = P(F_i = 1, i \in \mathcal{S}) P(D_i = 1, i \in \mathcal{S} \mid F_1, \ldots, F_N).$$

Then,

$$\begin{aligned} &P(F_i = 1, i \in \mathcal{S}) \\ &= \frac{\binom{m_1}{m_1}\binom{N_1 - m_1}{n_1 - m_1}\binom{m_2}{m_2}\binom{N_2 - m_2}{n_2 - m_2}\cdots\binom{m_H}{m_H}\binom{N_H - m_H}{n_H - m_H}}{\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_H}{n_H}} \\ &= \frac{\text{total number of stratified samples containing } \mathcal{S}}{\text{number of possible stratified samples}} \end{aligned}$$

Also,

$$\begin{aligned} P(D_i = 1, i \in \mathcal{S} \mid \mathbf{F}) &= P(M_h = m_h, h = 1, \ldots, H) P(D_i = 1, i \in \mathcal{S} | \mathbf{M} = \mathbf{m}) \\ &= \frac{\left( \begin{array}{c} N_1 \\ m_1 \end{array} \right) \cdots \left( \begin{array}{c} N_H \\ m_H \end{array} \right)}{\left( \begin{array}{c} N \\ m \end{array} \right)} \frac{1}{\left( \begin{array}{c} n_1 \\ m_1 \end{array} \right) \left( \begin{array}{c} n_2 \\ m_2 \end{array} \right) \cdots \left( \begin{array}{c} n_H \\ m_H \end{array} \right)}. \end{aligned}$$

Note that for each $h$,

$$\begin{aligned} & \frac{\left( \begin{array}{c} N_h - m_h \\ n_h - m_h \end{array} \right) \left( \begin{array}{c} N_h \\ m_h \end{array} \right)}{\left( \begin{array}{c} N_h \\ n_h \end{array} \right) \left( \begin{array}{c} n_h \\ m_h \end{array} \right)} \\ &= \frac{(N_h - m_h)!}{(n_h - m_h)!(N_h - n_h)!} \frac{N_h!}{m_h!(N_h - m_h)!} \frac{n_h!(N_h - n_h)!}{N_h!} \frac{m_h!(n_h - m_h)!}{n_h!} \\ &= 1. \end{aligned}$$

Consequently,

$$\begin{aligned} P(\mathcal{S}) &= P(F_i = 1, i \in \mathcal{S}) P(D_i = 1, i \in \mathcal{S} \mid F_1, \ldots, F_N) \\ &= \frac{1}{\left( \begin{array}{c} N \\ m \end{array} \right)} \prod_{h=1}^{H} \frac{\left( \begin{array}{c} N_h - m_h \\ n_h - m_h \end{array} \right) \left( \begin{array}{c} N_h \\ m_h \end{array} \right)}{\left( \begin{array}{c} N_h \\ n_h \end{array} \right) \left( \begin{array}{c} n_h \\ m_h \end{array} \right)} \\ &= \frac{1}{\left( \begin{array}{c} N \\ m \end{array} \right)}, \end{aligned}$$

so this procedure results in an SRS of size $m$.

# Chapter 13

# Estimating Population Size

**13.1** Students may answer this in several different ways. The maximum likelihood estimate is

$$\hat{N} = \frac{n_1 n_2}{m} = \frac{(500)(300)}{120} = 1250$$

with 95% CI (using likelihood ratio method) of [1116, 1422]. A bootstrap CI is [1103, 1456].

```
xmat <- cbind(c(1,1,0),c(1,0,1))
y <- c(120,380,180)
captureci(xmat,y)

bootout <- captureboot(converty(y,xmat[,1],xmat[,2]),
           crossprod(xmat[,1],y),nboot=999,nfunc=nmle)
```

**13.2** (a) The maximum likelihood estimate is

$$\hat{N} = \frac{n_1 n_2}{m} = \frac{(7)(12)}{4} = 21.$$

Chapman's estimate is

$$\tilde{N} = \frac{(n_1+1)(n_2+1)}{m+1} - 1 = \frac{(8)(13)}{5} - 1 = 19.8.$$

Because of the small sample sizes, we do not wish to employ a confidence interval that requires $\hat{N}$ or $\tilde{N}$ to be approximately normally distributed. Using the R function captureci gives $\hat{N} = 21$ with approximate 95% confidence interval [15.3, 47.2]. Alternatively, we could use the bootstrap to find an approximate confidence interval for $\tilde{N}$. (Theoretically, we could also use the bootstrap to find a confidence interval for $\hat{N}$ as well; in this data set, however, $m^*$ for resamples can be 0, so we only use the procedure with Chapman's estimator) The bootstrap gives a 95% confidence interval [12, 51] for $N$, using Chapman's estimator.

Here is the code from R:

```
xmat <- cbind(c(1,1,0),c(1,0,1))
y <- c(4,3,8)
captureci(xmat,y)

bootout <- captureboot(converty(y,xmat[,1],xmat[,2]),
           crossprod(xmat[,1],y),nboot=999,nfunc=nchapman)
```

(b) $\hat{N} = 27.6$ with approximate 95% confidence interval [24.1, 37.7].

(c) You are assuming that the two samples are independent. This means that fishers are equally likely to be captured on each occasion.

**13.3** We obtain $\hat{N} = 65.4$ with 95% CI [60.9, 74.4].

Here is the code from R:

```
xmat <- cbind(c(1,1,0),c(1,0,1))
y <- c(33,15,12)
captureci(xmat,y)
```

**13.4** (a) We treat the radio transmitter bears and feces sample bears as the two samples to obtain $\hat{N} = 483.8$ with 95% CI [413.7, 599.0].

```
xmat <- cbind(c(1,1,0),c(1,0,1))
y <- c(36,311-36,20)
captureci(xmat,y)
```

(b)$\hat{N} = 486.5$ with 95% CI [392.0, 646.4].

```
xmat <- cbind(c(1,1,0),c(1,0,1))
y <- c(28,239-28,57-28)
captureci(xmat,y)
```

(c)$\hat{N} = 450$ with 95% CI [427, 480].

```
xmat <- cbind(c(1,1,0),c(1,0,1))
y <- c(165, 311-165,239-165)
captureci(xmat,y)
```

**13.5** The model with all two-factor interactions has $G^2 = 3.3$, with 4 df. Comparing to a $\chi^2_4$ distribution gives a $p$-value 0.502. No simpler model appears to fit the data.

Using this model and function captureci in R, we estimate 3645 persons in the missing cell, with approximate 95% confidence interval [2804, 4725].

**13.6** (a)$\hat{N} = 336$, with 95% CI [288, 408]. $\tilde{N} = 333$ with 95% CI [273, 428]. The linearization-based standard errors are

$$\mathrm{SE}\,(\hat{N}) = \sqrt{\frac{n_1^2 n_2(n_2 - m)}{m^3}} = 37$$

and

$$\mathrm{SE}\,(\tilde{N}) = \sqrt{\frac{(n_1+1)(n_2+1)(n_1-m)(n_2-m)}{(m+1)^2(m+2)}} = 29.$$

```
xmat <- cbind(c(1,1,0),c(1,0,1))
y <- c(49,73,86)
captureci(xmat,y)

bootout <- captureboot(converty(y,xmat[,1],xmat[,2]),
           crossprod(xmat[,1],y),nboot=999,nfunc=nchapman)
```

(b) The following SAS code may be used to obtain estimates for the models.

```
data hep;
   input elist dlist tlist count;
   datalines;
0 0 1 63
0 1 0 55
0 1 1 18
1 0 0 69
1 0 1 17
1 1 0 21
1 1 1 28
;

proc means data=hep sum;
   var count;
   run;

proc print data=hep;
run;

proc catmod data=hep;
   weight count;
   model elist*dlist*tlist = _response_ /pred=freq ml=nr;
   loglin elist dlist tlist;
      /* Model of independent factors */
run;
```

```
proc genmod data=hep;
CLASS elist dlist tlist / param=effect;
MODEL count = elist dlist tlist / dist=poisson link=log type3;
run;

proc catmod data=hep;
   weight count;
   model elist*dlist*tlist = _response_ /pred=freq ml;
   loglin elist dlist tlist elist*dlist;
      /* Model with elist and dlist dependent */
run;

proc catmod data=hep;
   weight count;
   model elist*dlist*tlist = _response_ /pred=freq ml;
   loglin elist dlist tlist tlist*dlist;
      /* Model with tlist and dlist dependent */
run;

proc catmod data=hep;
   weight count;
   model elist*dlist*tlist = _response_ /pred=freq ml;
   loglin elist dlist tlist elist*tlist;
      /* Model with elist and tlist dependent */
run;

proc catmod data=hep;
   weight count;
   model elist*dlist*tlist = _response_ /pred=freq;
   loglin elist|dlist|tlist@2;
      /* Model with all 2-way intrxns */
run;
```

**13.7** (a) The assumption of independence of the two sources is probably met, at least approximately. The registry is from state and local health departments, while BDMP is from hospital data. Presumably, the health departments do not use hospital newborn discharge information when compiling their statistics. However, there might be a problem if congenital rubella syndrome is misclassified in both data sets, for instance, if both sources tend to miss cases.

We do not know how easily records were matched, but the paper said matching was not a problem.

The assumption of simple random sampling is probably not met. The BDMP was from a sample of hospitals, giving a cluster sample of records. In addition, selection of the hospitals for the BDMP was not random—hospitals were self-selected. It is

unclear how much the absence of simple random sampling in this source affects the results.

(b)

| Year | $\tilde{N}$ |
|---|---|
| 1970 | 244.33 |
| 1971 | 95 |
| 1972 | 48 |
| 1973 | 79.5 |
| 1974 | 44.5 |
| 1975 | 114 |
| 1976 | 41.67 |
| 1977 | 30.5 |
| 1978 | 62.33 |
| 1979 | 159 |
| 1980 | 31.5 |
| 1981 | 4 |
| 1982 | 35 |
| 1983 | 3 |
| 1984 | 3 |
| 1985 | 1 |

The sum of these estimates is 996.3333.

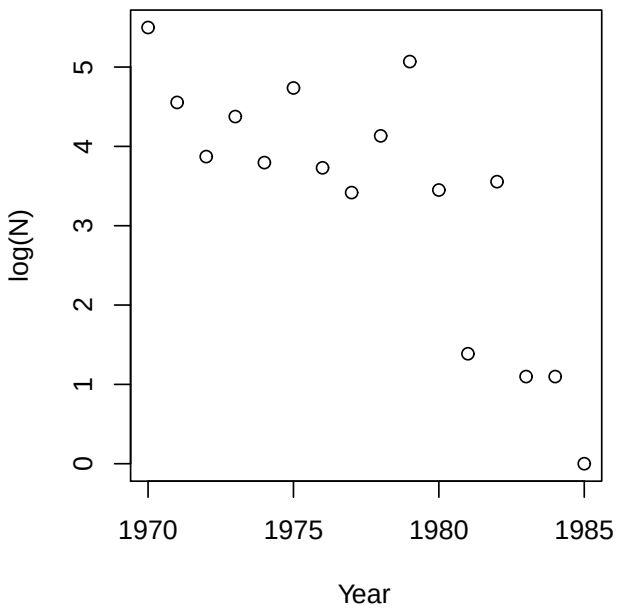
(c) Using the aggregated data, the total number of cases of congenital rubella syndrome between 1970 and 1985 is estimated to be

$$\tilde{N} = (263 + 1)(93 + 1)/(19 + 1) - 1 = 1239.8.$$

Equation (12.8) results in $\hat{V}(\tilde{N}) = 53343$, and in an approximate 95% confidence interval $1240 \pm 1.96\sqrt{(}53343) = [787, 1693]$. Using the bootstrap function gives a 95% confidence interval of [855, 1908].

The aggregated estimate should be more reliable—each yearly estimate is based on only a few cases and has large variability.

(d) Many methods could be used to assess whether the incidence of congenital rubella syndrome has changed, using these data. You could use a change-point test or divide the data into two groups and test whether the incidence is the same in both. The plot below shows the relation between year and $\log(\tilde{N})$, estimated in the table in part(b). The decrease after 1980 is apparent.




**13.8**

| Model | $G^2$ | df | $p$-value |
|---|---|---|---|
| Independence | 11.1 | 3 | 0.011 |
| 1*2 | 2.8 | 2 | 0.250 |
| 1*3 | 10.7 | 2 | 0.005 |
| 2*3 | 9.4 | 2 | 0.00 |

The model with interaction between sample 1 and sample 2 appears to fit well. Using that model, we estimate $\hat{N} = 2378$ with approximate 95% confidence interval [2142, 2664].

**13.9**

A positive interaction between presence in sample 1 and presence in sample 2 (as there is) suggests that some fish are "trap-happy"—they are susceptible to repeated trapping. An interaction between presence in sample 1 and presence in sample 3 might mean that the fin clipping makes it easier or harder to catch the fish with the net.

**13.10** (a) The maximum likelihood estimate is $\hat{N} = 73.1$ and Chapman's estimate is $\tilde{N} = 70.6$. A 95% confidence interval for $N$, using $\hat{N}$ and the function captureci, is [55.4, 124.1]. Another approximate 95% confidence interval for $N$, using Chapman's estimate and the bootstrap, is [52.7, 127.8].

**13.12** For the data in Example 13.1, $\hat{p} = \dfrac{20}{100} = \dfrac{1}{5}$ and a 95% confidence interval for $p$ is

$$0.2 \pm 1.96\sqrt{\frac{(0.2)(0.8)}{100}} = [0.12, 0.28].$$

The confidence limits $L(\hat{p})$ and $U(\hat{p})$ satisfy

$$P\{L(\hat{p}) \le p \le U(\hat{p})\} = 0.95.$$

Because $N = n_1/p$, we can write

$$P\{n_1/U(\hat{p}) \le N \le n_1/L(\hat{p})\} = 0.95.$$

Thus, a 95% confidence interval for $N$ is $[n_1/U(\hat{p}), n_1/L(\hat{p})]$; for these data, the interval is [718, 1645]. The interval is comparable to those from the inverted chi-square tests and bootstrap; like them, it is not symmetric.

**13.13** Note that

$$\begin{aligned}
\mathcal{L}(N-1|n_1,n_2) &= \frac{\binom{n_1}{m}\binom{N-1-n_1}{n_2-m}}{\binom{N-1}{n_2}} \\
&= \frac{\binom{n_1}{m}\binom{N-n_1}{n_2-m}\dfrac{N-n_1-(n_2-m)}{N-n_1}}{\binom{N}{n_2}\dfrac{N-n_2}{N}} \\
&= \mathcal{L}(N|n_1,n_2)\frac{N-n_1-n_2+m}{N-n_1}\frac{N}{N-n_2}.
\end{aligned}$$

Thus, if $N > n_1$ and $N > n_2$,

$$\begin{aligned}
\mathcal{L}(N) \ge \mathcal{L}(N-1) \qquad &\text{iff} \quad \frac{N-n_1-n_2+m}{N-n_1}\frac{N}{N-n_2} \le 1 \\
&\text{iff} \quad mN \le n_1 n_2.
\end{aligned}$$

Take $\hat{N}$ to be the integer part of $n_1 n_2/m$. Then for any integer $k \le \hat{N}$,

$$mk \le m\frac{n_1 n_2}{m} = n_1 n_2,$$

so $\mathcal{L}(k) \ge \mathcal{L}(k-1)$ for $k \ge \hat{N}$. Similarly, for $k > \hat{N}$ ($k$ integer),

$$mk > m\frac{n_1 n_2}{m} > n_1 n_2,$$

so $\mathcal{L}(k) < \mathcal{L}(k-1)$ for $k > \hat{N}$. Thus $\hat{N}$ is the maximum likelihood estimator of $N$.

**13.14** (a) Setting the derivative equal to zero, we have

$$m(\hat{N} - n_1) = (n_2 - m)n_1,$$

or

$$\hat{N} = \frac{n_2 n_1}{m}.$$

Note that the second derivative is

$$\begin{aligned}\frac{d^2 \log \mathcal{L}(N)}{dN^2} &= \frac{m}{N^2} - \frac{(n_2 - m)n_1(2N - n_1)}{N^2(N - n_1)^2} \\ &= \frac{mN^2 - 2Nn_1n_2 + n_1^2 n_2}{N^2(N - n_1)^2},\end{aligned}$$

which is negative when evaluated at $\hat{N}$.

(b) Noting that $E[m] = n_1 n_2/N$, the Fisher information is

$$\begin{aligned}\mathcal{I}(N) &= -E_m\left[\frac{\partial^2}{\partial N^2} \log \mathcal{L}(N/m, n_1, n_2)\right] \\ &= -E_m\left[\frac{mN^2 - 2Nn_1n_2 + n_1^2 n_2}{N^2(N - n_1)^2}\right] \\ &= -\frac{Nn_1n_2 - 2Nn_1n_2 + n_1^2 n_2}{N^2(N - n_1)^2} \\ &= \frac{n_1 n_2}{N^2(N - n_1)}.\end{aligned}$$

Consequently, the asymptotic variance of $\hat{N}$ is

$$V(\hat{N}) = \frac{1}{\mathcal{I}(N)} = \frac{N^2(N - n_1)}{n_1 n_2}.$$

**13.15** Substituting $C - n_1$ for $n_2$ in the variance, we have

$$g(n_1) = V(\hat{N}) = \frac{N^2(N - n_1)}{n_1(C - n_1)}.$$

Taking the derivative,

$$\begin{aligned}\frac{dg}{dn_1} &= -\frac{N^2}{n_1(C - n_1)} - \frac{N^2(N - n_1)(C - 2n_1)}{n_1^2(C - n_1)^2} \\ &= -\frac{N^2}{n_1^2(C - n_1)^2}[n_1(C - n_1) + (N - n_1)(C - 2n_1)].\end{aligned}$$

Equating the derivative to 0 gives

$$n_1^2 - 2Nn_1 + NC = 0,$$

or

$$n_1 = \frac{2N \pm \sqrt{4N^2 - 4NC}}{2}.$$

Since $n_1 \leq N$, we take

$$n_1 = N - \sqrt{N(N - C)}$$

and

$$n_2 = C - N + \sqrt{N(N-C)}$$

**13.16** (a) $X$ is hypergeometric;

$$P(X=m) = \frac{\binom{n_1}{m}\binom{N-n_1}{n_2-m}}{\binom{N}{n_2}}.$$

(b) In the following, we let $q = n_2 + 1$

$$\begin{aligned}
E[\tilde{N}] &= E\left[\frac{(n_1+1)(n_2+1)}{X+1} - 1\right] \\
&= \sum_{m=0}^{n_2} \frac{\binom{n_1}{m}\binom{N-n_1}{n_2-m}}{\binom{N}{n_2}} \frac{(n_1+1)(n_2+1)}{m+1} - 1 \\
&= \sum_{m=0}^{n_2} (N+1) \frac{\binom{n_1+1}{m+1}\binom{N+1-(n_1+1)}{n_2+1-(m+1)}}{\binom{N+1}{n_2+1}} - 1 \\
&= (N+1) \sum_{k=1}^{q} \frac{\binom{n_1+1}{k}\binom{N-n_1}{q-k}}{\binom{N+1}{q}} - 1 \\
&= (N+1)\left[\sum_{k=0}^{q} \frac{\binom{n_1+1}{k}\binom{N-n_1}{q-k}}{\binom{N+1}{q}} - \frac{\binom{N-n_1}{q}}{\binom{N+1}{q}}\right] - 1.
\end{aligned}$$

The first term inside the brackets is $\sum_k P(Y=k)$ for $Y$ a hypergeometric random variable; it thus equals 1. If $n_2 \geq N - n_1$, then $q > N - n_1$ and $\binom{N-n_1}{q} = 0$. Hence, if $n_2 \geq N - n_1$, $E[\tilde{N}] = N$.

# Chapter 14

# Rare Populations and Small Area Estimation

**14.2** (a) Note that $\bar{y}_1 = \dfrac{\sum_{i\in\mathcal{S}_1} r_i y_i}{\sum_{i\in\mathcal{S}_1} r_i}$, so using properties of ratio estimation [see equation (4.10)],

$$V(\bar{y}_1) \approx \frac{1}{n_1(N_1-1)p_1^2}\sum_{i=1}^{N_1}(r_i y_i - r_i\bar{y}_{1U})^2 = \frac{1}{n_1(N_1-1)p_1^2}(M_1-1)S_1^2 \approx \frac{S_1^2}{n_1 p_1}.$$

Consequently,

$$\begin{aligned} V(\hat{\bar{y}}_d) &= A^2V(\bar{y}_1) + (1-A)^2V(\bar{y}_2) \\ &\approx \frac{A^2S_1^2}{n_1p_1} + \frac{(1-A)^2S_2^2}{n_2p_2}. \end{aligned}$$

(b) With these assumptions, we have $A = \dfrac{N_1p_1}{Np}$, $1-A = \dfrac{N_2p_2}{Np}$, $n_1p_1 = kf_2p_1N_1$, $n_2p_2 = f_2p_2N_2$, and

$$\begin{aligned} V(\hat{\bar{y}}_d) &\approx \frac{S_1^2}{f_2}\left(\frac{A^2}{kN_1p_1} + \frac{(1-A)^2}{N_2p_2}\right) \\ &= S_1^2\left[\left(\frac{N_1p_1}{Np}\right)^2\frac{1}{kf_2N_1p_1} + \left(\frac{N_2p_2}{Np}\right)^2\frac{1}{f_2N_1p_1}\right] \\ &= \frac{S_1^2}{(Np)^2f_2}\left(\frac{N_1p_1}{k} + N_2p_2\right). \end{aligned}$$

The constraint on the sample size is $n = n_1 + n_2 = kf_2N_1 + f_2N_2$; solving for $f_2$, we have

$$f_2 = \frac{n}{N_1k + N_2}.$$

Consequently, we wish to minimize

$$\begin{aligned} V(\hat{\bar{y}}_d) &\approx \frac{S_1^2}{(Np)^2 f_2}\left(\frac{N_1 p_1}{k} + N_2 p_2\right) \\ &= \frac{S_1^2}{(Np)^2 n}(N_1 k + N_2)\left(\frac{N_1 p_1}{k} + N_2 p_2\right), \end{aligned}$$

or, equivalently, to minimize

$$g(k) = (N_1 k + N_2)\left(\frac{N_1 p_1}{k} + N_2 p_2\right).$$

Setting the derivative

$$\frac{dg}{dk} = N_1 N_2 p_2 - \frac{N_1 N_2 p_1}{k^2}$$

to 0 gives $k^2 = p_1/p_2$.

**14.3** (a) The estimator is unbiased because each component is unbiased for its respective population quantity. The variance formula follows because the random variables for inclusion in sample $A$ are independent of the random variables for inclusion in sample $B$.

(b) We take the derivative of the variance with respect to $\theta$.

$$\begin{aligned} \frac{d}{d\theta}V(\hat{t}_{y,\theta}) &= \frac{d}{d\theta}\left\{V\left[\hat{t}_a^A\right] + \theta^2 V\left[\hat{t}_{ab}^A\right] + 2\theta\text{Cov}\left[\hat{t}_a^A, \hat{t}_{ab}^A\right]\right. \\ &\qquad \left. +(1-\theta)^2 V\left[\hat{t}_{ab}^B\right] + V\left[\hat{t}_b^B\right] + 2(1-\theta)\text{Cov}\left[\hat{t}_b^B, \hat{t}_{ab}^B\right]\right\} \\ &= 2\theta V\left[\hat{t}_{ab}^A\right] + 2\text{Cov}\left[\hat{t}_a^A, \hat{t}_{ab}^A\right] - 2(1-\theta)V\left[\hat{t}_{ab}^B\right] - 2\text{Cov}\left[\hat{t}_b^B, \hat{t}_{ab}^B\right] \end{aligned}$$

Setting the derivative equal to 0 and solving gives the optimal value of $\theta$.

**14.7** (a) We write

$$\tilde{\theta}_d(a) - \theta_d = a(\mathbf{x}_d^T\boldsymbol{\beta} + v_d + e_d) + (1-a)\mathbf{x}_d^T\boldsymbol{\beta} - (\mathbf{x}_d^T\boldsymbol{\beta} + v_d).$$

Then

$$E[\tilde{\theta}_d(a) - \theta_d] = E[a(v_d + e_d) - v_d] = 0.$$

(b)

$$\begin{aligned} V[\tilde{\theta}_d(a) - \theta_d] &= E\left\{[a(v_d + e_d) - v_d]^2\right\} \\ &= (a-1)^2\sigma_v^2 + a^2\psi_d \end{aligned}$$

since $e_d$ and $v_d$ are independent.

$$\frac{d}{da}[(a-1)^2\sigma_v^2 + a^2\psi_d] = 2(a-1)\sigma_v^2 + 2a\psi_d;$$

setting this equal to 0 and solving for $a$ gives $a = \sigma_v^2/(\sigma_v^2 + \psi_d) = \alpha_d$. The minimum variance achieved is

$$\begin{aligned} V[\tilde{\theta}_d(\alpha_d) - \theta_d] &= (\alpha_d - 1)^2\sigma_v^2 + \alpha_d^2\psi_d \\ &= \left(\frac{\psi_d}{\sigma_v^2 + \psi_d}\right)^2 \sigma_v^2 + \left(\frac{\sigma_v^2}{\sigma_v^2 + \psi_d}\right)^2 \psi_d \\ &= \frac{\psi_d^2\sigma_v^2}{(\sigma_v^2 + \psi_d)^2} + \frac{\psi_d\sigma_v^4}{(\sigma_v^2 + \psi_d)^2} \\ &= \frac{\psi_d\sigma_v^2(\psi_d + \sigma_v^2)}{(\sigma_v^2 + \psi_d)^2} \\ &= \alpha_d\psi_d. \end{aligned}$$

**14.8** Here is SAS code for construction the population and samples:

```
options ls=78 nodate nocenter;

data domainpop;
   do strat = 1 to 20;
   do psu = 1 to 4;
     do j = 1 to 3;
       y = strat;
   dom = 1;
   output;
end; end;
do psu = 5 to 8;
     do j = 1 to 3;
       y = strat;
   dom = 2;
   output;
end; end;
end;

proc sort data=domainpop;
   by strat psu;

proc print data=domainpop;
run;

/* Select SRS of 2 psus from each stratum */

data psuid;
   do strat = 1 to 20;
   do psu = 1 to 8;
```

```
   output;
   end; end;

proc surveyselect data=psuid out=psusamp1 sampsize=2 seed=425;
   strata strat;

proc sort data=psusamp1;
   by strat psu;

proc print data=psusamp1;
run;

/* Merge back with data */

data samp1 ;
   merge psusamp1 (in=Insample) domainpop ;
    /* When a data set contributes an observation for
       the current BY group, the IN= value is 1. */
   by strat psu;
   if Insample ; /*delete obsns not in sample */
run;

proc print data=samp1;
run;

/* Here is the correct analysis */

proc surveymeans data=samp1 nobs mean sum clm clsum;
   stratum strat;
   cluster psu;
   var y;
   weight SamplingWeight;
   domain dom;
run;

/*Now do incorrect analysis by deleting observations not in domain*/

data samp1d1;
   set samp1;
   if dom = 1;

proc surveymeans data=samp1d1 nobs mean sum clm clsum;
   stratum strat;
   cluster psu;
   var y;
```

```
   weight SamplingWeight;
run;


data samp1d2;
   set samp1;
   if dom = 2;

proc surveymeans data=samp1d2 nobs mean sum clm clsum;
   stratum strat;
   cluster psu;
   var y;
   weight SamplingWeight;
run;
```

# Chapter 15

# Survey Quality

**15.2** This is a stratified sample, so we use formulas from stratified sampling to find $\hat{\phi}$ and $\hat{V}(\hat{\phi})$.

| Stratum | $N_h$ | $n_h$ | yes | $\hat{\phi}_h$ | $\frac{N_h}{N}\hat{\phi}_h$ | $\frac{N_h - n_h}{N_h}\frac{N_h^2}{N^2}\frac{s_h^2}{n_h}$ |
|---|---|---|---|---|---|---|
| Undergrads | 8972 | 900 | 123 | 0.1367 | 0.1077 | $7.34 \times 10^{-5}$ |
| Graduates | 1548 | 150 | 27 | 0.1800 | 0.0245 | $1.66 \times 10^{-5}$ |
| Professional | 860 | 80 | 27 | 0.3375 | 0.0255 | $1.47 \times 10^{-5}$ |
| Total | 11380 | 1130 | 177 | | 0.1577 | $1.05 \times 10^{-4}$ |

Thus $\hat{\phi} = 0.1577$ with $\hat{V}(\hat{\phi}) = 1.05 \times 10^{-4}$. The probability $P$ that a person is asked the sensitive question is the probability that a red ball is drawn from the box, 30/50. Also,

$$p_I = P(\text{white ball drawn} \mid \text{red ball not drawn}) = 4/20.$$

Thus, using (12.10),

$$\hat{p}_S = \frac{\hat{\phi} - (1-P)p_I}{P} = \frac{0.1577 - (1-.6)(.2)}{.6} = 0.130$$

and

$$\hat{V}(\hat{p}_S) = \frac{1.05 \times 10^{-4}}{(0.6)^2} = 2.91 \times 10^{-4}$$

so the standard error is 0.17.

**15.3** (a)

$$\begin{aligned} P(\text{“1”}) &= P(\text{“1”} \mid \text{sensitive})p_s + P(\text{“1”} \mid \text{not sensitive})(1-p_s) \\ &= \theta_1 p_s + \theta_2(1-p_s) \end{aligned}$$

(b) Let $\hat{p}$ be the proportion of respondents who report “1.” Let

$$\hat{p}_s = \frac{\hat{p} - \theta_2}{\theta_1 - \theta_2}.$$

(We must have $\theta_1 \neq \theta_2$.)

(c) If an SRS is taken,

$$\begin{aligned} V(\hat{p}_s) &= \frac{1}{(\theta_1 - \theta_2)^2} V(\hat{p}) \\ &= \frac{1}{(\theta_1 - \theta_2)^2} \frac{p(1-p)}{n-1}. \end{aligned}$$

# Appendix A: Probability Concepts Used in Sampling

**A.1**

$$\begin{aligned} P\text{ (match exactly 3 numbers)} &= \frac{\binom{5}{3}\binom{30}{2}}{\binom{35}{5}} = \frac{(10)(435)}{324{,}632} = \frac{4350}{324{,}632} \\ P\text{ (match at least 1 number)} &= 1 - P\text{ (match no numbers)} \\ &= 1 - \frac{\binom{5}{0}\binom{30}{5}}{\binom{35}{5}} \\ &= 1 - \frac{142{,}506}{324{,}632} = \frac{182{,}126}{324{,}632}. \end{aligned}$$

**A.2**

$$\begin{aligned} P\text{ (no 7s)} &= \frac{\binom{3}{0}\binom{5}{4}}{\binom{8}{4}} = \frac{5}{70} \\ P\text{ (exactly one 7)} &= \frac{\binom{3}{1}\binom{5}{3}}{\binom{8}{4}} = \frac{30}{70} \\ P\text{ (exactly two 7s)} &= \frac{\binom{3}{2}\binom{5}{2}}{\binom{8}{4}} = \frac{30}{70}. \end{aligned}$$

**A.3 Property 1:** Let $Y = g(x)$. Then

$$P(Y = y) = \sum_{x:g(x)=y} P(X = x)$$

so, using (A.3),

$$\begin{aligned} E[Y] &= \sum_y yP(Y=y) \\ &= \sum_y y \sum_{x:g(x)=y} P(X=x) \\ &= \sum_y \sum_{x:g(x)=y} g(x)P(X=x) \\ &= \sum_x g(x)P(X=x). \end{aligned}$$

**Property 2:** Using Property 1, let $g(x) = aX + b$. Then

$$\begin{aligned} E[aX+b] &= \sum_x (ax+b)P(X=x) \\ &= a\sum_x xP(X=x) + b\sum_x P(X=x) \\ &= aE[X] + b. \end{aligned}$$

**Property 3:** If $X$ and $Y$ are independent, then $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all $x$ and $y$. Then

$$\begin{aligned} E[XY] &= \sum_x \sum_y xyP(X=x, Y=y) \\ &= \sum_x \sum_y xyP(X=x)P(Y=y) \\ &= \left[\sum_x xP(X=x)\right]\left[\sum_y yP(Y=y)\right] \\ &= (EX)(EY). \end{aligned}$$

**Property 4:**

$$\begin{aligned} \mathrm{Cov}\,[X,Y] &= E[(X-EX)(Y-EY)] \\ &= E[XY - Y(EX) - X(EY) - (EX)(EY)] \\ &= E[XY] - (EX)(EY). \end{aligned}$$

**Property 5:** Using Property 4,

$$\begin{aligned}
&\mathrm{Cov}\left[\sum_{i=1}^{n} a_i X_i + b_i, \sum_{j=1}^{m} c_j Y_j + d_j\right] \\
&= E\left[\sum_{i=1}^{n}\sum_{j=1}^{m}(a_i X_i + b_i)(c_j Y_j + d_j)\right] \\
&\quad -E\left[\sum_{i=1}^{n}(a_i X_i + b_i)\right] E\left[\sum_{j=1}^{m}(c_j Y_j + d_j)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}[a_i c_j E(X_i Y_j) + a_i d_j EX_i + b_i c_j EY_j + b_i d_j] \\
&\quad -\sum_{i=1}^{n}\sum_{j=1}^{m}[a_i E(X_i) + b_i][c_j E(Y_j) + d_j] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i c_j [E(X_i Y_j) - (EX_i)(EY_j)] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i c_j \mathrm{Cov}\,[X_i, Y_j].
\end{aligned}$$

**Property 6:** Using Property 4,

$$V[X] = \mathrm{Cov}\,(X, X) = E[X^2] - (EX)^2.$$

**Property 7:** Using Property 5,

$$\begin{aligned}
V[X+Y] &= \mathrm{Cov}\,[X+Y, X+Y] \\
&= \mathrm{Cov}\,[X, X] + \mathrm{Cov}\,[Y, X] + \mathrm{Cov}\,[X, Y] + \mathrm{Cov}\,[Y, Y] \\
&= V[X] + V[Y] + 2\mathrm{Cov}\,[X, Y].
\end{aligned}$$

(It follows from the definition of Cov that $\mathrm{Cov}\,[X, Y] = \mathrm{Cov}\,[Y, X]$.)

**Property 8:** From Property 7,

$$\begin{aligned}
V\left[\frac{X}{\sqrt{V(X)}} + \frac{Y}{\sqrt{V(Y)}}\right] &= V\left[\frac{X}{\sqrt{V(X)}}\right] + V\left[\frac{Y}{\sqrt{V(Y)}}\right] \\
&\quad +2\mathrm{Cov}\left[\frac{X}{\sqrt{V(X)}}, \frac{Y}{\sqrt{V(Y)}}\right] \\
&= 2 + 2\mathrm{Corr}\,[X, Y].
\end{aligned}$$

Since the variance on the left must be nonnegative, we have $2 + 2\ \mathrm{Corr}\,[X, Y] \geq 0$, which implies $\mathrm{Corr}\,[X, Y] \geq -1$.

Similarly, the relation

$$0 \leq V\left[\frac{X}{\sqrt{V(X)}} - \frac{Y}{\sqrt{V(Y)}}\right] = 2 - 2\ \mathrm{Corr}\,[X, Y]$$

implies that $\text{Corr}\,[X, Y] \leq 1$.

**A.4** Note that $Z_i^2 = Z_i$, so $E[Z_i^2] = E[Z_i] = n/N$. Thus,

$$\begin{aligned} V[Z_i] &= E[Z_i^2] - [E(Z_i)]^2 \\ &= \frac{n}{N} - \left(\frac{n}{N}\right)^2 \\ &= \frac{n(N-n)}{N^2}. \end{aligned}$$

$$\begin{aligned} \text{Cov}\,[Z_i, Z_j] &= E[Z_i Z_j] - (EZ_i)(EZ_j) \\ &= \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 \\ &= \frac{n(n-1)N - n^2(N-1)}{N^2(N-1)} \\ &= -\frac{n(N-n)}{N^2(N-1)}. \end{aligned}$$

**A.5**

$$\begin{aligned} \text{Corr}\,[\bar{x}, \bar{y}] &= \frac{\text{Cov}\,[\bar{x}, \bar{y}]}{\sqrt{V[\bar{x}]V[\bar{y}]}} \\ &= \frac{\frac{1}{n}(1 - \frac{n}{N})RS_xS_y}{\sqrt{[\frac{1}{n}(1 - \frac{n}{N})S_x^2][\frac{1}{n}(1 - \frac{n}{N})S_y^2]}} \\ &= R. \end{aligned}$$